TREBALL DE FI DE GRAU

Doble Grau en Economia i Estadística

Títol: Anàlisi del consum i els costos ocults de la carn a les llars espanyoles

Autor: Cristina Quesada Grau

Director: Lluís Bermúdez Morata

Departament: Departament de Matemàtica Econòmica, Financera i Actuarial

Convocatòria: Juny 2022

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

ANÀLISI DEL CONSUM I ELS COSTOS OCULTS DE LA CARN A LES LLARS ESPANYOLES

CRISTINA QUESADA GRAU

TREBALL FINAL DE GRAU

TUTOR: LLUÍS BERMÚDEZ MORATA
DEPARTAMENT DE MATEMÀTICA ECONÒMICA, FINANCERA I ACTUARIAL

CONVOCATÒRIA JUNY 2022

UNIVERSITAT DE BARCELONA – UNIVERSITAT POLITÈCNICA DE CATALUNYA

UNIVERSITAT DE BARCELONA
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Agraïments

Als meus pares, a la meva àvia i a la meva germana pels seus savis consells, per la seva comprensió, paciència i ànims. Gràcies per estar sempre al meu costat, estar orgullosos de mi i animar-me a seguir endavant.

Al meu tutor Lluís Bermúdez, per ser el meu guia, pel suport, acompanyament, i organització durant tot el procés d'elaboració.

A la Universitat de Barcelona i a la Universitat Politècnica de Catalunya per brindar-me la millor formació possible durant els cinc anys de grau.

A tots vosaltres, moltes gràcies per tot.

Resum

Des de la segona meitat del segle XX fins a l'actualitat s'han anat introduint nous models de gestió ramadera que es basen en l'alimentació complementària dels animals, passant d'un sistema de producció extensiu, en el qual s'utilitzen mètodes tradicionals que repliquen els ecosistemes naturals perquè siguin favorables per al desenvolupament de l'animal, a un sistema de producció intensiva on el principal objectiu és incrementar la producció de carn i altres productes animals en el menor temps possible. En conseqüència, el consum de carn s'ha convertit en un dels grans debats nutricionals de l'actualitat, abastant diferents aspectes com la salut, la nutrició, el medi ambient, el benestar animal i l'ètica. Per això, sorgeix la necessitat de conèixer el perfil de les llars que la consumeixen i dels quals no, per a comprendre com els factors domèstics, incideixen en els costos ambientals associats a la carn.

Paraules clau: Gestió Ramadera, Sistema de Producció, Consum de Carn, Debats Nutricionals, Medi Ambient, Perfil, Factors Domèstics, Costos Ambientals.

Abstract

From the second half of the 20th century to the present, new livestock management models, which are based on complementary feeding of animals have been introduced, moving from an extensive production system, in which traditional methods that replicate natural ecosystems are used so that they are favourable for the animal's development, to an intensive production system where the main objective is to increase the production of meat and other animal products in the shortest time possible. Consequently, meat consumption has become one of the great nutritional debates of today, covering different aspects such as health, nutrition, the environment, animal welfare and ethics. For this reason, there is a need to know the profile of the households that consume it and those that do not, in order to understand how domestic factors, affect the environmental costs associated with meat.

Key words: Livestock Management, Production System, Meat Consumption, Nutritional Debates, Environment, Profile, Domestic Factors, Environmental Costs.

Índex de contingut

Índex de Figures

Índex de Taules

1. El SECTOR CARNI ESPANYOL

1.1. Aspectes generals

En l'àmbit espanyol, la indústria càrnia ocupa la quarta posició en importància respecte als grans sectors industrials, situant-se per darrere de la indústria automobilística, la indústria del petroli i combustibles o subministrament d'energia i de la indústria química o metal·lúrgica. Pel que fa a la indústria espanyola d'aliments i begudes, lidera la primera posició amb una xifra de negocis de 31.727 milions d'euros, equivalent al 28,5% de tot el sector alimentari espanyol [(1)]. El teixit empresarial del sector carni es troba constituït per 2.800 empreses, gran part d'elles petites i mitjanes empreses relacionades amb activitats als escorxadors, sales de despulles i indústries d'elaborats carnis.

En termes econòmics, la xifra de negoci l'any 2021 suposa el 2,55% del PIB total espanyol, el 17,22% del PIB de la branca industrial i el 4,66% de la facturació total de la indústria a Espanya. Actualment, el sector industrial proporciona 105.396 llocs de treball els quals representen un 28,9% de l'ocupació total de la indústria alimentària espanyola.

A escala nacional i internacional, tot i que el sector carni ha estat marcat per la Covid-19, durant els últims anys s'ha assolit el record anual amb 3,24 milions de tones de carn i 212.443 tones de productes elaborats, per un valor econòmic de 9.107 milions d'euros, obtenint així una balança comercial positiva del 712% [(1)]. És per això que les exportacions càrnies espanyoles tenen un gran pes dins del sector carni.

1.2. Estructura del sector carni

Els sectors més importants de la indústria càrnia són el porcí, boví, oví i caprí i l'avícola. El **sector porcí** representa aproximadament 14% de la Producció Final Agrària i el 39% de la Producció Final Ramadera i a escala europea, ocupa la segona posició després d'Alemanya. Gràcies als mercats exteriors, en els últims anys el sector porcí ha experimentat un creixement notable en l'àmbit productiu, de censos i en el nombre d'explotacions. Aquest augment en la producció ha permès un increment de les exportacions convertint-se així en un element clau per l'equilibri de mercat. El **sector boví** representa el 5,7% de la Producció Final Agrària i el 15,3% respecte a la Producció Final Ramadera i en l'àmbit europeu, Espanya ocupa la tercera posició després de França

i Alemanya. Durant els últims anys, a causa d'una disminució del consum intern, la internacionalització del sector ha estat clau pel seu creixement. Referent al **sector oví i caprí**, aquest representa l'11% de la Producció Final Ramadera a Espanya i dins de la Unió Europea també ocupa la segona posició en termes de rellevància. A més, juga un paper molt important en termes de composició del territori, la conservació de l'entorn i en la creació de llocs de treballs en les zones rurals. Finalment, el **sector avícola,** que representa l'11,8% de la Producció Final Ramadera, es troba constituït principalment per la carn de pollastre i de gall d'indi, europeament és el segon país productor de carn de pollastre després del Regne Unit [(2)].

El comportament en la compra i el consum ha anat evolucionant al llarg del temps en funció de les condicions socioeconòmiques del país i la situació del sector. Els últims dos anys han estat marcats per la pandèmia i el confinament, el qual s'ha vist reflectit en les xifres; en termes de volum, l'any 2020 el consum de carn i productes carnis va augmentar en un 9,4% mentre que en el 2021 va caure en un 8,4%. Associat a la despesa, en 2020 es va experimentar un augment del 12% tot i que en el 2021 va caure en un 6,2% enfront l'any anterior. Pel que fa al nivell de consum per càpita es va desplomar aproximadament en un 12%, arribant als 45,5 kg.

Els patrons de consum de les llars espanyoles varien en funció de l'època de l'any. En el cas de la **carn de boví**, aquest és major en mesos freds i menor en els mesos d'estiu, tot i que de manera genèrica té un comportament estable al llarg de l'any. Referent al **consum de carn de porcí**, a conseqüència del confinament a causa de la crisi sanitària, a l'any 2020 es va experimentar un ascens en el consum de carn fresca i carn transformada. Per altra banda, el **consum de carn d'oví i caprí** va tenir un comportament constant durant tot l'any menys al mes de desembre on va arribar als seus valors màxims. Finalment, el **consum de carn avícola** també es va veure repercutit, durant el 2020 va ser ascendent gràcies a la situació extraordinària del país, però aquest augment no va compensar el descens d'anys anteriors.

1.3. Els costos ocults associats a la carn

Actualment, l'adopció de dietes no tradicionals està augmentant per diverses raons (religioses, ètiques, ecològiques, econòmiques…). El motiu principal del canvi

d'alimentació és fonamentalment la sostenibilitat, pel fet que s'han utilitzat menys recursos naturals, no són tan nocius pel medi ambient i promouen la justícia social a escala mundial [(3)].

Per altra banda, el procés de producció de la indústria càrnia té associat dos tipus de costos; els costos de producció i els costos ocults. Els **costos de producció** (també anomenats costos d'operació) fan referència a les despeses necessàries per mantenir projectes, línies de processament o equips en funcionament. Per altra banda, els **costos ocults** són aquelles despeses no imputables, ni directament ni indirectament a cap element que generi valor per a l'organització, és a dir, es tracten de despeses que no són productives o necessàries per al funcionament de l'empresa i solen passar desapercebudes per als sistemes comptables i el compte de resultats [(4)].

Des de la dècada dels anys 90 fins a l'actualitat, el canvi climàtic ha estat un dels temes que ha anat guanyant importància. Tot i que el clima ha estat present en les últimes dues reunions dirigides per les Nacions Unides (París l'any 2015 i Bonn al 2017), l'amenaça de la superpoblació no ho ha estat i aquest últim fenomen afecta tant al canvi climàtic, com a la pèrdua de diversitat biològica, a la seguretat alimentària, l'aigua, a les malalties, a la contaminació i a l'energia.

Així com els diferents tipus d'aliments tenen diferents preus en el mercat, també existeixen diferents costos ambientals associats a la producció de cada tipus d'aliment. Això és el que es denomina **petjada ambiental** [(5)]. Els costos ocults associats a la ramaderia són principalment dos: el cost atmosfèric i el cost dels recursos hídrics. En un segon pla es troben el cost de la biodiversitat i del sòl. Però aquests últims no queden recollits en aquest document.

Contaminació atmosfèrica

La **petjada de carboni** es defineix com la totalitat de gasos d'efecte hivernacle (GEI) emesos per efecte directe o indirecte per un individu, organització, esdeveniment o producte. A Espanya, el sector agrari és el segon major emissor de gasos d'efecte hivernacle, després del sector del transport. Els principals gasos emesos pel sector són el diòxid de carboni (CO_2), l'òxid nitrós (N_2O) i el metà (CH_4) [(6)]. La mesura kilograms de diòxid de carboni equivalent (kg CO2-eq) és una manera de representar l'emissió total de

gasos d'efecte hivernacle (GEI) en una única mesura a través de la conversió que té en compte la massa dels gasos i la capacitat calorífica (7).

El Protocol de Kyoto el qual va entrar en vigor el 16 de febrer de 2005 va posar de manifest la Convenció Marco de les Nacions Unides sobre el Canvi Climàtic on comprometia als països industrialitzats a reduir les emissions de GEI. Amb la intenció d'aconseguir els objectius (marcats pel Protocol) al menor cost possible va sorgir el mercat de carboni. Aquests mercats es basen en la venda o adquisició de bons de carboni[1] a un determinat preu. El funcionament d'aquest mercat és el següent; els bons són repartits entre les principals empreses emissores dels gasos d'efecte hivernacle les quals per llei, estan obligades a emetre una quantitat igual o inferior a la quantitat de bons que posseeixen (8). Tanmateix, aquests poden ser venuts o comprats i és aquí on entra en joc els mercats de carboni. De tal manera que si una empresa no consumeix els seus bons, els pot vendre a una altra empresa que sí que els consumeixi superant així la quantitat permesa que se li havia assignat a l'inici (9).

Les principals fonts d'emissions (10) produïdes pel sector ramader són:

- La **fermentació entèrica** és un procés que es dona en l'aparell digestiu dels remugants i monogàstrics[2], i està lligada amb la producció de metà. La quantitat de gas produït per l'animal dependrà de la constitució del seu aparell digestiu i de la seva alimentació. Aquesta font representa el 44% del total de les emissions del sector.
- Els diferents sistemes de **gestió dels fems** donen lloc a diferents nivells d'emissió de metà i òxid nitrós. El primer d'ells, es genera en la descomposició anaeròbica de la matèria orgànica i el segon, en el procés de descomposició de l'amoníac dels fems. Aquesta font d'emissió representa el 10% del total de les emissions del sector.
- La **producció de pinsos** per al bestiar dona lloc a l'emissió diòxid de carboni i òxid nitrós. El primer d'ells s'origina en la fabricació de fertilitzants i pesticides pels cultius, del processat i el transport així com l'expansió de pastures i terres de cultiu destinades a l'alimentació del bestiar. Mentre que el segon d'ells, s'origina per l'ús de fertilitzants de nitrogen i la concentració de fems. Aquesta font representa el 41% del total de les emissions del sector.

[1] Documents que capaciten al propietari a emetre una determinada quantitat de CO2 i GEI
[2] Espècie corresponent al regne animal que es caracteritza per tenir un sol estómac o ventre

- El **consum d'energia** que s'utilitza en el procés de fabricació de fertilitzants, ús de la maquinària, el processat i transport de cultius per tal d'alimentar als animals, el processat, l'embassat , l'empaquetat i el transport de productes, genera emissions. Del total de les emissions del sector, aquesta font representa el 5%.

Cost de l'aigua

Estudis recents han demostrat que la indústria ramadera té associada un elevat consum de recursos hídrics. Concretament, el 27% de la petjada hídrica humana mundialment està associada amb la producció de productes animals i el 4% amb l'ús domèstic de l'aigua (11).

Hoekstra defineix el concepte de **petjada hídrica** com un indicador de l'ús de l'aigua en relació amb els béns de consum. Concretament, la petjada hídrica associada a un producte és el volum d'aigua dolça necessària per produir-lo, mesurat en les diferents fases de la producció (mesurada en termes d'aigua consumida (evaporada) o contaminada) i es desglossa en tres components:

- **Petjada hídrica blava:** volum d'aigua dolça que s'evapora en les aigües superficials i subterrànies.
- **Petjada hídrica verda**: volum d'aigua evaporada de l'aigua de la pluja emmagatzemada en el sòl.
- **Petjada hídrica grisa:** volum d'aigua necessària per diluir els contaminants fins al punt que la qualitat de l'aigua es mantingui per sobre de les normes de qualitat.

La cadena de subministrament de productes carnis comença des del cultiu de pinsos fins al consumidor final. En cada una de les etapes implicades existeix un consum d'aigua de manera directa i indirecta.

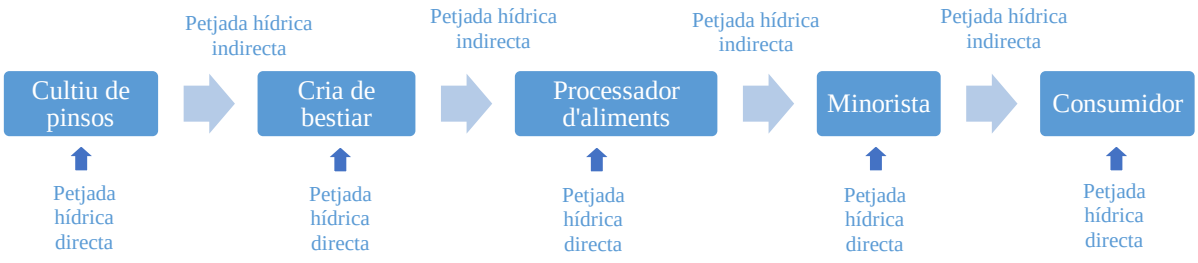


Figura 1. *La petjada hídrica directa i indirecta a cada etapa de la cadena de subministrament d'un producte animal. Font: Arjen Y. Hoekstra.*

Un estudi realitzat per Mekonnen y Hoekstra a l'any 2010 [(11)], va demostrar sota la condició d'un valor nutricional equivalent, que la petjada hídrica del producte animal era superior a la d'un producte vegetal . Globalment, la petjada hídrica de la producció animal és de 2.422 milions de m^3/any; el 87% correspon a la petjada hídrica verda, el 6% a la blava i el 7% restant a la grisa.

LA MITJANA GLOBAL DE LA PETJADA HÍDRICA

	Petjada hídrica per unitat de pes, m^3/Tn			
Aliment	**Verda**	**Blava**	**Gris**	**Total**
Cultiu de sucre	130	52	15	197
Verdures	194	43	85	322
Arrels de fècula	327	16	43	387
Fruita	726	147	89	962
Cereals	1.232	228	184	1.644
Cultius d'olis	2.023	220	121	2.364
Llegums	3.180	141	734	4.055
Fruits secs	7.016	1.367	680	9.063
Llet	863	86	72	1.020
Ous	2.592	244	429	3.265
Pollastre	**3.545**	**313**	**467**	**4.325**
Mantega	4.695	465	393	5.553
Porcí	**4.907**	**459**	**622**	**5.988**
Oví i caprí	**8.253**	**457**	**53**	**8.763**
Boví	**14.414**	**550**	**451**	**15.415**

***Taula 1**. Mitjana global de la petjada hídrica. Font: Arjen Y. Hoekstra*

Segons la FAO, el consum mitjà de calories per persona és de 3.400 kcal/dia en els països industrialitzats, el qual el 30% d'aquestes corresponen a productes d'origen animal. La disparitat entre els efectes d'una dieta vegetariana i una dieta càrnia són molt grans, ja que 1 kcal de producte d'origen animal requereix aproximadament en mitjana 2,5 L d'aigua mentre que un producte d'origen vegetal requereix en mitjana 0,5 L d'aigua. Per tant, **els consumidors podrien reduir la petjada hídrica reduint el volum de carn consumit o seleccionant aquells productes carnis que tinguin una petjada hídrica menor.**

2. MACHINE LEARNING

El *Machine Learning* (ML) o l'aprenentatge automàtic és una branca de la intel·ligència artificial (IA) que, a través dels algoritmes, dota als ordinadors de la capacitat per identificar patrons en dades massives i elaborar-ne prediccions [(12)].

Durant els últims anys, el ML s'ha convertit en una disciplina aplicada en tots els àmbits d'investigació tant a escala acadèmica com industrial. Concretament, l'agricultura juga un paper fonamental en l'economia espanyola i amb la contínua expansió de la població humana, la pressió del sector agrícola tendirà a augmentar en un futur pròxim [(13)]. Davant l'escenari on les dades canvien constantment segons les situacions socioeconòmiques del país, l'aprenentatge automàtic pot ser un instrument útil gràcies a la seva gran capacitat adaptativa.

Tal com es mostra a la ***Figura 2***, els algoritmes més comuns del *Machine Learning* es poden classificar en dues grans categories: l'aprenentatge supervisat i l'aprenentatge no supervisat [(14)].

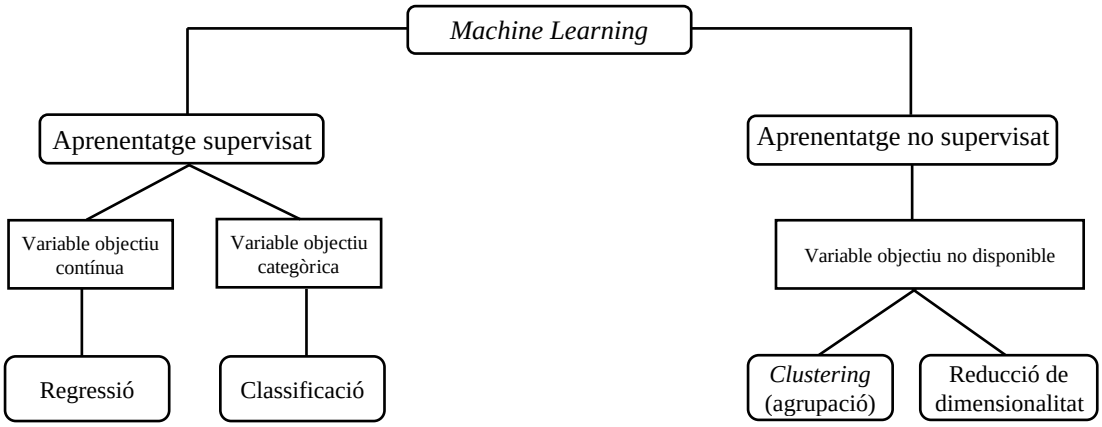


Figura 2. *Classificació dels models de Machine Learning. Font pròpia.*

En l'**aprenentatge supervisat**, el model s'entrena a partir d'un conjunt d'exemples en què tant els resultats d'entrada com els de sortida són coneguts [(15)]. L'algoritme s'alimenta de les dades d'entrenament les quals inclouran les solucions desitjades, anomenades etiquetes. Aquest aprendrà la relació entre la variable resposta i els predictors amb l'objectiu de predir exactament les respostes per dades futures que no hagin estat involucrades en el procés d'entrenament, és a dir, aquelles en què l'etiqueta és desconeguda.

La tipologia de prova emprada en l'aprenentatge supervisat dependrà del caràcter de la variable resposta que es desitgi estudiar. Si es tracta d'una variable qualitativa, s'anomenaran problemes de classificació, mentre que si és quantitativa, s'anomenaran problemes de regressió. Tanmateix, la categoria dels predictors és totalment independent a l'hora d'escollir el model.

En **l'aprenentatge no supervisat,** les dades etiquetades no estan a l'abast per l'entrenament, és a dir, només es coneixen les dades d'entrada però no les de sortida**.** Mitjançant l'exploració de l'estructura de les dades, l'objectiu de l'algoritme serà extreure informació rellevant sense la necessitat de conèixer la variable resposta [(14)].

Principalment, hi ha dos tipus d'aprenentatge no supervisat; el *clustering* o agrupació, el qual organitza la informació en grups similars i, la reducció de dimensionalitat, la qual elimina aquelles variables que contenen informació redundant obtenint així un subespai més reduït i al mateix temps que continguin gran part de la informació rellevant [(15)].

2.1. Elements bàsics

2.1.1. *Overfitting i underfitting*

Els conceptes *d'overfitting* (sobre ajust) i *underfitting* (falta d'ajust) es troben relacionats amb les causes principals d'un baix rendiment del model produït durant l'entrenament d'aquest. En termes generals, es produeix el sobre ajust quan el model s'ajusta perfectament a les dades d'entrenament i acaba memoritzant els patrons i les fluctuacions de soroll de les dades i com a conseqüència, l'algoritme no és capaç de predir correctament noves observacions [(16)]. Les principals causes d'aquest problema solen ser degudes al fet que no hi ha suficients dades en el conjunt d'entrenament o quan l'entrenament d'un model dura molt de temps [(17)]. Per altra banda, la falta d'ajust en un model es dona quan aquest és incapaç de recollir correctament la relació entre les variables d'entrada i de sortida, generant així una elevada taxa d'error tant en el conjunt d'entrenament com en el conjunt de test. Se solen donar quan els models són molt simples o bé quan els models no s'han entrenat suficient temps [(18)].

Com a conseqüència, apareix un problema d'equilibri entre el biaix, que ens permet quantificar en mitjana com de lluny es troben les prediccions respecte als valors reals, i

la variància, que fa referència a quant canvia el model segons les dades utilitzades en el procés d'entrenament. En el cas del sobre ajust tendeix a reduir el biaix però a augmentar la variància, mentre que la falta d'ajust tendeix a augmentar el biaix, però a reduir la variància. Per tant, el millor model serà aquell que aconsegueixi un major equilibri entre el biaix i la variància [19].

Per detectar aquests problemes cal establir un conjunt d'observacions, de les quals es conegui la variable resposta, però que el model no hagi "vist", és a dir, que no hagin participat en l'estimació del model. D'aquesta manera, se separa de manera aleatòria el conjunt de dades disponibles en el subconjunt d'entrenament i en el conjunt de test. El conjunt d'entrenament representa la majoria de les dades (al voltant del 80%) amb les que s'entrena el model i la qualitat d'aquest serà directament proporcional a la qualitat de les dades. Mentre que el conjunt de test fa referència a una petita part del conjunt de dades (aproximadament el 20%) i s'utilitza per comprovar si el model que s'ha generat a partir de les dades d'entrenament funciona correctament, és a dir, si les prediccions de la variable resposta del model per un cas totalment nou són correctes o no.

Com a solució al problema *d'overfitting* o *underfitting*, s'utilitzen els mètodes d'assemblatge els quals combinen prediccions de diferents models amb l'objectiu d'aconseguir un equilibri entre el biaix i la variància. Els mètodes més usats són el ***bagging*** i el ***boosting*** [19]. El primer d'ells ajusta diversos models i cadascun d'ells fa servir un subconjunt diferent de dades d'entrenament. En canvi, el segon, ajusta diversos models seqüencialment, de manera que cada un dels models aprèn dels errors comesos per l'anterior model.

2.1.2. Problema de desequilibri de classes

La classificació en l'entorn de l'aprenentatge estadístic fa referència als models predictius en el que es prediu l'etiqueta de classe. Aquestes etiquetes poden pertànyer a dues o més classes.

El desequilibri de classes es produeix quan la majoria de les dades pertanyen a una etiqueta de classe. Els algoritmes de ML assumeixen que les dades es troben repartides de forma equitativa entre les classes. Per tant, quan es dona el desequilibri de classe, el

classificador de l'algoritme tendeix a estar esbiaixat cap a la classe majoritària. Com a conseqüència, l'algoritme no és capaç de classificar correctament la classe minoritària.

Generalment, existeixen dues mètriques [20] que permeten detectar el problema:

1. **Dependents del llindar (*Threshold-dependent*):** Aquest grup inclou mètriques com l'exactitud i la precisió, entre d'altres. Per conèixer el seu valor, requereixen el càlcul d'una matriu de confusió utilitzant un límit per calcular les probabilitats predites de cada classe. Davant la situació del problema de classes desequilibrades, aquestes mètriques solen ser insuficients, ja que el programari estadístic assumeix un llindar del 0,50, fet que fa que el model predigui que totes les observacions pertanyen a la classe majoritària. Per aquesta raó no es faran servir en la pertinent anàlisi.
2. **Invalidesa del llindar (*Threshold-invariant*):** Engloba mètriques com l'àrea sota la corba ROC (AUC), que quantifica la taxa de veritables positius en funció de la taxa de falsos positius segons diferents llindars de classificació. En el cas del desequilibri de classes, aquesta mètrica és una bona mesura per avaluar els diferents models predictius.

Per solucionar la qüestió sobre el desequilibri de classes, es poden emprar dues estratègies [21]:

1. **A priori. Mostreig:** És l'estratègia més directa i, consisteix a modificar la distribució de les classes al conjunt d'entrenament per tal d'equilibrar les proporcions. Hi ha dues aproximacions diferents: el sobre mostreig i el submostreig. El primer d'ells consisteix a replicar de manera aleatòria els casos de la classe minoritària, mentre que el segon, elimina aleatòriament casos de la classe majoritària.
2. **A posteriori. Punts de tall al voltant del percentatge real:** Els algoritmes permeten modificar la manera com es realitza l'aprenentatge i esbiaixar-los (canviant el *threshold o* punt de tall) cap a les classes minoritàries, aquest fet rep el nom d'aprenentatge sensible als costos.

2.1.3. Mètriques

L'avaluació de qualsevol algoritme en l'àmbit del *Machine Learning* és una part essencial de qualsevol projecte. És per això que en aquest apartat s'explicaran les mètriques més comunes segons si es tracta d'un problema de classificació o de regressió.

2.1.3.1.Mètriques de classificació

La matriu de confusió és molt útil per mesurar la capacitat predictiva del model, ja que engloba quatre escenaris possibles entre els valors reals i els valors predits. La matriu de confusió es defineix com:

		Classe real	
		$Y = 1$	$Y = 0$
Classe predita	$\hat{Y} = 1$	**Verdaders positius (TP)**	**Falsos positius (FP)**
	$\hat{Y} = 0$	**Falsos negatius (FN)**	**Verdaders negatius (TN)**

***Figura 3**. Matriu de confusió. Font pròpia.*

Els verdaders positius (TP) fan referència al nombre de casos en què el valor real i predit són 1; els verdaders negatius (TN) corresponen al nombre de casos en què el valor real i predit són 0; els falsos positius (FP) són el nombre de casos en què el valor real pren valor 0 i el valor predit pren valor 1 i finalment, els falsos negatius (FN) fan referència al nombre de casos en què el valor real pren valor 1 i el valor predit pren valor 0.

A partir de la matriu de confusió es poden calcular les següents mètriques [22]:

- **Exactitud (AC):** proporció de resultats predits com a positius i negatius entre el total de prediccions**,** és a dir, representa la taxa de prediccions correctes sobre el total i, per tant, com major sigui la dispersió menor serà l'exactitud.

$$AC = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precisió (P):** proporció dels resultats positius classificats correctament entre el total de prediccions positives, és a dir, fa referència a com de prop està el resultat de la predicció del veritable valor.

$$P = \frac{TP}{TP + FP}$$

- **Sensibilitat (Sens):** proporció de resultats predits com a positius entre els resultats positius observats, és a dir, representa la taxa de verdaders negatius.

$$Sens = \frac{TP}{TP + FN}$$

- **Especificitat (Sp)**: proporció de resultats predits com a negatius entre els resultats negatius observats, és a dir, representa la taxa de verdaders negatius.

$$Sp = \frac{TN}{FP + TN}$$

No obstant això, com s'ha mencionat anteriorment, quan es dona el problema de classes desequilibrades, l'Exactitud i la Precisió no són una mesura adequada. Com a alternativa, se sol recórrer a la Sensibilitat i l'Especificitat.

A partir de la Sensibilitat i l'Especificitat es pot traçar la corba ROC, la qual mesura el rendiment del model, descriu i compara la precisió de les prediccions. Per altra banda, també es pot calcular l'àrea sota la corba ROC (AUC) definida com la probabilitat de què el model classifiqui a l'atzar una observació positiva per sobre d'una negativa.

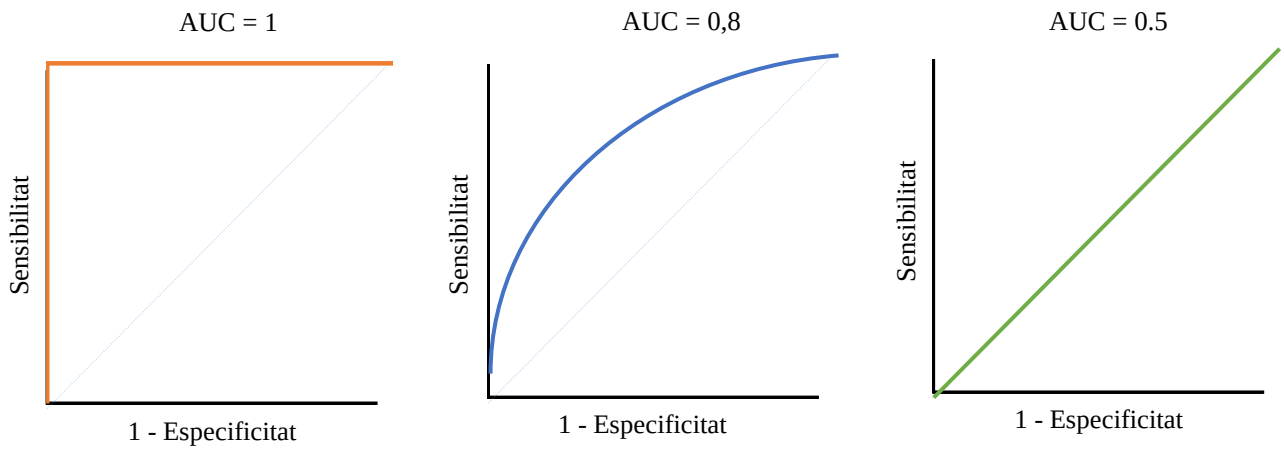


Figura 4. *Representació gràfica de la corba ROC. Font pròpia.*

Tal com es posa de manifest a la ***Figura 4***, aquells models que proporcionin corbes més properes a la cantonada superior esquerra o prenguin valors AUC superiors a 0,5,

indiquen un rendiment bo. En canvi, com més proper estigui la corba a la diagonal de 45 graus o quan el valor AUC sigui igual o inferior a 0,5, el rendiment es considerarà dolent.

2.1.3.2.Mètriques de regressió

En aquesta secció s'exposaran les mètriques [23] més populars per avaluar els models de regressió i el seu respectiu càlcul.

- **Error Quadràtic Mig (MSE):** és la mètrica més simple. Mesura l'error quadrat mitjà de cada una de les prediccions, és a dir, per cada punt, calcula la diferència al quadrat entre el valor predit i el valor real. Finalment, es realitza la mitjana d'aquests valors.

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

- **Arrel de l'Error Quadràtic Mig (RMSE):** es realitza l'arrel quadrada del MSE. D'aquesta manera es penalitza amb major força aquells errors de major magnitud.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

- **Error Absolut Mitjà (MAE):** es calcula com la mitjana de les diferències absolutes entre els valors reals i les prediccions.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- **Coeficient de determinació (R^2):** reflecteix la bondat de l'ajust d'un model i es calcula com el coeficient entre la variància explicada pel model i la variància total. El valor del coeficient està comprès entre 0 i 1, sent 0 el valor mínim, el qual el model no és capaç d'explicar la relació entre la variable resposta i els seus predictors i sent 1 el valor màxim, el qual el model explica perfectament la relació entre la variable que es vol estudiar i els seus predictors.

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y}_j)^2}$$

- **Coeficient de determinació ajustat (R^2 ajustat):** és una correcció del coeficient de determinació i representa el grau d'efectivitat que tenen els predictors per explicar la variable resposta.

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

2.2.Tècniques de ML

Entre les nombroses tècniques de *Machine Learning* existents podem destacar principalment la Regressió Lineal Múltiple, la Regressió Logística, els Arbres de Decisió, *Bagging*, *Random Forest* i *Gradient Boosting*. A continuació, es realitzarà una breu descripció de cadascuna d'elles.

2.2.1. Regressió Lineal Múltiple

La Regressió Lineal Múltiple és una tècnica de modelat estadístic utilitzada per descriure una variable resposta contínua en funció de més d'una variable predictora. L'equació general que correspon al model és la següent [24]:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

On els coeficients β s'interpreten com l'efecte mitjà sobre la variable resposta, Y, davant un increment unitari de la variable predictora, X_k, i el terme ϵ representa el terme d'error del model. Referent a l'estimació dels paràmetres, aquesta s'efectua mitjançant el mètode de Mínims Quadrats Ordinaris (MQO) el qual pren la següent expressió:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

2.2.2. Regressió logística

La Regressió Logística és un model d'aprenentatge supervisat que permet estimar la probabilitat d'una variable qualitativa en funció d'una o diverses variables numèriques o categòriques [25]. En general, aquests models són adequats quan la variable resposta és polinòmica, és a dir, quan admet diferents categories de resposta. Concretament, és especialment útil quan la variable resposta és dicotòmica, és a dir, que pren dos possibles valors. Aquest últim cas és el més comú i el que s'abordarà en aquest treball.

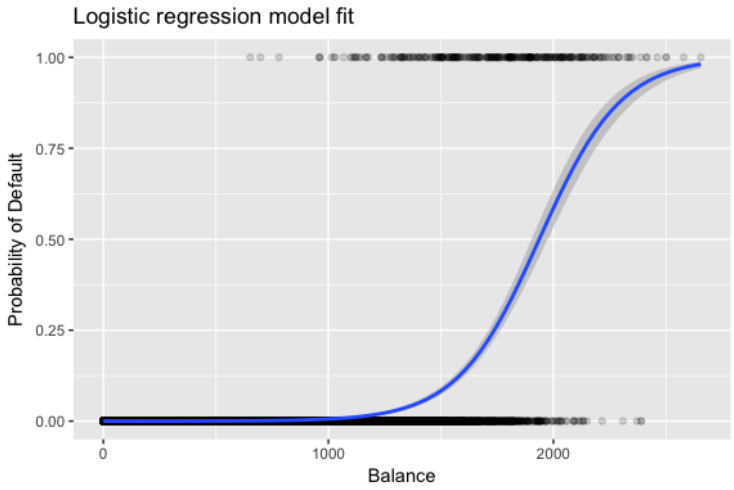


Figura 5. *Ajust del model de la Regressió Logística. Font: Bradley Boehmke & Brandon, Hands-On Machine Learning with R, 2020*

Sigui Y una variable dependent binària (amb dos possibles valors: 0 i 1), la probabilitat de resposta positiva (Y=1) i un conjunt k de variables independents (X_1, X_2, ..., X_k) observades amb la finalitat de predir o explicar el valor de Y:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

On l'estimació dels paràmetres es realitza a partir del mètode de màxima versemblança. Si el predictor és continu, els coeficients s'interpreten com el canvi en el logaritme natural de les probabilitats de l'esdeveniment de referència respecte a un increment unitari del predictor. Si el predictor és categòric, els coeficients s'interpreten com el canvi en el logaritme natural de les probabilitats quan passa del nivell de referència a un altre nivell de la variable predictora.

Per poder interpretar els paràmetres obtinguts, es descriu l'equació del model en termes d'*Odds i Odds Ratio (OR)*. Els *odds* representen la proporció de probabilitat que passi un esdeveniment i el seu complementari:

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon).$$

Per tant, el **valor de la probabilitat de resposta positiva en funció de les variables predictores** es pot trobar aplicant la inversa del logaritme natural:

$$\pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon)}.$$

L'*Odds Ratio (OR)* és una mesura d'associació entre dues variables i indica la fortalesa de la relació entre dues variables i pren la següent expressió:

$$OR_i = \exp(\beta_i).$$

Una de les principals aplicacions del model de regressió logística és classificar la variable resposta binària en funció dels valors que prenguin els predictors. Per aconseguir-ho és necessari establir un *threshold* o punt de tall el qual a partir d'aquella probabilitat es considerarà que la variable pertany a una classe o a una altra. Si no s'indica el contrari, s'assumirà que el *threshold* pren valor de 0,5 [26].

2.2.3. Arbres de decisió

Els arbres de decisió formen part dels algoritmes d'aprenentatge supervisat i són models predictius no paramètrics formats per regles binàries (Si/No o Compleix/No compleix) que tenen com a finalitat distribuir les observacions segons els seus atributs i predir el valor de la variable resposta.

L'algoritme consisteix a dividir l'espai de característiques en una sèrie de regions més petites (que no se superposen) amb valors de resposta semblants fent ús d'un conjunt de regles de divisió.

Gràficament, es representa mitjançant un conjunt de nodes, fulles i branques tal com es mostra a la ***Figura 6***. En el node principal es troba l'atribut o variable amb la qual es comença el procés de classificació, els nodes interns fan referència a cadascuna de les preguntes sobre l'atribut en particular del problema i cada possible resposta apareix representada per un node secundari del qual surten les branques que mostren els possibles valors que pot prendre l'atribut. Per últim, es troba el node final o node fulla que correspon amb la decisió la qual determina la classe de la variable resposta [27].

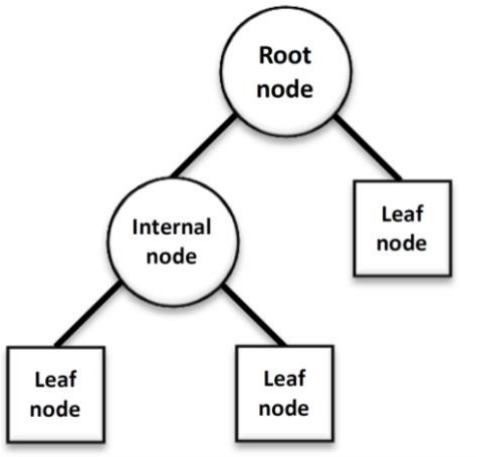


Figura 6. *Estructura del arbre de decisió. Font: ResearchGate*

Un dels principals problemes a l'hora de construir arbres de decisió és l'elevada quantitat de nodes generats i sovint pot conduir a un problema de sobre ajust. Per evitar-ho és necessari aplicar la poda que consisteix a retallar l'arbre amb l'objectiu de reduir la seva dimensió. Un altre inconvenient associat a aquest tipus d'algoritme és la seva inestabilitat, ja que qualsevol petit canvi en les dades pot suposar un arbre totalment diferent [15].

Alguns dels avantatges principals d'aquest tipus de models són la seva fàcil interpretació i no requereixen un preprocessament de les dades massa exigent[28]. També són capaços de treballar tant amb variables qualitatives com continues i si l'arbre no és excessivament gran, aquest pot visualitzar-se.

2.2.4. *Bagging*

El *bagging* o també anomenat agregació *Bootstrap* és una tècnica d'aprenentatge supervisat que ajuda a millorar el rendiment i la precisió dels algoritmes. En termes generals, s'utilitza per tractar el desequilibri biaix - variància, reduir la variància dels models de predicció i minimitzar el sobre ajust de les dades. Es fa servir tant per models de regressió com de classificació.

La idea intuïtiva de l'algoritme consisteix que a partir la creació de mostres bootstrap del conjunt d'entrenament, s'aplica l'algoritme de classificació o regressió a cadascuna d'elles. En el context de la regressió es realitzaran noves prediccions fent la mitjana de les prediccions individuals, mentre que en el context de la classificació les prediccions es realitzaran tenint en compte la classe més freqüent.

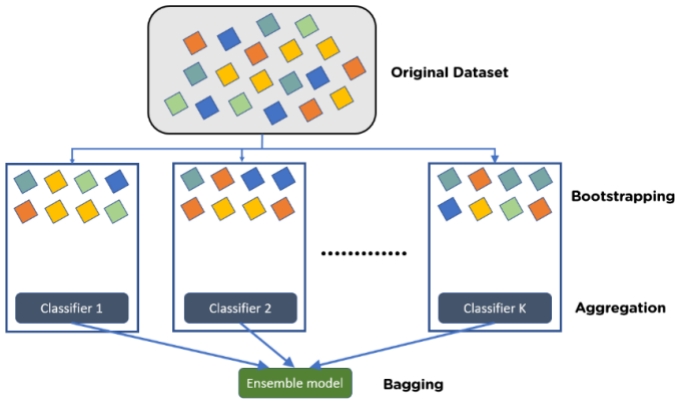


Figura 7. *Algoritme Bagging Font: PLURALSIGHT.*

2.2.5. *Random Forest*

El *Random Forest* forma part del conjunt d'algoritmes d'aprenentatge supervisat i és considerat una variant del procés *bagging*.

L'algoritme es basa a dividir la mostra del conjunt d'entrenament de manera aleatòria, i s'entrenarà el model per cadascun dels subconjunts, obtenint així tants models com subconjunts creats. Finalment, es combinaran tots els resultats dels models proporcionant el resultat final. En els problemes de regressió, el resultat final s'obté fent la mitjana de totes les prediccions mentre que en els problemes de classificació s'aconsegueix per majoria (*Majority-Voting*) [29].

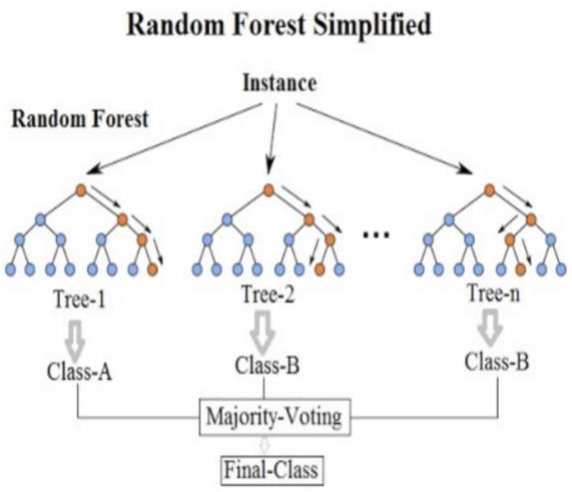


***Figura 8**. Algoritme Random Forest Font: Springboard Blog.*

La principal diferencia entre els algoritmes *Random Forest* i *bagging* recau en el fet que les mostres Bootstrap generades en el *bagging* introdueixen un element d'aleatorietat que provoca que tots els arbres siguin diferents, però de vegades no són suficientment diferents i aquests tenen estructures molt similars. Aquest fenomen rep el nom de correlació entre arbres i es dona en dues ocasions; quan és un model adequat per descriure la relació entre les variables predictores i la variable resposta i, quan una de les variables predictores és especialment rellevant. A conseqüència d'un elevat grau de correlació entre els arbres, el procés *bagging* no aconsegueix reduir prou la variància i, per tant, no millora el model. En canvi, els algoritmes basats en *Random Forest*, solucionen el problema realitzant una selecció de *n* predictors abans d'avaluar cap divisió. D'aquesta forma sent *p* el nombre total de predictors, una mitjana de *(p-m)/p* divisions no tindran en compte el predictor amb major influència, així permetent que la resta de predictors puguin ser seleccionats. Com a resultat, s'aconseguirà desfer la correlació i, per tant, reduir la seva variància.

Els algoritmes basats en *Random Forest* tenen una gran capacitat predictiva, ja que, per una banda, tenen totes els avantatges dels arbres de decisió y per l'altra, redueixen la inestabilitat i correlació entre els arbres.

2.2.6. *Gradient Boosting*

El *Gradient Boosting* és un tipus d'algoritme d'aprenentatge automàtic supervisat que s'utilitza en problemes de classificació i regressió.

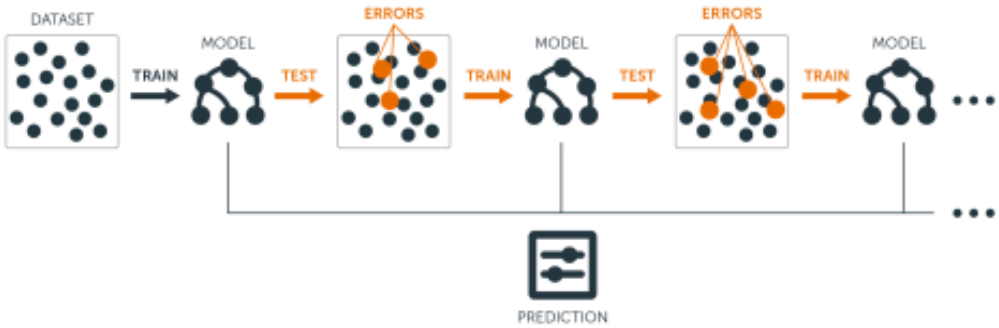


Figura 9. *Algoritme Gradient Boosting. Font: Bradley Boehmke & Brandon, Hands-On Machine Learning with R, 2020.*

L'algoritme *Gradient Boosting* genera un conjunt d'arbres de decisió en seqüència en els que cada arbre aprèn de l'anterior i el millora. La idea principal recau en afegir nous models al conjunt de forma seqüencial, és per això que l'algoritme comença per un model dèbil i a mesura que es van construint nous arbres, cada un d'ells corregeix els errors de l'anterior, augmentant així el seu rendiment.

La contribució seqüencial es basa en el procés d'optimització per descens de gradient i té com a objectiu minimitzar una funció de costos.

3. CAS EMPÍRIC

3.1. Base de dades

3.1.1. Fonts estadístiques

Les fonts d'informació rellevants per dur a terme la pertinent anàlisi han estat principalment dues; **l'Enquesta de Pressupostos Familiars (EPF),** elaborada per l'Institut Nacional d'Estadística (INE) i, el **Model d'Avaluació Ambiental de la Ramaderia Mundial (GLEAM)** dut a terme per l'Organització de les Nacions Unides per l'Alimentació i l'Agricultura (FAO).

Enquesta de Pressupostos Familiars (EPF) (30)

L'Enquesta de Pressupostos Familiars (EPF) realitzada per l'Institut Nacional d'Estadística té com a objectiu principal proporcionar informació sobre la naturalesa i distribució de les despeses de consum de les llars, així com mostrar les diferents característiques de les condicions de vida de les llars espanyoles. Al llarg del temps, l'enquesta ha anat adoptant diverses formes en termes de periodicitat i ha anat evolucionant en diversos aspectes com el tipus de població, la mida mostral, el nivell de desagregació de les despeses, els sistemes de recollida o el disseny dels qüestionaris.

Des de 2006 fins al 2015, la classificació utilitzada per codificar les despeses va ser la COICOP (*Classification of Individual Consumption by Purpose*). A partir del 2016 fins a l'actualitat, la nova codificació s'ha denominat ECOICOP (*European Classification of Individual Consumption by Purpose).* La lleugera diferència entre totes dues recau principalment en els conceptes que s'incloïa en cada parcel·la de despesa. Tant la desagregació de despeses mitjançant el COICOP com l'ECOICOP comprenen dotze grans grups. El treball se centra exclusivament en el primer gran grup corresponent al Grup 1 d'Aliments i Begudes No Alcohòliques, format per diferents subgrups que es troben referenciats a *l'Annex 1*.

Les unitats d'anàlisi de l'EPF són les llars privades residents en els habitatges familiars principals. Per les unitats de mostreig, es pren com a unitat primària la secció censal i com a última unitat, els habitatges familiars principals pertanyents a la secció censal. L'enquesta de pressupostos familiars actua en diversos àmbits d'investigació; en l'àmbit poblacional on la població objectiu són les llars privades i els membres que la conformen, en l'àmbit geogràfic la qual comprèn tot el territori espanyol i, en l'àmbit temporal en què

el període d'estudi és d'un any natural. La mida mostral engloba un total de 2.275 seccions censals (unitats primàries) i de cada una d'elles se seleccionen 10 habitatges (unitats secundàries) dels quals es recull la informació de les llars residents. Per dur a terme l'enquesta s'ha utilitzat un mostreig bietàpic amb estratificació de les unitats de primera etapa (seccions censals en què està dividit el territori nacional en el moment de l'enquesta on s'ha seleccionat una mostra independent dins de cada comunitat autònoma). Finalment, la informació recollida s'ha dut a terme per mètode mixt d'entrevista personal i anotació directa per part de la llar.

Model d'Avaluació Ambiental de la Ramaderia Mundial (GLEAM)

El Model Global d'Avaluació Ambiental de Ramaderia és un entorn del Sistema d'Informació Geogràfica que segueix la metodologia de l'anàlisi del cicle de vida i és capaç de simular activitats i processos de les cadenes de subministrament del sector ramader. El model es va realitzar l'any 2010 i els principals objectius eren quantificar la producció i l'ús de recursos naturals de la ramaderia i, identificar l'impacte mediambiental del sector per tal de promoure iniciatives que pal·liessin la problemàtica [31].

Principalment, va ser dissenyat per avaluar diversos aspectes mediambientals com són el consum de pinsos, les emissions de gasos d'efecte hivernacle, l'ús i la degradació del sòl, el consum d'aigua i nutrients i la interacció amb la biodiversitat.

L'àmbit territorial d'estudi està comprès a escala subnacional, regional i global. Les espècies ramaderes tractades en l'estudi principalment són la carn i llet de boví, búfals, oví i caprí, carn porcina i carn i ous de pollastre. Els gasos d'efecte hivernacle associats a cadascuna de les etapes de producció són el metà (CH_4), el diòxid de carboni (CO_2) i l'òxid nitrós (N_2O). Finalment, el mètode emprat pel càlcul d'emissions d'origen animal (fermentacions entèriques i gestió de fems), és el TIER 2 el qual inclou dades detallades de les característiques del país com són els atributs de la producció ramadera per categories, la Ingesta d'Energia Bruta, els factors de conversió del metà, les taxes específiques d'excreció de Nitrogen i els sistemes de gestió dels fems [32].

3.1.2. *Software* utilitzat

Respecte als recursos informàtics que han estat utilitzats per dur a terme el treball han estat el programa *Rstudio* per tal de tractar, analitzar i visualitzar els resultats, i l'Excel per emmagatzemar les dades. Per modelitzar i predir els diferents models en l'àmbit del ML s'ha fet ús de la llibreria *caret*. També s'han usat les llibreries *fitdistrplus* i *gamlss* per ajustar diferents distribucions i crear models no paramètrics. En últim lloc, totes les figures gràfiques de l'anàlisi han estat elaborades a partir de la llibreria *ggplot2* i, tot i que s'han emprat diverses llibreries, les més usades han estat *tidyr*, *dplyr*, *pROC*, *vcd* i *recipes*.

3.1.3. Estructura i descriptiva inicial de les dades

A partir de les dades proporcionades per les diferents fonts estadístiques s'ha elaborat una base de dades final. És per això que a continuació s'explicarà l'estructura de les dades originals i seguidament, es detallarà l'estructura de la base de dades resultant.

La informació present en l'Enquesta de Pressupostos Familiars es troba en format de microdades mentre que les del Model d'Avaluació Ambiental de Ramaderia (GLEAM) es troben en taules prèviament estructurades.

Enquesta de Pressupostos Familiars [30]

Les Enquestes de Pressupostos Familiars referents al període d'estudi són les dels anys 2006, 2016 i 2020. La informació present en cada una d'elles es troba desglossada en tres fitxers:

- **Fitxer de la llar:** proporciona informació concreta de les llars espanyoles recollides a la mostra on cada llar s'identifica mitjançant la clau primària **NÚMERO**. La informació es troba distribuïda en els següents vuit blocs:
  I. **Informació general:** s'inclouen variables relatives a la comunitat autònoma, la mida del municipi, la densitat de població, les claus de col·laboració... de cada una de les llars.
  II. **Característiques relatives a la llar:** recull el nombre d'individus de cada llar que reuneixen un atribut específic com per exemple, si són majors o menors de 14 o 18 anys, si són membres independents, el nombre de membres actius, ocupats, estudiants... L'enquesta realitzada l'any 2006 recull informació en l'àmbit d'ocupació i activitat així com el nombre de membres

ocupats o actius a la llar. També inclou preguntes de caràcter subjectiu amb la finalitat de conèixer aspectes específics sobre la situació econòmica de la llar. En canvi, les enquestes dels anys 2016 i 2020 la informació es troba recollida en variables segmentades en onze tipologies diferents.

III. **Característiques relatives al sustentador principal:** recullen les mateixes variables que apareixen al fitxer dels membres de la llar, però centrades en el sustentador principal, algunes de les variables recollides són la situació socioeconòmica i professional, el sector de l'activitat en què treballa, el tipus de contracte, entre d'altres.

IV. **Característiques de l'habitatge principal:** s'inclouen variables com són els règims de tinença, qüestions relacionades amb la tipologia de l'edifici, la zona residencial... També es recull informació referent a l'equipament de l'habitatge com la disponibilitat d'aigua calenta i calefacció.

V. **Altres habitatges a disposició de la llar:** recull informació sobre els habitatges de la llar que no són domicili principal. També té en consideració algunes de les mateixes variables recollides per l'habitatge principal com les variables relacionades amb l'equipament de l'habitatge.

VI. **Despeses de consum de la llar:** en aquest bloc es recull la despesa total la qual inclou la despesa monetària i no monetària. A més, recull la despesa monetària, d'autoconsum, d'auto subministrament, del salari en espècie i del lloguer imputat per un motiu diferent del treball. Al fitxer de despeses es pot trobar aquesta mateixa informació però amb un major grau de desagregació.

VII. **Ingressos regulars mensuals de la llar:** les variables incloses en aquest apartat proporcionen informació relacionada amb el nombre de membres de la llar que reben ingressos així com les respectives tipologies de fonts.

VIII. **Nombre de dinars i sopars bisetmanals:** recullen informació del nombre d'àpats dels membres de la llar on s'exclouen el servei domèstic, hostes o convidats.

Les variables seleccionades per l'elaboració de la base de dades final han estat ANOENC, NUMERO, IMPEXAC, DENSIDAD, CCAA i COMITOT.

- **Fitxer dels membres de la llar:** facilita informació específica dels individus que conformen la llar. Cada registre correspon a un sol membre de la llar i s'identifica mitjançant el **NÚMERO** del fitxer de la llar i el **NORDEN** que correspon amb el

número identificador dins de cada llar. Alguns aspectes rellevants d'aquest fitxer són els següents:

- Atès que el servei domèstic, hostes i convidats poden ser membres de les llars, s'ha inclòs la variable CATEGMH, la qual classifica als membres de la llar entre el servei domèstic, hostes, convidats i la resta de membres no pertanyents a cap de les categories esmentades anteriorment.
- La nacionalitat dels membres de la llar queda recollida a la variable NACIONA, però en cas que la nacionalitat del membre sigui estrangera la variable PAÍS indica si l'individu pertany a la Unió Europea, a la resta d'Europa o bé a la resta del món.
- Per cada un dels membres que conforma la llar, s'indica si és perceptor o no d'ingressos, i en cas afirmatiu, s'indica l'import o bé l'interval de l'import net percebut així com l'origen d'aquest.

Les variables seleccionades la creació de la base de dades han estat ANOENC, NUMERO, EDAD, SEXO, NACION, NUMOCU i ESTUDIOS.

- **Fitxer de despeses:** recopila informació de totes les despeses efectuades per les diferents llars de la mostra seleccionada. El nivell de desagregació de les despeses queda recollit en la classificació COICOP/HBS o ECOICOP/HBS (5 dígits). Cada registre de la base de dades, representa la tipologia de despesa que hagi efectuat la llar i, s'identifica a partir de la concatenació de la variable **NÚMERO**, que correspon al número de la llar i la variable **CÓDIGO**, que equival al codi de despesa.

  Les principals variables recollides per cada llar i tipologia de despesa són la variable **GASTO** i la variable **CANTIDAD**. Totes dues es troben elevades a un factor temporal i poblacional.

  Altres variables que queden recollides en un segon pla, però també es troben elevades temporalment i poblacionalment són les variables GASTOMON, GASTONOM1, GASTONOM2, GASTONOM3, GASTONOM4, GASTONOM5 on la primera d'elles representa la despesa monetària i la resta corresponen a la desagregació de les despeses no monetàries com l'autoconsum, el salari en espècie excepte el lloguer, lloguer imputat a l'habitatge en propietat o cedida per raó diferent del treball i el lloguer imputat a l'habitatge cedit per raó de treball.

La necessitat d'aplicar el factor d'elevació temporal és causat per la metodologia de l'enquesta, ja que el període d'estudi (durada al qual van referides les dades de l'enquesta, és d'un any) és diferent del període de referència (durada que son observades les adquisicions de béns i serveis destinats al consum). Per tant, en l'EPF s'apliquen els factors d'elevació temporal per tal d'obtenir una estimació de la despesa o quantitat consumida durant el període d'un any.

D'altra banda, s'utilitzen factors d'elevació espacial per tal d'elevar les dades mostrals a la població, de manera que el factor d'elevació espacial d'una llar de la mostra és el nombre de llars de la població que representa aquesta llar mostral.

Les variables seleccionades han estat ANOENC, NUMERO, CODIGO i GASTO. Les variables de cada un dels fitxers es troben referenciades a *l'Annex 2.*

Els valors absents o *missings* poden ser un problema a l'hora de generar la base de dades resultant i la seva anàlisi posterior. Si la quantitat de valors *missings* és molt petita amb relació a la mida del conjunt de dades, s'eliminaran aquells valors absents. Però pel contrari, si la quantitat de valors absents és elevada, serà necessari imputar-los mitjançant l'ús de la funció *mice*. Aquesta funció utilitza un algoritme de tal manera que identifica el tipus de variable i a partir de la informació de la resta de variables imputa els valors absents.

En el cas del fitxer de la llar, referent al 2006, 2016 i 2020 presenten 2, 1 i 15 valors *missings* respectivament. A causa de la gran dimensió del conjunt de dades, aquestes seran eliminades. Amb relació al fitxer dels membres de la llar, aquest presenta 9.277, 9.882 i 8.009 *missings* respectivament i, per tant, aquests valors s'imputaran emprant la funció explicada anteriorment. Finalment, el fitxer de despeses no presenta cap valor *missing* i no serà necessari tractar-los.

Model d'Avaluació Ambiental de la Ramaderia Mundial (GLEAM)

La descripció es basa en tres aspectes claus; en primer lloc, es troba la classificació dels sistemes de producció segons les diferents condicions agroecològiques, el ús dels pinsos i el farratge. En segon lloc, té en compte l'Anàlisi del Cicle de Vida la qual proporciona una visió completa dels processos de producció, per la qual cosa permet identificar les

etapes més importants on es generen els majors impactes mediambientals. Finalment, el model identifica les principals fonts d'emissions relacionades amb la producció, el processament i el transport de pinsos del sector ramader. Les fonts identificades en el model són la fermentació entèrica, la gestió dels fems i el consum d'energia.

La principal variable a estudiar és EMISION INTENSITY que correspon amb els kilograms de CO_2-eq emesos per cada kg de proteïna (kg CO_2-eq · kg proteïna^{-1}) de l'animal. La variable es calcula de la següent manera:

$$\boldsymbol{EMISION\ INTENSITY}(kg\ CO_2 - eq\ \cdot\ kg\ protein^{-1}) = \frac{Total\ GHG\ emissions\ (kg\ CO_2 - eq)}{Production\ (kg\ protein)}$$

- $\boldsymbol{Total\ GHG\ emissions} = Total\ CO_2 emissions + Total\ CH_4\ emissions + Total\ N_2O\ emissions$
  - $\boldsymbol{Total\ CO_2 emissions} = Feed, CO_2 + LUC: soy\ \&\ palm, CO_2 + Direct\ energy, CO_2 + Indirect\ Energy, CO_2 + Postfarm, CO_2$
  - $\boldsymbol{Total\ CH_4\ emissions} = Enteric\ fermentation, CH_4 + Manure\ management, CH_4$
  - $\boldsymbol{Total\ N_2O\ emissions} = Feed: fertilizer\ \&\ crop\ residues, N_2O + Feed: applied\ \&\ deposited\ manure, N_2O + Manure\ management, N_2O$

Base de dades resultant

A partir de l'Enquesta de Pressupostos Familiars, el Model d'Avaluació Ambiental de la Ramaderia Mundial i la taula proporcionada per l'article d'Arjen Y. Hoekstra, [11] s'elaboraran dues bases de dades finals:

A) Base de dades I. Consum de la llar: Mitjançant l'Enquesta de Pressupostos Familiars (EPF) es crearan diverses variables auxiliars amb l'objectiu d'analitzar la distribució de les despeses segons el tipus d'alimentació de les diferents llars espanyoles.

B) Base de dades II. Costos Ocults de la carn: A partir de la informació proporcionada en el Model GLEAM i l'article Arjen Y. Hoekstra [11], juntament amb el valor del cost de l'Aigua a Espanya, així com el cost per tona de CO_2, es crearà una base de dades que tindrà com a finalitat quantificar el cost ocult segons el tipus de carn.

A continuació, s'explicarà cada una de les variables i com ha estat calculada respectivament:

Base de dades I. Consum de la llar

A partir de les variables dels fitxers referents a l'Enquesta de Pressupostos Familiars i els respectius anys, s'han seleccionat i calculat una sèrie de variables auxiliars obtenint així la base de dades amb la que es treballarà.

Variables identificadores

- **ANOENC:** Any de l'enquesta corresponent. (2006; 2016; 2020)
- **NUMERO:** Número seqüencial d'ordre de la llar. (00001-25000)
- **CODIGO:** Codi de despesa. (COICOP/ ECOICOP)

Variables econòmiques

- **GASTOT:** Despesa monetària i no monetària total en euros elevada temporal i poblacionalment.
- **IMPEXAC:** Import exacte en euros dels ingressos mensuals nets totals de la llar.

Variables sociodemogràfiques

- **SEXO:** Percentatge de dones a la llar. (0-100)
- **EDAD:** Edat mitjana dels membres de la llar en anys. (0-85)
- **NACION:** Percentatge de membres amb nacionalitat espanyola. (0-100)
- **DENSITAT:** Densitat de la població. (1: Zona densament poblada; 2: Zona intermèdia; 3: Zona disseminada).
- **CCAA:** Comunitat autònoma segons PIB per càpita. (1: CCAA amb PIB per càpita superior a 25.000 euros; 2: CCAA amb PIB per càpita entre 25.000 i 20.000 euros; 3: CCAA amb PIB per càpita inferior a 20.000 euros).
- **COMITOT:** Número de Dinars i Sopars efectuats durant la bisetmana mostral. (0-999)
- **NUMOCUP:** Número d'ocupats a la llar. (0: Cap ocupat; 1: Un ocupat; 2: Dos ocupats; 3: Tres ocupats; 4: Quatre o més ocupats).
- **ECIVILSP:** Estat Civil del Sustentador Principal. (1: Solter; 2: Casat; 3: Vidu; 4: Separat o Divorciat).
- **ZONARES:** Zona Residencial. (1: Urbana; 2: Rural).
- **NMIEMBR1:** Percentatge de membres de la llar entre 0 i 15 anys. (0-100)

- **NMIEMBR2:** Percentatge de membres de la llar entre 16 i 64 anys. (0-100)
- **NMIEMBR3:** Percentatge de membres de la llar de 65 anys o més. (0-100)
- **ESTUDIOS1:** Percentatge de membres de la llar sense Estudis Primaris. (0-100)
- **ESTUDIOS2:** Percentatge de membres de la llar amb Estudis Obligatoris. (0-100)
- **ESTUDIOS3:** Percentatge de membres de la llar amb Estudis Superiors. (0-100)

Els càlculs de les variables es troben mencionats a *l'Annex 3.*

Base de dades I. Costos Ocults de la carn.

Referent a la petjada hídrica, l'article *The hidden water resource use behind meat and dairy de Arjen Y. Hoekstra* posa de manifest la mitjana global (L/kg) de la petjada hídrica segons la tipologia d'espècie, per tant, només serà necessari fer la conversió de L/kg a m^3/kg i d'aquesta manera obtindrem els metres cúbics consumits per cada kg de carn. A partir de la dada proporcionada per l'Institut Nacional d'Estadística sobre el cost mitjà d'aigua, que pren valor 1,91 €/m^3, obtindrem el cost ocult de l'aigua en €/kg segons el tipus d'espècie animal:

$$\boldsymbol{Cost\ Ocult\ Aigua}\ (\boldsymbol{kg}/€) = Petjada\ hídrica\ (\mathrm{m}^3/kg) * \text{Preu Aigua}\ (€)$$

Respecte a la petjada de carboni, a partir del model GLEAM, el qual estima la intensitat d'emissions de gasos d'efecte hivernacle (kg CO_2-eq·proteina^{-1}) segons les especies animals, la taula de l'article *Arjen Y. Hoekstra,* el qual mostra el contingut nutricional, (proteïna (g/kg)) i l'empresa SENCO2 que mostra el preu mitjà anual (€/Tn) del CO_2-eq segons els anys estudiats (preu mitjà anual l'any 2006 és de 17,64 €/Tn, al 2016 és de 5,35 €/Tn i al 2020 de 24,75 €/Tn). Fàcilment, serà possible calcular el cost ocult referent a l'impacte atmosfèric com:

$$\boldsymbol{Cost\ Ocult\ Atmosfèric}\ (\boldsymbol{kg}/€) = \frac{EMISION\ INTENSITY\ (\text{kg CO2} - \text{eq} \cdot \text{kg proteina}^{-1}) * Proteïna\ (\text{g/kg})}{1000} * Preu\ CO_2(€)$$

Per tant, el Cost Ocult Total en euros per cada kilogram de carn es calcularà de la següent manera:

$$\boldsymbol{PREU\_KG\_MA} = Cost\ Ocult\ Aigua\ (kg/€) + Cost\ Ocult\ Atmosfèric\ (kg/€)$$

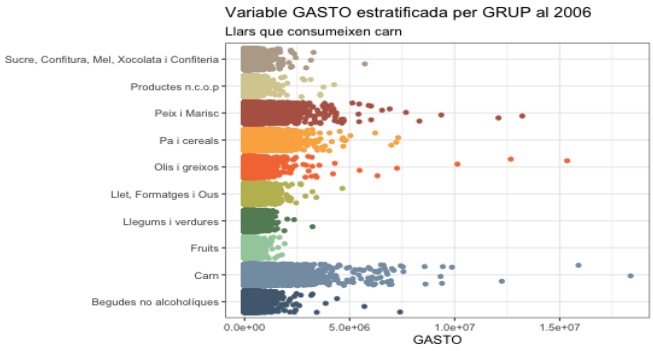
Per altra banda, una variable que haurem de tenir en compte per l'estudi serà el preu de mercat de cada kg de carn. Aquesta informació la trobarem al Ministeri d'Agricultura, Pesca i Alimentació.

A partir dels càlculs realitzats, la Base de Dades II resultant estarà formada per les següents variables:

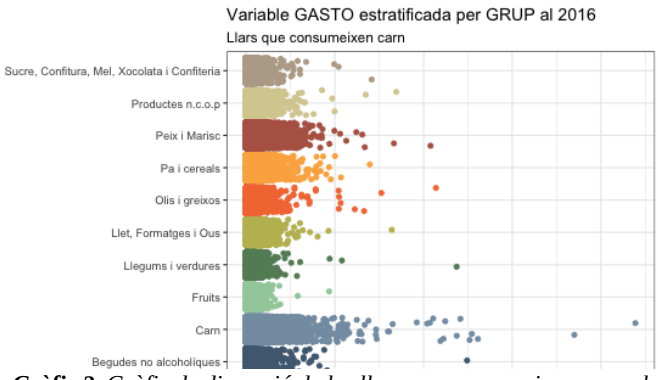
- **ANY:** Any de referència (2006; 2016; 2020)
- **SUBGRUP:** tipologia de carn (Boví; Porcí; Oví i Caprí; Au)
- **PREU_KG_MA:** Preu Ocult/Mediambiental en euros per cada kilogram de carn
- **PREU_KG_CARN:** Preu de mercat en euros per cada kilogram de carn**.**

Conèixer com es comporten i distribueixen les dades és essencial per tenir una idea general de l'àmbit en el qual es treballa. És primordial conèixer com es distribueix el consum entre les llars espanyoles durant els diferents anys corresponents. Per això, a continuació s'estudiarà com es distribueix la variable continua GASTO segons la variable categòrica GRUP i s'agruparan en si les llars consumeixen o no carn.

De manera general, en els ***Gràfics 1-3,*** s'observen que el comportament de les llars que consumeixen carn són molt similars entre els diferents anys, on l'ordre dels grups de més a menys freqüent és, el grup de Carn, Peix i Marisc, Pa i cereals, Olis i greixos, Begudes no alcohòliques, Sucre, Confitures, Mel, Xocolata i Confiteria, Productes n.c.o.p, Llet, formatge i ous, Llegums i verdures i Fruits. Referent a l'any 2006, els grups d'aliments més consumits per les llars són el grup del Peix i Marisc, seguit del Pa i Cereals i la Carn, sent aquest últim al qual destinen una

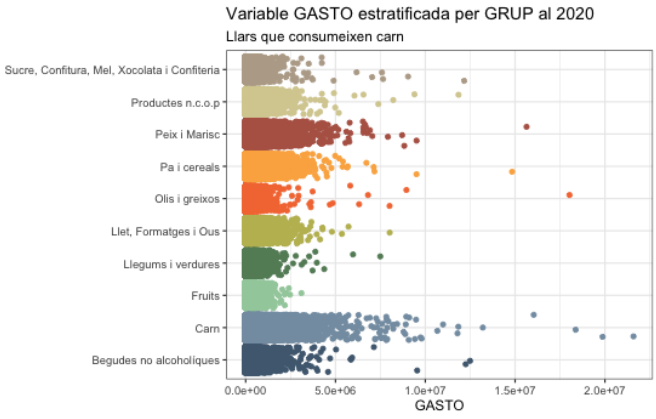


Gràfic 1*. Gràfic de dispersió de les llars que consumeixen carn al 2006.*



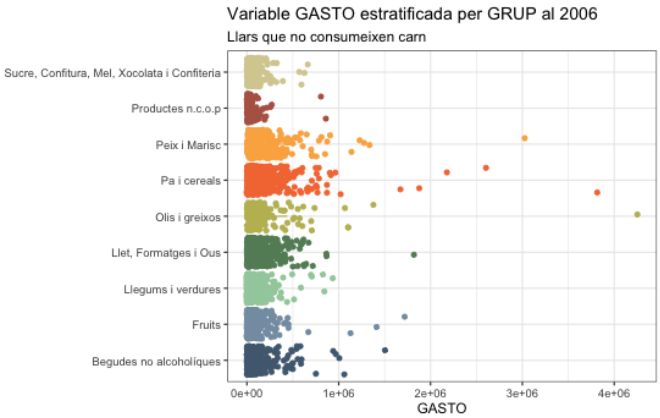
Gràfic 2*. Gràfic de dispersió de les llars que consumeixen carn al 2016.*

major despesa. I els grups menys consumits per les llars són els grups de Llegums i verdures i Fruits. Referent al 2016 i al 2020, segueixen el mateix ordre que l'any 2006, però si es comparen els respectius anys, s'observa que la despesa realitzada per les llars durant el transcurs del temps en el grup de Llegums i Verdures i Fruits, va augmentant mentre que la despesa de carn roman aproximadament constant.

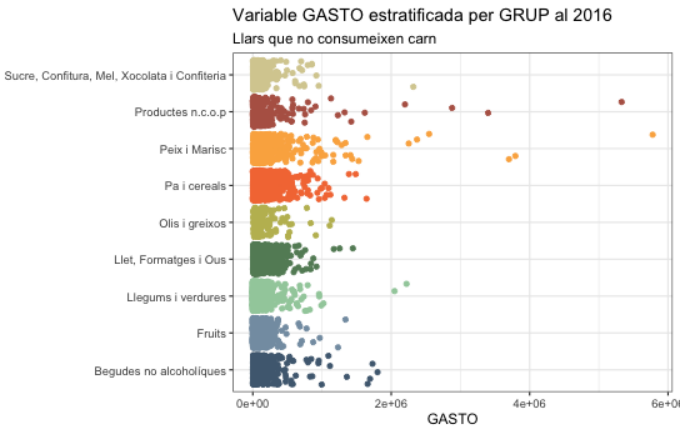


***Gràfic 3**. Gràfic de dispersió de les llars que consumeixen carn al 2020.*

Respecte als gràfics anteriors, els ***Gràfics 4-6*** no contemplen el grup de Carn. Gràficament, el grup de Pa i cereals, Peix i Marisc, Llet, Formatge i ous, Begudes no alcohòliques, Llegums i verdures, Fruits, Olis i greixos, Sucre, Confitures, Mel, Xocolata i Confiteria i Productes n.c.o.p corresponen a l'ordre dels grups de més a menys freqüent per aquelles llars que no consumeixen carn. Referent a l'any 2006, el Pa i cereals i el Peix i marisc, són els grups més consumits per les llars espanyoles. En canvi, el grup Productes n.c.o.p i Sucre, confitura, Mel, Xocolata i Confiteria, corresponen als grups menys consumits. Respecte a l'any 2016, l'ordre dels grups més consumits és el mateix que el de l'any 2006,

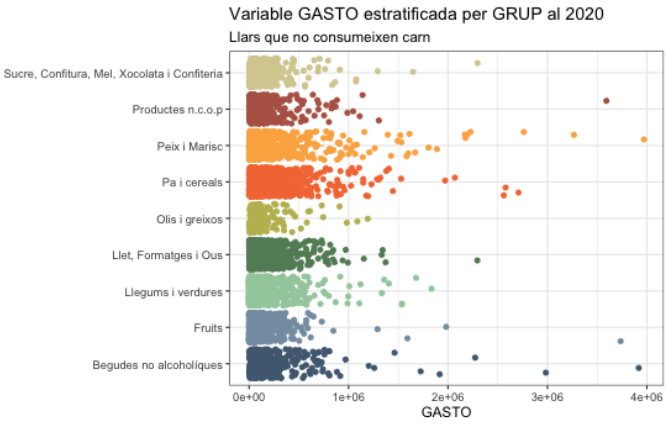


***Gràfic 4.** Gràfic de dispersió de les llars que no consumeixen carn al 2006.*



***Gràfic 5.** Gràfic de dispersió de les llars que no consumeixen carn al 2016.*

mentre que el de menys consumits canvia, sent l'ordre, el grup dels Olis i Greixos i Sucre, Confitura, Mel, Xocolata, i Confiteria. En canvi, l'any 2020 els grups menys consumits per les llars són els Olis i Greixos i els Fruits. Si es comparen els respectius anys, s'observa que el nombre de llars que no consumeixen carn ha anat augmentant durant els anys i, també han anat canviant els patrons de consum d'aquestes llars.



Gràfic 6. *Gràfic de dispersió de les llars que no consumeixen carn al 2020.*

Com a resultat, podem concloure que no hi ha diferències rellevants respecte als diferents anys i, per tant, s'agruparan totes les llars sense tenir en compte la variable temporal.

3.2. Problema de classificació: modelització sobre el consum de productes carnis de les llars espanyoles

Els problemes de classificació es caracteritzen per tenir una variable qualitativa o categòrica com a resposta. El fet de realitzar prediccions sobre una variable qualitativa d'una observació, se'n diu classificar aquesta observació, és a dir, predir la seva categoria o classe. Normalment, els mètodes capaços de classificar, el que fan és predir la probabilitat d'una observació de pertànyer a cadascuna de les categories.

En aquest apartat s'abordaran dos objectius: per una banda, determinar quins són els factors explicatius que fan que les llars espanyoles consumeixin o no carn i quin és l'efecte d'aquests sobre la variable resposta. Per altra banda, s'avaluarà la capacitat predictiva dels diferents models i es duran a terme prediccions sobre aquests.

Les dades utilitzades per dur a terme l'anàlisi, contenen la informació relativa a la Base de Dades I, especificada anteriorment, la qual està formada per un conjunt de 19 variables i un total 52.871 llars que van participar en l'Enquesta de Pressupostos Familiars del 2006, 2016 i 2020, on les llars són diferents per cada any en qüestió.

A continuació, cal especificar quina és la variable resposta i quines són les explicatives o predictores del problema de classificació. Pel que fa a la **variable resposta**, és a dir, la variable d'interès es definirà com:

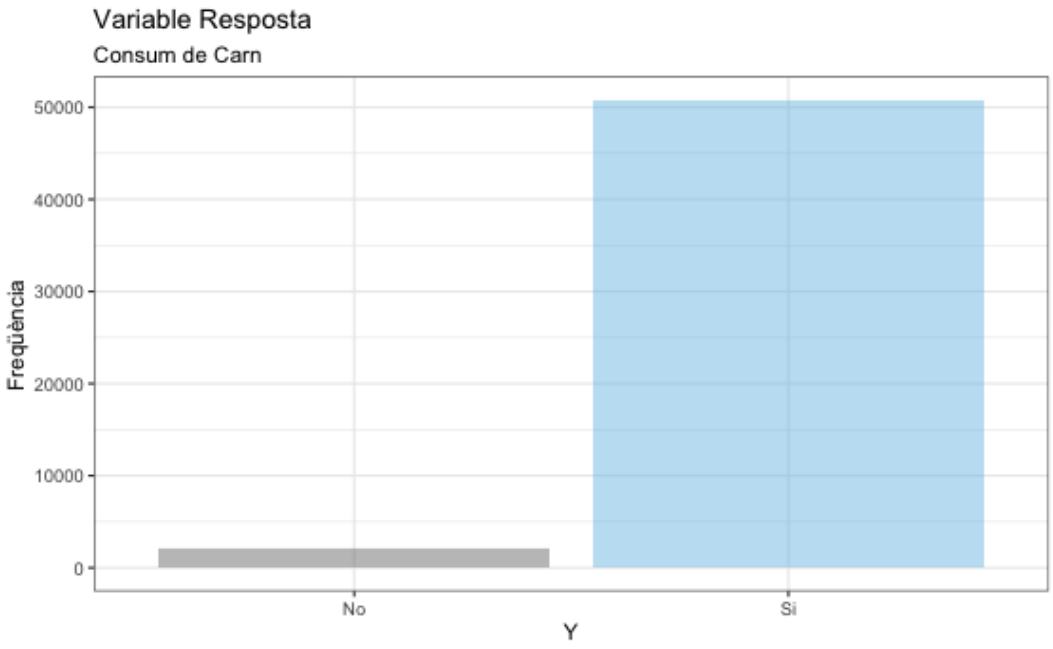
$$Y = \begin{cases} \mathbf{1} & \boldsymbol{Si\ la\ llar\ consumeix\ carn} \\ \mathbf{0} & \boldsymbol{Si\ la\ llar\ no\ consumeix\ carn} \end{cases}$$

Les **variables explicatives** seran; GASTOT, SEXO, EDAD, NACION, DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARES, NMIEMBR1, NMIEMBR2, NMIMEBR3, ESTUDIOS1, ESTUDIOS2 i ESTUDIOS3. Cal especificar que després de realitzar la normalització de la base de dades les variables ANOENC, GASTO i CODIGO han sigut excloses de l'anàlisi.

3.2.1. Anàlisi exploratòria de les dades

Abans de realitzar qualsevol càlcul amb un nou conjunt de dades o entrenar un model predictiu, és essencial fer-ne una exploració descriptiva de les variables. Aquest procés permet entendre millor la informació continguda en cadascuna de les variables, així com detectar possibles anomalies.

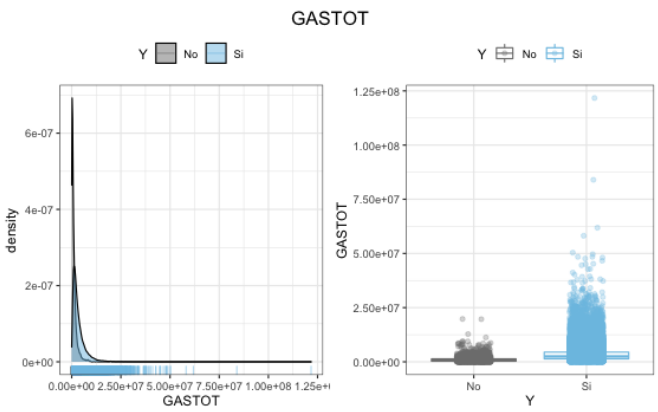
Quan es crea un model, és molt important estudiar la distribució de la variable resposta, ja que, al cap i a la fi, és la variable que es vol explicar i predir.



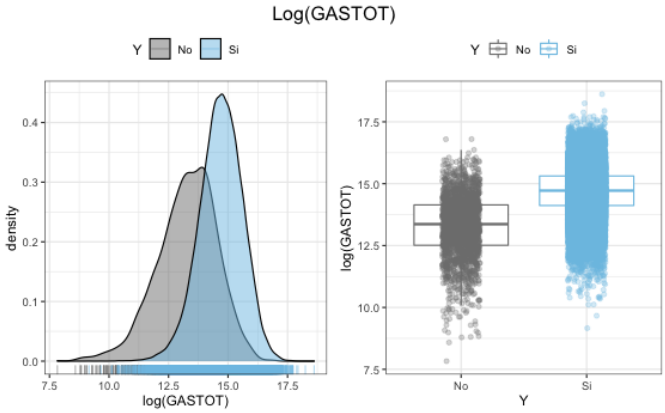
Gràfic 7. *Diagrama de barres de la variable resposta.*

En termes numèrics, del total de la mostra hi ha 50.798 llars que consumeixen carn i 2.073 que no. En el ***Gràfic 7*** es mostra un diagrama de barres de la variable resposta el qual posa de manifest que clarament les classes no es troben balancejades. És a dir, el nombre de registres per a una de les classes és inferior a la resta. Quan el desequilibri és petit, no suposa cap problemàtica, però quan el desequilibri és gran, sí. Tal com es mostra en aquest cas, el 4% de les llars no consumeixen carn mentre que el 96% sí. Per tant, estem davant d'un problema de desequilibri de classes (*Class Imbalance Problem*). En el nostre cas, la solució a la qüestió, es realitzarà mitjançant el sobre mostreig i el submostreig.

Com que l'objectiu de l'estudi és predir quines llars consumeixen carn i quines no, l'anàlisi de cada variable predictora es farà segons el valor que prengui la variable resposta. Emprant aquesta estratègia, es poden començar a conèixer quines variables estan més relacionades amb la variable resposta, és a dir, amb el fet de consumir o no carn.

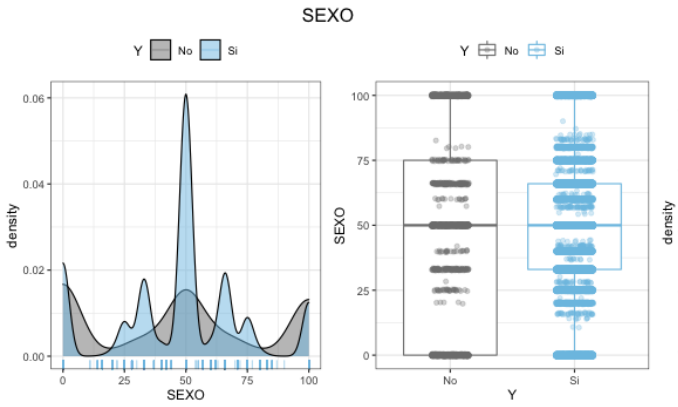


Gràfic 8*. Corba de densitat i boxplot de la variable GASTOT.*

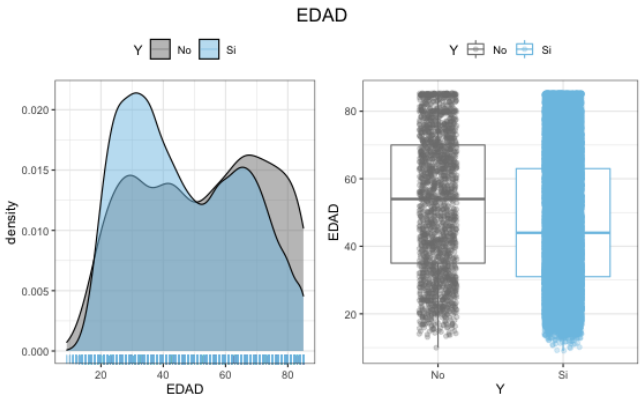


Gràfic 9. *Corba de densitat i boxplot de la variable log(GASTOT).*

El ***Gràfic 8*** mostra una distribució molt asimètrica de la variable GASTOT, on moltes llars tenen una despesa baixa i molt poques, tenen una despesa molt alta. Aquest tipus de distribució se sol visualitzar millor després de realitzar una transformació logarítmica, tal com es mostra en el ***Gràfic 9***, on les dades indiquen que la despesa mitjana de les llars que van consumir carn és superior a les llars que no van consumir-ne.



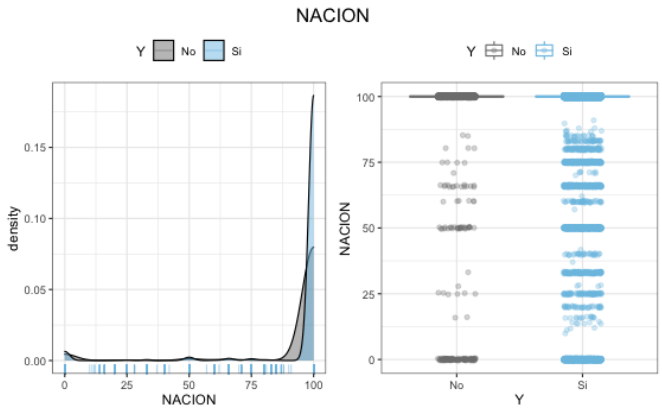
Gràfic 10. *Corba de densitat i boxplot de la variable SEXO.*



Gràfic 11*. Corba de densitat i boxplot de la variable EDAD.*

En el ***Gràfic 10*** es pot observar que gran part de les llars consumidores de carn estan compostes principalment per un 50% de dones i la resta de llars, entre el 25%-45% i el 65%-78% de dones. Per altra banda, aquelles que no consumeixen carn, el percentatge de dones es distribueix entre el 0%, 50% i el 100% de dones.

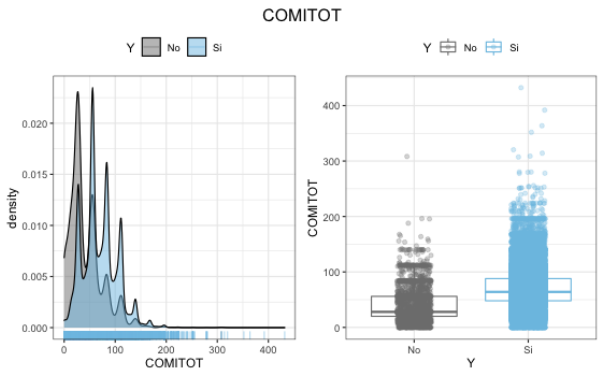
El ***Gràfic 11*** determina que l'edat mitjana de les llars sembla molt similar entre els dos grups, amb dues excepcions: en el rang d'edat comprés entre els 19 i 40 anys, el percentatge de consumir carn és molt més elevat. Mentre que, a l'extrem oposat, a partir dels seixanta anys, la probabilitat de no consumir carn és més elevada. Quan la informació d'una variable contínua resideix en si se superen determinats límits, els models predictius solen obtenir millors resultats si la variable es categoritza en intervals. Aquesta s'analitzarà juntament amb la resta de variables qualitatives.



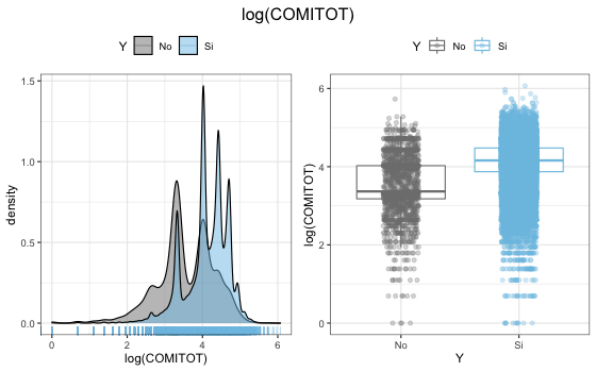
Gràfic 12. *Corba de densitat i boxplot de la variable NACION.*

En el ***Gràfic 12*** s'evidencia que la variable NACION té una distribució molt asimètrica, independentment de si les llars consumeixen carn com si no. A més, quasi totes les llars estan compostes pel 100% d'individus amb nacionalitat espanyola i, com és d'esperar, molt poques llars estaran formades únicament per individus amb nacionalitat estrangera. La raó per la qual succeeix aquest fenomen és perquè l'enquesta s'ha realitzat dins del territori espanyol i

en cas que tinguin doble nacionalitat es consideraran dins de la categoria de nacionalitat espanyola. Per tant, majoritàriament, quasi totes les llars tindran nacionalitat espanyola.

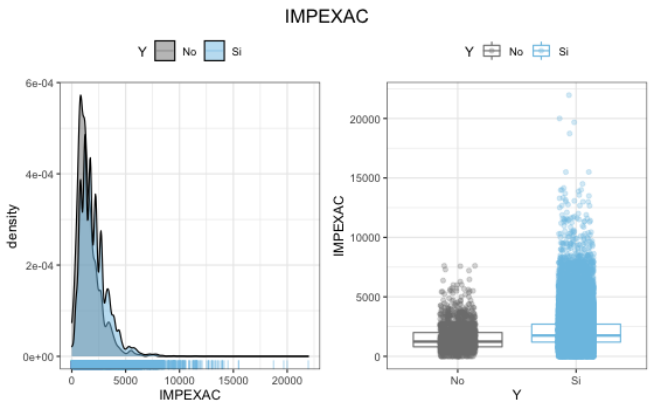


Gràfic 13. *Corba de densitat i boxplot de la variable COMITOT.*

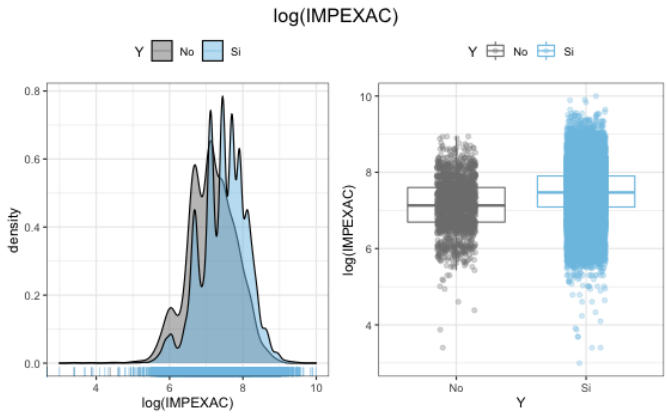


Gràfic 14. *Corba de densitat i boxplot de la variable log(COMITOT).*

El ***Gràfic 13*** mostra que de manera genèrica, el nombre de dinars i sopars bisetmanals de gran part de les llars es troba comprès entre 0 i 200, mentre que molt poques llars prenen valors molt elevats de la variable. Aquelles llars que consumeixen carn tenen un major nombre de dinars i sopars que no pas els que no mengen carn. En presentar l'asimetria, és recomanable visualitzar la variable realitzant una transformació logarítmica. El ***Gràfic 14*** permet observar que la distribució sembla ser molt similar entre els dos grups, però quan el valor de la variable oscil·la entre 1 i 54, la probabilitat de no consumir carn és major, mentre que, quan el valor de la variable és superior a 54, la probabilitat de consumir carn és superior.



Gràfic 15. *Corba de densitat i boxplot de la variable IMPEXAC.*

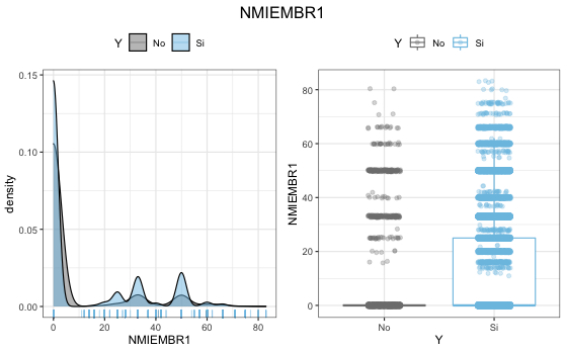


Gràfic 16. *Corba de densitat i boxplot de la variable log(IMPEXAC).*

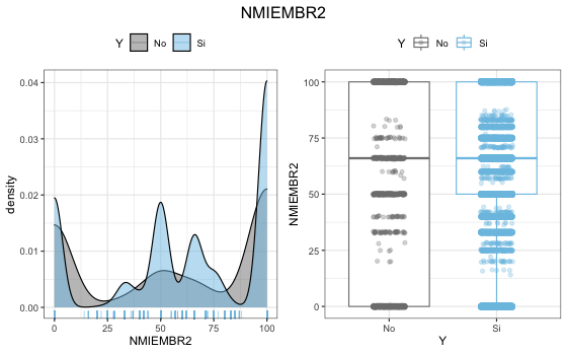
Referent a la variable IMPEXAC, tal com s'observa al ***Gràfic 15***, gran part de les llars reben uns ingressos mensuals nets entre 0 € i 5.000 € mentre que, poques llars reben salaris molt elevats, provocant una distribució asimètrica. És per això, que s'aplicarà una transformació logarítmica de la variable amb l'objectiu de facilitar la seva interpretació.

La distribució de les llars del ***Gràfic 16*** és molt semblant entre totes dues classes. Es mostra que aquelles llars que consumeixen carn tenen uns ingressos lleugerament superiors als que no consumeixen, és a dir, les dades indiquen que per salaris més baixos, la probabilitat de no consumir carn és més elevada que consumir-ne. En canvi, per ingressos més elevats, la probabilitat de consumir carn és més elevada. Com que la informació de la variable resideix en si superen determinats valors, serà aconsellable discretitzar-la en intervals i analitzar-la com a variable categòrica.

El ***Gràfic 17*** recull informació sobre la variable NMIEMBR on es mostren distribucions molt similars entre els dos grups. S'observa que gran part de les llars, tant si consumeixen carn com si no, no estan compostes per llars amb individus entre 0 i 15 anys, o bé aquest percentatge és imperceptible. A més, hi ha un petit col·lectiu que està constituït pel 35% i el 50% d'individus de la llar entre 0 i 15 anys. Aquest fet es podria atribuir a aquelles llars plurifamiliars les quals estan constituïdes per fills adolescents. Concretament, es pot veure que la probabilitat d'aquelles llars de consumir carn és superior que no consumir-ne.

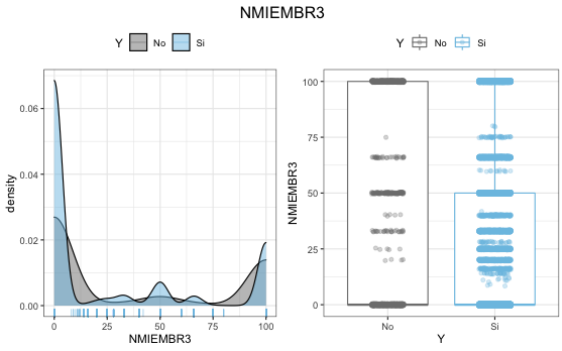


Gràfic 17. *Corba de densitat i boxplot de la variable NMIEMBR1.*



Gràfic 18*. Corba de densitat i boxplot de la variable NMIEMBR2.*

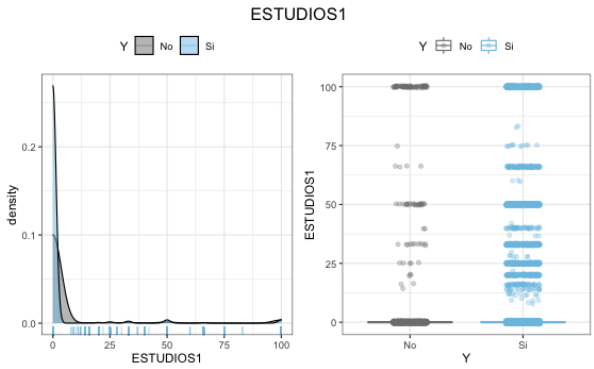
El ***Gràfic 18*** mostra que gran part de les llars que consumeixen carn estan totalment compostes per individus entre 16 i 64 anys i, l'altre percentatge restant es troba distribuït entre el 0% i el 25%-75%. D'altra banda, aquelles llars que no consumeixen carn, el 5%-25%, 80%-95% i el 100% dels individus de les llars tenen



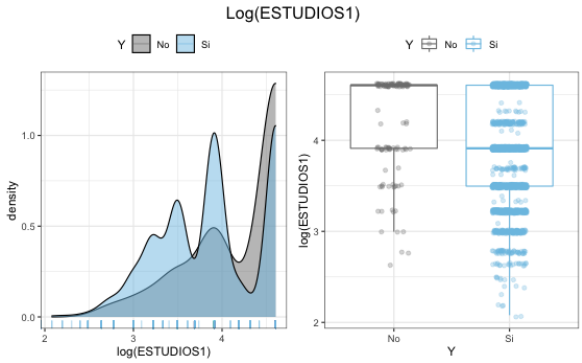
Gràfic 19*. Corba de densitat i boxplot de la variable NMIEMBR3.*

entre 16 i 64 anys. Tal com s'ha explicat anteriorment, quan la informació d'una variable contínua depèn de si se superen determinats límits, és recomanable categoritzar la variable en intervals.

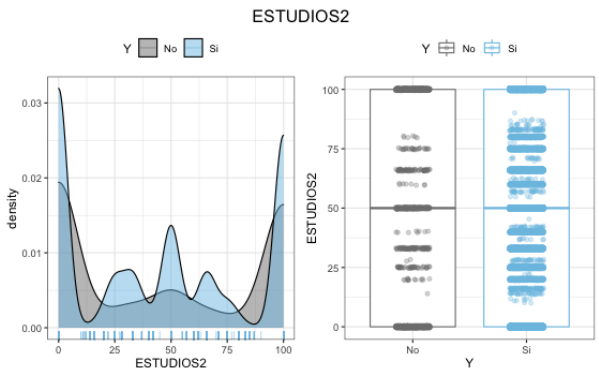
Respecte al ***Gràfic 19***, s'observa que el percentatge de llars on no hi ha cap membre major de seixanta-quatre anys, la probabilitat de consumir carn és més elevada, però si la llar es troba formada pel 10%-15% i 75%-98% dels membres de la llar majors de seixanta-quatre anys, la probabilitat de no consumir carn és superior. Per tant, en aquest cas també és aconsellable discretitzar la variable i tractar-la com a variable categòrica.



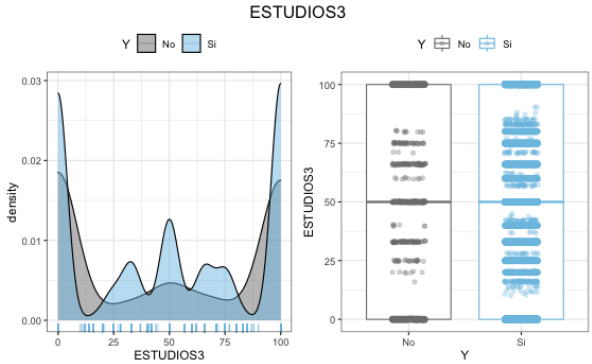
Gràfic 20. *Corba de densitat i boxplot de la variable ESTUDIOS1.*



Gràfic 21. *Corba de densitat i boxplot de la variable log(ESTUDIOS1).*



Gràfic 22. *Corba de densitat i boxplot de la variable ESTUDIOS2.*



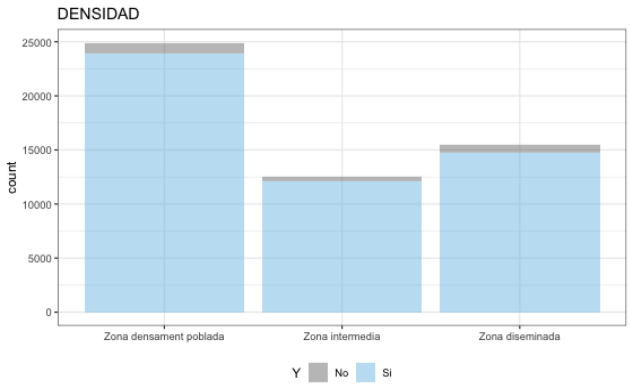
Gràfic 23. *Corba de densitat i boxplot de la variable ESTUDIOS3.*

Referent al ***Gràfic 20***, la variable ESTUDIOS1 presenta una distribució asimètrica on majoritàriament tots els individus de la llar tenen estudis primaris o superiors i, unes poques llars no tenen estudis. Aquest últim fet podria atribuir-se a aquelles llars compostes per individus amb una edat avançada i que no han anat a l'escola. Aquest tipus de distribució se sol visualitzar millor després d'aplicar una transformació logarítmica, tal com es mostra en el ***Gràfic 21***, la qual indica que si el percentatge dels individus de la llar es troba compres entre el 3% i 54%, la probabilitat de consumir carn és superior a no

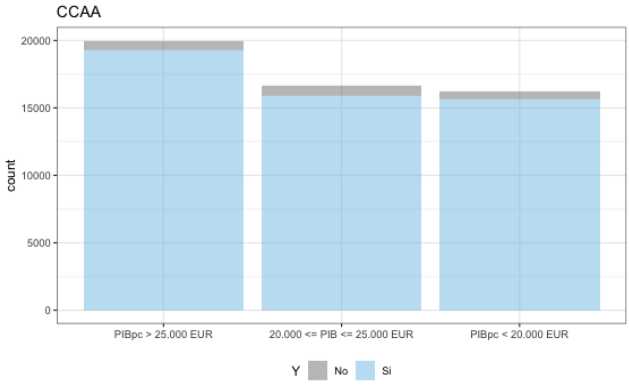
consumir-ne, però això canvia quan el valor de la variable pren valors superiors al 54%, on la probabilitat de no consumir carn és superior.

En el ***Gràfic 22*** s'observa que les llars constituïdes pel 0%, 25%-75% i el 100% dels individus de cada llar, tenen estudis primaris i la probabilitat de consumir carn és superior a no consumir-ne. D'altra banda, per llars formades entre el 5%-25% i el 80%-95% d'individus amb estudis primaris, la probabilitat de no consumir carn és més elevada que sí consumir-ne. Com la informació d'una variable radica en si se superen determinats valors, serà adequat categoritzar la variable contínua en intervals.

Al ***Gràfic 23,*** es mostra que les llars constituïdes pel 0%, 25%-75% i el 100% dels membres de la llar tenen estudis superiors i la probabilitat de consumir carn és més elevada que no consumir-ne. En canvi, les llars compostes pel 5%-25% i el 80%-95% dels individus de la llar, la probabilitat de no consumir carn és major que consumir-ne. En aquest cas també serà recomanable discretitzar la variable en intervals i analitzar-la com a categòrica.



Gràfic 24. *Diagrama de barres de la variable DENSIDAD.*

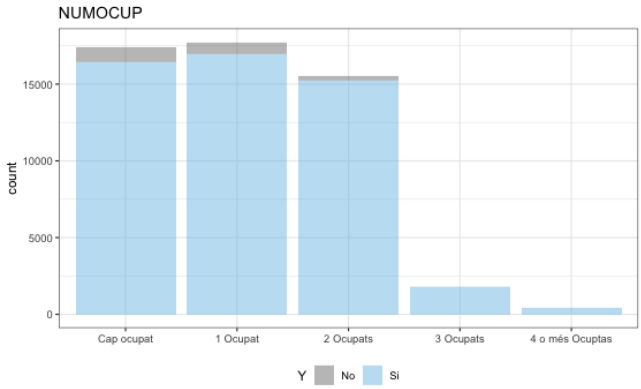


Gràfic 25*. Diagrama de barres de la variable CCAA.*

El ***Gràfic 24*** mostra que gran part de les llars, independentment si consumeixen o no carn, viuen en zones densament poblades, seguides de zones disseminades i finalment de zones intermèdies. Per altra banda, en el ***Gràfic 25***, s'observa que aquelles llars que consumeixen carn, gran part d'elles viuen en comunitats autònomes amb un PIB per càpita superior als 25.000 € i, la resta es troben distribuïdes de forma igualitària entre, comunitats autònomes amb un PIB per càpita entre 20.000 € i 25.000 € i, un PIB per càpita inferior a 20.000 €. Referent a les llars que no consumeixen carn, aquestes es reparteixen de manera igualitària entre els tres tipus de comunitats autònomes, és a dir, el

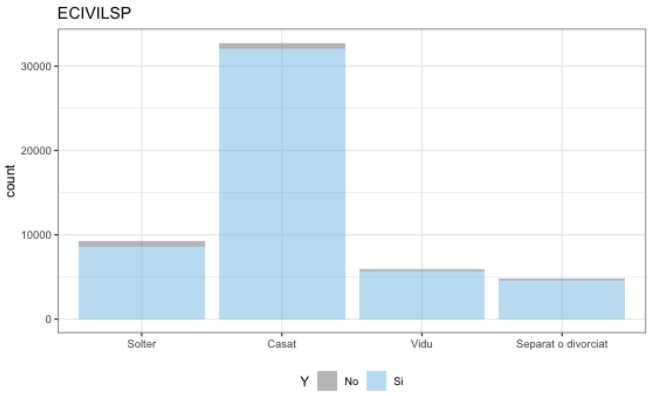
nombre de llars que no consumeixen carn serà el mateix independentment al tipus de comunitats autònomes al qual pertanyin.

El ***Gràfic 26*** indica que, la majoria de llars que no consumeixen carn no hi ha cap ocupat o només hi ha 1 ocupat i, a les llars on només hi ha 2 ocupats, el nombre de llars que no consumeix carn és menor. Finalment, aquelles compostes per 3 ocupats, 4 o més ocupats, no hi ha cap llar que no consumeixi carn. Referent al nombre de llars que sí que consumeixen carn, en la majoria de casos, no hi ha cap ocupat, o bé hi ha un o dos ocupats.



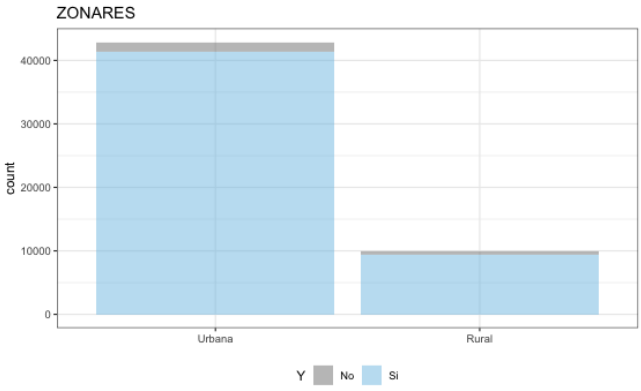
Gràfic 26. *Diagrama de barres de la variable NUMOCUP.*

En el ***Gràfic 27***, s'observa que gran part de les llars que consumeixen carn, el sustentador principal es troba casat. Mentre que, aquelles llars que no consumeixen carn, es troba solter o casat.



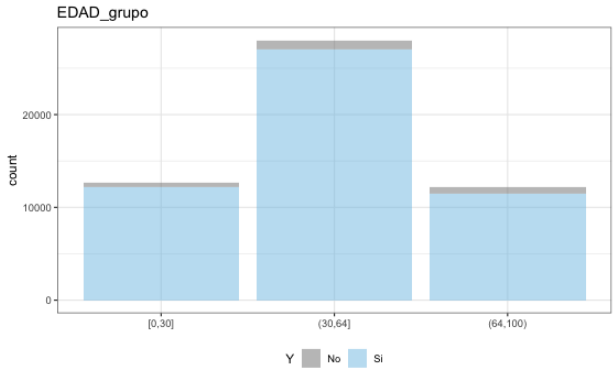
Gràfic 27*. Diagrama de barres de la variable ECIVILSP.*

Pel que fa al ***Gràfic 28***, tant les llars consumidores de carn com les que no, viuen en zona una residencial urbana.

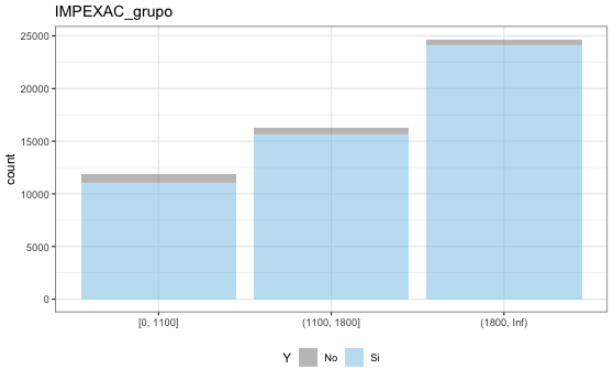


Gràfic 28*. Diagrama de barres de la variable ZONARES.*

A continuació, s'interpretaran les variables que anteriorment eren contínues, però s'han discretitzat en intervals, convertint-se així en variables categòriques.

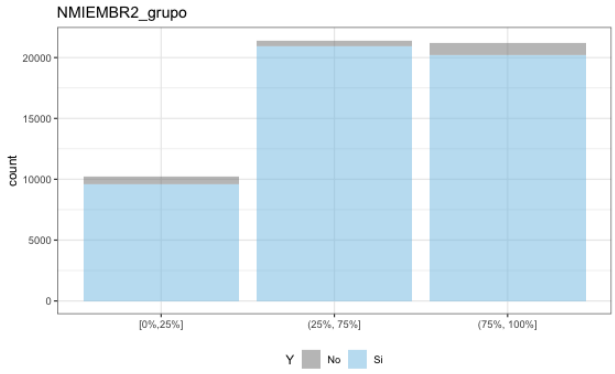


Gràfic 29. *Diagrama de barres de la variable EDAD_grupo.*

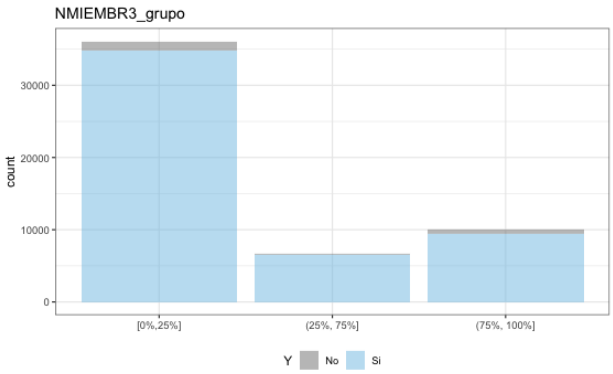


Gràfic 30. *Diagrama de barres de la variable IMPEXAC_grupo.*

En el ***Gràfic 29*** s'observa que independentment de si les llars consumeixen carn o no, majoritàriament l'edat mitjana de les llars es troba compresa entre els 31 i 64 anys. Gràficament, es posa de manifest que per aquelles llars que no consumeixen carn, hi ha un major nombre d'aquestes, que prenen un valor entre 64 i 100 en comptes de 0 i 30 anys. Amb relació al ***Gràfic 30***, aquelles llars que no consumeixen carn tenen uns ingressos nets més baixos que no pas les llars que sí que consumeixen carn.

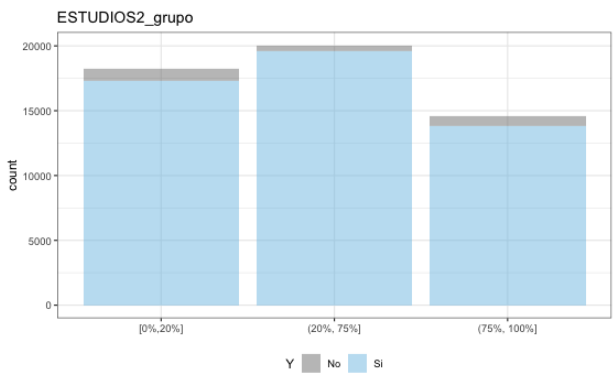


Gràfic 31. *Diagrama de barres de la variable NMIEMBR1_grupo.*

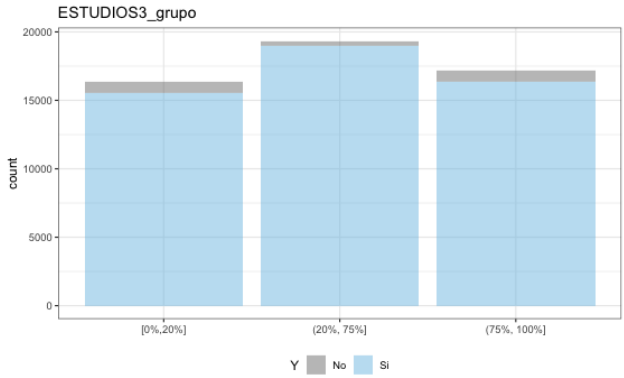


Gràfic 32. *Diagrama de barres de la variable NMIEMBR3_grupo.*

Per una banda, en el ***Gràfic 31*** s'indica que aquelles llars que no consumeixen carn estan constituïdes generalment pel 75% - 100% d'individus entre 31 i 64 anys. En canvi, referent a les llars que sí que consumeixen carn, aquestes es troben compostes majoritàriament entre el 25%-100%. Per altra banda, el ***Gràfic 32*** mostra que per les llars que no consumeixen carn, el 0-25% dels individus tenen més de seixanta-quatre anys i, en segon pla queden aquelles llars formades pel 75%-100%. Respecte a les llars que sí que consumeixen carn, gran part d'elles no són estan constituïdes per individus majors de seixanta-quatre anys.



Gràfic 33. *Diagrama de barres de la variable ESTUDIOS2_grupo.*



Gràfic 34. *Diagrama de barres de la variable ESTUDIOS3_grupo.*

Pel que fa al ***Gràfic 33***, per aquelles llars que sí que consumeixen carn, les classes es troben força balancejades, sent la classe més freqüent aquelles llars constituïdes per un 20-75% d'individus amb estudis obligatoris. En canvi, les llars que no consumeixen carn, les classes també es consideren balancejades, però a diferència de l'anterior, les classes més freqüents són aquelles llars compostes pel 0-20% i el 75%-100% dels individus amb estudis obligatoris. En últim lloc, en el ***Gràfic 34*** es pot observar que, per aquelles llars consumidores de carn, la classe més freqüent són aquelles compostes per un 20-75% d'individus amb estudis superiors. Altrament, per les llars que no consumeixen carn, les classes més freqüents estan constituïdes pel 0-20% i el 75%-100% dels membres de la llar amb estudis superiors.

3.2.2. Preparació de les dades

La representació gràfica de la distribució de les variables en funció de si les llars van consumir carn o no, contribueix a tenir una idea de quines variables poden ser bones predictores per al model i quines d'elles poden ser redundants.

El preprocessament de dades té com a objectiu, que les dades puguin ser admeses en els algoritmes de ML. És per això que, aquesta etapa engloba transformacions que millorin els resultats. Alguns dels aspectes més importants a tenir en compte en aquesta fase són la imputació de valors absents, l'exclusió de variables amb variància propera a zero, la identificació de predictors correlacionats, l'estandardització i l'escalat de variables numèriques i finalment, la binarització de variables qualitatives.

3.2.2.1.Imputació de valors absents

Gran part dels algoritmes en ML no accepten observacions que tinguin valors *missings*. Tal com s'ha especificat en l'apartat 3.1.3, per tal de crear la base de dades resultant, prèviament s'han hagut d'imputar els valors absents mitjançant el paquet *mice*. Per tant, la base de dades resultant no conté cap valor absent i no serà necessari tenir-ho en compte en aquest apartat.

3.2.2.2.Exclusió de variables amb variància igual o propera a zero

Quan una variable presenta un valor de variància igual o proper a zero, és a dir, que el valor de la variable per totes les observacions és el mateix o gairebé el mateix, afegeix al model més soroll que no pas informació, per aquesta raó serà recomanable no incloure-la en el model.

Per detectar aquest fenomen serà necessari utilitzar la funció *nearZeroVar*, que permet identificar aquelles variables amb variància igual o propera a zero. Els resultats es mostren a la següent taula.

Variable	ZeroVar	NZV
GASTOT	FALSE	FALSE
SEXO	FALSE	FALSE
EDAD_grupo	FALSE	FALSE
NACION	**FALSE**	**TRUE**
DENSIDAD	FALSE	FALSE
CCAA	FALSE	FALSE
COMITOT	FALSE	FALSE
NUMOCUP	FALSE	FALSE
IMPEXAC_grupo	FALSE	FALSE
ECIVILSP	FALSE	FALSE
ZONARES	FALSE	FALSE
ESTUDIOS1	**FALSE**	**TRUE**
ESTUDIOS2_grupo	FALSE	FALSE
ESTUDIOS3_grupo	FALSE	FALSE

***Taula 2**. Detecció de variables amb variància igual o propera a zero.*

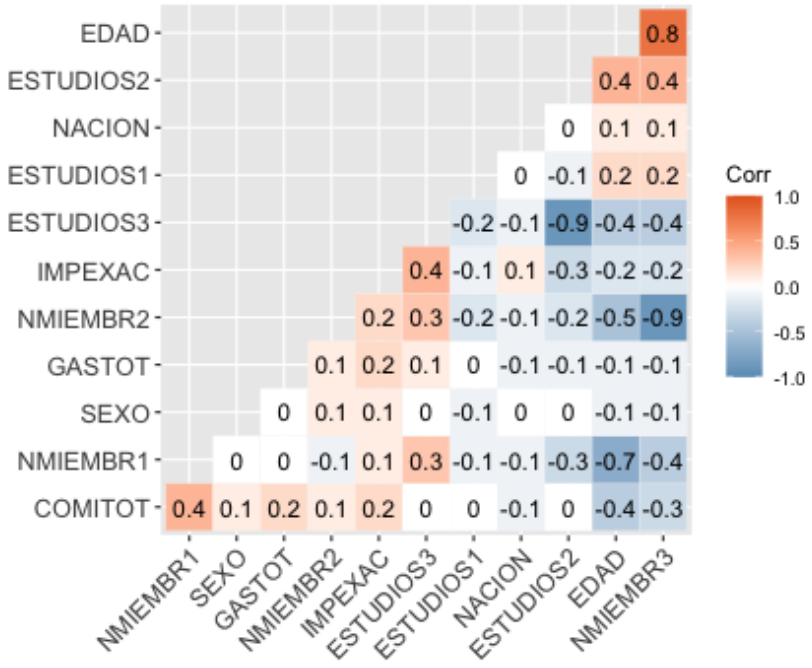
La primera columna de la ***Taula 2***, fa referència als predictors inclosos en el model. La segona, indica si la variable predictora presenta o no una variància igual a zero i la tercera, si presenta o no una variància propera a zero. A partir dels resultats obtinguts de la taula, es pot concloure que no se'n detecta cap variable amb variància zero, però sí amb

variància propera a zero, aquestes variables predictores són les variables NACION i ESTUDIOS1 i, per tant, no s'inclouran en el model.

3.2.2.3.Identificació de predictors correlacionats

Un dels aspectes a tenir en compte a l'hora de construir el model són les correlacions entre les variables. Si dues variables numèriques es troben altament correlacionades, aquestes poden afegir informació redundant al model, i per aquesta raó, no convé incorporar-les. Si això succeeix, es recomana excloure una de les dues o bé, crear una nova variable la qual reculli la informació de les dues. Finalment, si algun dels nivells d'una variable qualitativa té molt poques observacions en comparació amb la resta de nivells, podria donar lloc a errors durant la validació creuada o mostreig *bootstrap,* ja que podria ser que algunes particions no continguessin cap observació d'aquesta classe. En aquests casos, es poden eliminar les observacions del grup amb menys observacions o bé, assegurar que, quan es produeixin les particions, es garanteixi que tots els nivells estiguin representats en cadascuna d'elles.

A efectes pràctics, les correlacions entre les variables contínues es mostren en el ***Gràfic 35.***



Gràfic 35. *Matriu de correlacions de les variables contínues.*

Aquelles variables que presenten una major correlació són EDAD:NMIEMBR3, ESTUDIOS3:ESTUDIOS2, NMIEMBR2:NMIEMBR3, EDAD:NMIEMBR1. La variable EDAD, NMIEMBR1, NMIEMBR2 i NMIMEBR3 recullen la mateixa informació. Per aquesta raó, decidim eliminar les variables NMIEMBR1, NMIMEB2 i NMIEMBR3, ja que la variable EDAD recull la informació de les tres variables anteriors i recordem que ja s'havia categoritzat anteriorment rebent el nom EDAD_grupo. D'aquesta manera, és possible eliminar les correlacions entre EDAD:NMIEMBR3, NMIEMBR2:NMIEMBR3, EDAD:NMIEMBR1. Referent a la correlació entre les variables ESTUDIO2:ESTUDIO3, recordem que també es troba categoritzada i, per tant, no s'ha de tractar com a variable numèrica.

En el cas de les variables qualitatives, s'assegurarà que tots els nivells de les variables estiguin representats en cadascuna de les mostres.

3.2.2.4.Estandardització de variables numèriques

Quan els predictors són de caràcter numèric, és aconsellable centrar les variables predictores perquè tinguin una mitjana de 0 i, la seva interpretació sigui més senzilla. Per altra banda, és recomanable escalar les variables perquè quan una variable presenta una escala molt gran (com per exemple la variable GASTOT) els coeficients del model tindran un ordre molt petit, en canvi, si s'escala la variable (mesura la variable GASTOT en milers), els coeficients del model tindran un major ordre. És a dir, la convenció d'estandarditzar i escalar les variables predictores existeix principalment perquè les unitats dels coeficients siguin les mateixes.

El procés d'estandardització i escalat es podrà incloure en la modelització del conjunt d'entrenament que s'explicarà més endavant.

3.2.2.5.Binarització de variables qualitatives

Gran part dels models requereixen que totes les variables predictores siguin contínues. Com a conseqüència, serà necessari transformar les variables categòriques en variables numèriques per tal que l'algoritme funcioni. La tècnica més comuna s'anomena procés de binarització o *one hot encoding*. Aquest consisteix a crear noves variables *dummy* amb cadascun dels nivells de les variables qualitatives. Per exemple, la variable

anomenada ZONARES la qual conté els nivells zona residencial Urbana i Rural, esdevindran dues noves variables (ZONARES_Urbana, ZONARES_Rural), totes amb el valor 0 excepte la que coincideix amb l'observació, que prendrà valor 1.

Per tant, la base de dades resultant després de realitzar la binarització, estarà formada per 23 variables predictores i la variable resposta. A continuació, es mostra una taula detallada amb totes les variables.

Variable	**Descripció**
GASTOT	Despesa monetària i no monetària total en euros elevat temporal i poblacionalment.
SEXO	Percentatge de dones a la llar. (0-100)
COMITOT	Número de Dinars i Sopars bisetmanals. (0-999)
EDAD_grupo_X.64..100.	Edat mitjana dels membres de la llar entre els 64 i 100 anys. (1: Sí; 0: No)
EDAD_grupo_X.0.30.	Edat mitjana dels membres de la llar entre els 0 i 30 anys. (1: Sí; 0: No)
DENSIDAD_X2	Zona amb densitat poblacional intermèdia. (1: Sí; 0: No)
DENSIDAD_X3	Zona amb densitat poblacional intermèdia disseminada. (1: Sí; 0: No)
CCAA_X2	CCAA amb PIB per càpita entre 25.000 i 20.000 euros. (1: Sí; 0: No)
CCAA_X3	CCAA amb PIB per càpita inferior a 20.000 euros. (1: Sí; 0: No)
NUMOCUP_X1	Un ocupat a la llar. (1: Sí; 0: No)
NUMOCUP_X2	Dos ocupats a la llar. (1: Sí; 0: No)
NUMOCUP_X3	Tres ocupats a la llar. (1: Sí; 0: No)
NUMOCUP_X4	Quatre o més ocupats a la llar. (1: Sí; 0: No)
IMPEXAC_grupo_X.1800..Inf.	L'import en euros dels ingressos mensuals nets totals de la llar es major de 1800 €. (1: Sí; 0: No)
IMPEXAC_grupo_X.0..1100.	L'import dels ingressos mensuals nets totals de la llar es troba entre 0 € i 1800 €. (1: Sí; 0: No)
ECIVILSP_X2	Estat Civil del Sustentador Principal Casat. (1: Sí; 0: No)

ECIVILSP_X3	Estat Civil del Sustentador Principal Vidu. (1: Sí; 0: No)
ECIVILSP_X4	Estat Civil del Sustentador Principal Separat o Divorciat. (1: Sí; 0: No)
ZONARES_X2	Zona Residencial Rural. (1: Sí; 0: No)
ESTUDIOS2_grupo_X.20...75..	Entre el 20% i 75% dels membres de la llar tenen Estudis Obligatoris. (1: Sí; 0: No)
ESTUDIOS2_grupo_X.75...100..	Més del 75% dels membres de la llar tenen Estudis Obligatoris. (1: Sí; 0: No)
ESTUDIOS3_grupo_X.20...75..	Entre el 20% i 75% dels membres de la llar tenen Estudis Superiors. (1: Sí; 0: No)
ESTUDIOS3_grupo_X.75...10..	Més del 75% dels membres de la llar tenen Estudis Superiors. (1: Sí; 0: No)
Y	Consum de carn. (1: Sí; 0: No)

***Taula 3**. Descripció de les variables de la Base de Dades Resultant*

3.2.3. Models predictius

Un cop construïda i preprocessada la base de dades, ja es podran implementar els models predictius en l'àmbit de l'aprenentatge automàtic. Per dur a terme la modelització, s'utilitzarà el paquet *caret* disponible a *RStudio*.

Els models implementats seran: la Regressió Logística, l'Arbre de Decisió Simple, *Bagging*, *Random Forest* i *Gradient Boosting*. Tots aquests seran avaluats mitjançant *Cross – Validation* (CV) i, fonamentalment s'utilitzaran les corbes ROC i l'AUC com a mètrica de validació. Finalment, veurem quin dels models s'ajusta millor a la variable resposta binària, consum de carn de les llars espanyoles.

Com s'ha esmentat en els apartats anteriors, en aquest punt s'introdueixen les alternatives referents al problema del desequilibri de classes, l'escalat i el centrat de les variables. En referència a la solució del desequilibri de classes, si el sobre mostreig o submostreig es realitza a priori, durant l'ajust del model, les mostres generades no reflectirien el desequilibri de classes pel fet que ja hauria sigut tractat abans d'avaluar el model i com a conseqüència, donaria lloc a estimacions "optimistes". L'alternativa proposada, és dur a terme el submostreig/sobre mostreig dins del procediment de remostreig, això també s'aplicarà pel *preprocessing*. El principal inconvenient d'aquesta alternativa, és el temps

computacional i, en alguns casos pot complicar l'anàlisi. Tot i això, en el nostre cas utilitzarem aquesta última estratègia.

El procés de modelització de les dades en l'àmbit de ML es divideix tres fases:

1. **Entrenament**

Per evitar el problema de sobre ajust del model explicat anteriorment, és necessari dividir la base de dades en el conjunt d'entrenament i el conjunt test. La mida adequada de les particions depèn de la quantitat de dades disponibles i la seguretat en l'estimació de l'error. És per això que 80%-20% sol donar bons resultats, on el 80% de les dades correspon al conjunt d'entrenament i el 20% restant, correspon al conjunt test. El repartiment s'ha de realitzar de manera aleatòria i és important verificar que la distribució de la variable resposta és similar tant al conjunt d'entrenament com al conjunt del test.

	Valor Variable Resposta (Y)	
	No	Sí
Conjunt d'entrenament	0.03922171	0.96077829
Conjunt de test	0.03915634	0.96084366

Taula 4. *Freqüències relatives de la variable resposta.*

La funció *trainControl* del paquet *caret* permet avaluar, mitjançant el remostreig, l'efecte dels paràmetres de l'ajust del model en el rendiment, escollir el model òptim entre els paràmetres i estimar el rendiment del model a partir del conjunt d'entrenament.

El primer argument de la funció *trainControl* correspon amb el mètode de remostreig, en el nostre cas utilitzarem la *K Fold – Crosss Validation*[3]. El segon argument indica el nombre de plecs (*folds)* i de repeticions, en el nostre cas es faran servir tres validacions creuades de deu *folds* cadascuna. El tercer argument serà *classProb* on prendrà un valor lògic que determina si les probabilitats de classe s'han de calcular per les mostres emprades durant el remostreig, en el nostre cas serà igual *True*. En últim lloc, l'argument *summaryFuntion* és una funció que computa resums del rendiment alternatiu, en tractar-se d'un problema de classificació, en el nostre cas farem ús de *twoClassSummary*.
Per defecte la funció *trainControl* utilitzarà el conjunt d'entrenament original, però també té l'opció d'especificar el tipus de remostreig desitjat. L'argument per incloure el

[3] Mètode estadístic utilitzat per estimar l'habilitat dels models d'aprenentatge automàtic.

submostreig i sobre mostreig es realitza mitjançant la instrucció *sampling = down* i *sampling = up* respectivament.

A continuació, s'ajustarà el model al conjunt d'entrenament a partir de la funció *train*: el primer argument correspon a l'equació del model; el segon, al conjunt de dades d'entrenament; el tercer, a la funció *preProcess* que permet centrar, escalar i imputar, però aquesta última no s'utilitzarà, ja que els *missings* s'han imputat anteriorment; el quart fa referència a la mètrica que es vol utilitzar, que en aquest cas serà la corba ROC i finalment, l'argument *trControl* el qual crida a la funció *trainControl* prèviament creada.

2. Validació

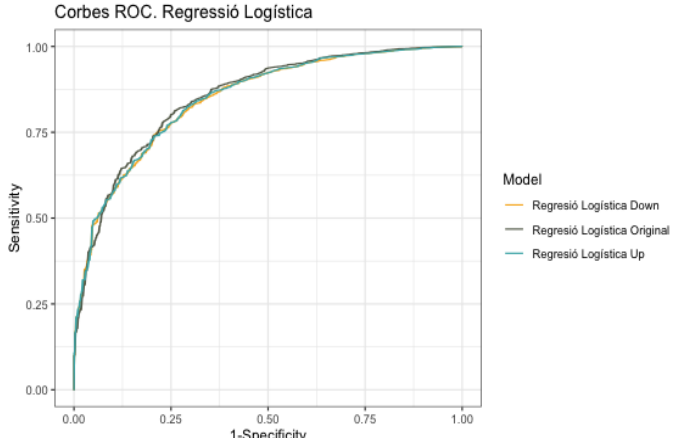
Una vegada generat el model a partir del conjunt d'entrenament, s'avaluarà amb el 20% restant de les dades. A partir de la matriu de confusió de les dades de test, s'utilitzarà l'Exactitud, Precisió, Sensibilitat, Especificitat i la corba ROC per mesurar la capacitat predictiva del model enfront de noves observacions.

3. Prova

Finalment, es faran prediccions per provar el model amb el conjunt test, és a dir, amb les dades que el model no ha vist.

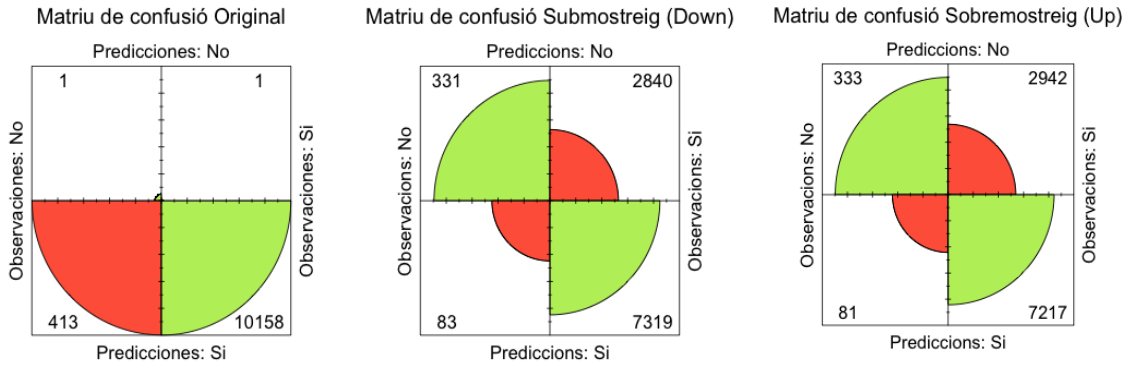
3.2.3.1.Regressió logística

Per definir l'entrenament de la regressió logística dins la funció *train* caldrà especificar el mètode i la família a la qual pertany, mitjançant els arguments *glm* i *binomial* respectivament. A continuació, a partir de la corba ROC i la matriu de confusió, es procedirà a realitzar la validació del model.



***Gràfic 36**. Corbes ROC de la Regressió Logística.*

A simple vista i de manera genèrica, totes les corbes es troben allunyades de la recta de 45 graus i si es comparen les corbes entre elles, no s'observen grans diferències.



Gràfic 37. *Matrius de Confusió de la Regressió Logística.*

La interpretació referent a les matrius de confusió és la següent; la matriu de confusió original mostra que del conjunt test, hi ha 1 llar que el model prediu que no consumeix carn quan realment no en consumeix, 1 llar que el model prediu que no consumeix carn quan realment sí que consumeix, 413 llars que el model prediu que sí que consumeixen, però realment no en consumeixen i 10.158 llars que el model prediu que sí que consumeixen i realment sí que en consumeixen. Referent a la matriu de confusió de submostreig, hi ha 331 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 2.840 llars que el model prediu que no consumeixen carn quan realment sí que consumeixen, 83 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 7.318 llars que el model prediu que si consumeixen carn i realment sí que en consumeixen. Pel que fa a la matriu de confusió de sobre mostreig, hi ha 333 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 2.942 llars que el model prediu que no consumeixen carn quan realment sí que consumeixen, 81 llars que el model prediu que sí que consumeixen carn, però realment no en i 7.217 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen.

En termes generals, respecte a la matriu de confusió de la base de dades original, quan s'aplica el remostreig es percep una millora considerable en la classificació de les observacions. Si es comparen les matrius de confusió de sobre mostreig i submostreig, no mostren grans diferències entre elles.

Analíticament, les mètriques de validació de la base de dades original, del submostreig i del sobre mostreig es mostren a ***Taula 5.***

	Exactitud	Precisió	Sensibilitat	Especificitat
Original	0.9608	0.96093	0.9999	0.0024
Submostreig	0.7235	0.9987	0.7204	0.7995
Sobre mostreig	0.7141	0.9889	0.7104	0.8043

Taula 5. *Mètriques de Validació de la Regressió Logística*

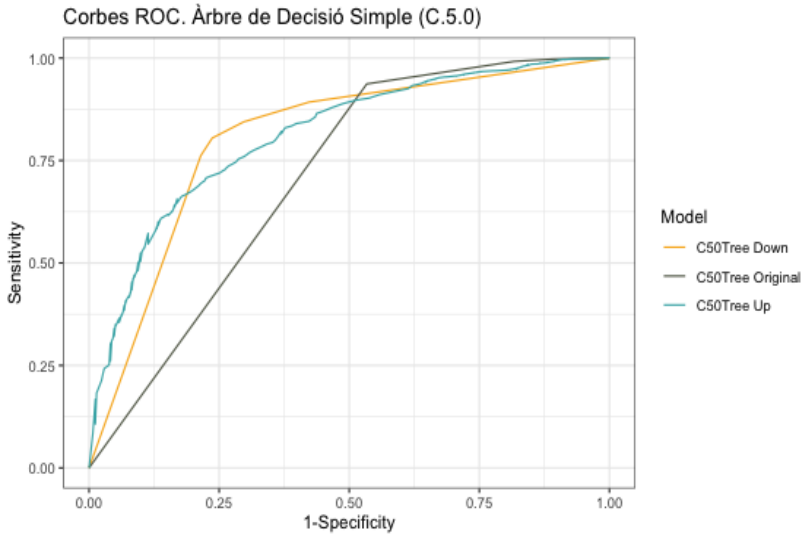
El conjunt de dades original presenta valors molt elevats en termes d'Exactitud, Precisió i Sensibilitat. A conseqüència del desequilibri de classes, l'exactitud i la precisió no seran apropiades a l'hora d'avaluar el model, ja que gran part de les llars consumeixen carn i l'algoritme classifica totes les observacions com a positives, provocant que no sigui capaç de detectar aquelles llars que no consumeixen carn. Com a alternativa, s'avaluaran la Sensibilitat i l'Especificitat. El primer d'ells, pren un valor molt proper a un i el segon, a zero. És a dir, el model és perfectament capaç d'identificar els verdaders positius, però no és capaç d'identificar els veritables negatius.

En el cas del sobre mostreig, s'obtenen uns valors elevats d'Exactitud, Precisió, Sensibilitat i Especificitat. A diferència del cas anterior, sí que podem utilitzar totes les mètriques, ja que no presenta un problema de desequilibri de classes. Generalment, el model és capaç de classificar la variable resposta correctament i de predir de manera correcta tant les observacions positives com negatives, és a dir, aquelles llars que consumeixen carn i les que no en consumeixen.

Si comparem els dos últims, s'assoleix que el mètode de sobre mostreig és millor en termes de Precisió i Especificitat, mentre que el submostreig és millor en termes d'Exactitud i Sensibilitat.

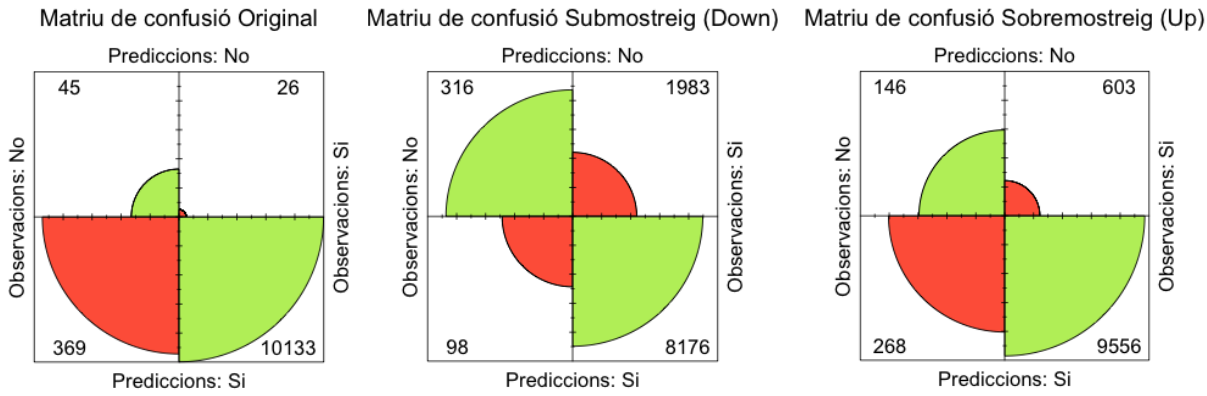
3.2.3.2.Arbre de decisió

L'especificació dins de la funció *train* per l'arbre de decisió simple serà únicament el mètode *C5.0Tree*. Cal esmentar que, aquest tipus de model no conté cap hiperparàmetre. Un cop definit l'entrenament del model, es procedirà a avaluar-lo gràficament i analíticament.



Gràfic 38. *Corbes ROC de l'Arbre de Decisió.*

Gràficament, s'observa que els models emprant el remostreig, es troben més troben allunyades de la recta de quaranta-cinc graus, que no pas el model amb la base de dades original. Per tant, aquest últim model és pitjor que els altres dos.



Gràfic 39*. Matrius de Confusió de l'Arbre de Decisió.*

La interpretació de les matrius de confusió obtingudes després d'aplicar l'algoritme és la següent; la matriu de confusió original la qual presenta el desequilibri de classes mostra que del conjunt test hi ha 45 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 26 llars que el model prediu que no consumeixen carn quan realment sí que en consumeixen, 369 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 10.133 llars que el model prediu que sí que consumeixen i realment sí que en consumeixen. En relació amb la matriu de confusió de submostreig, hi ha 316 llars que el model prediu que no consumeixen carn quan realment

no en consumeixen, 1.983 llars que el model prediu que no consumeixen carn quan realment sí que en consumeixen, 98 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 8.176 llars que el model prediu que sí que consumeixen i realment sí que en consumeixen. Pel que fa a la matriu de confusió de sobre mostreig, hi ha 146 llars que el model prediu que no consumeixen carn quan realment no consumeixen, 603 llars que el model prediu que no consumeix carn quan realment sí que consumeixen, 268 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 9.556 llars que el model prediu que sí que consumeixen i realment sí que en consumeixen.

En termes numèrics, els resultats es mostren a la següent taula:

	Exactitud	Precisió	Sensibilitat	Especificitat
Original	0.9626	0.9649	0.9974	0.1087
Submostreig	0.8032	0.9882	0.8048	0.7633
Sobre mostreig	0.9176	0.9727	0.9406	0.3527

Taula 6. *Mètriques de Validació de l'Arbre de Decisió.*

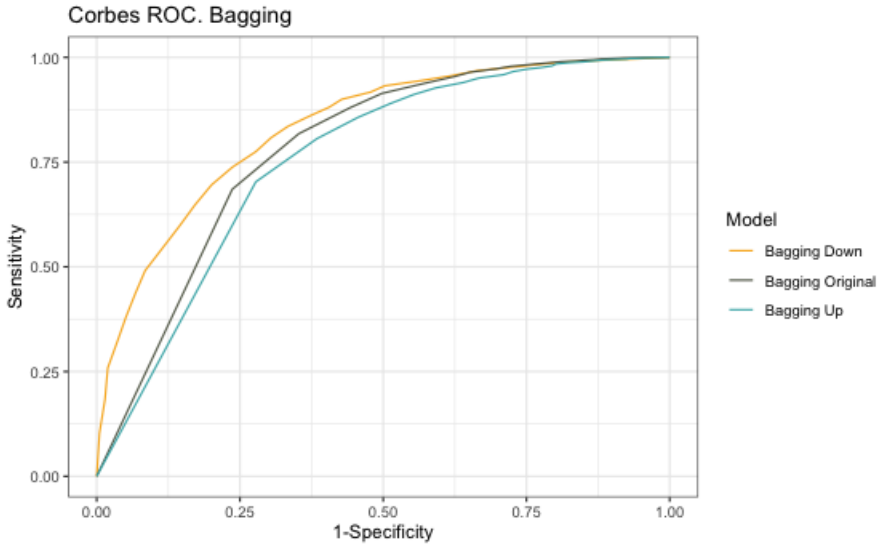
Referent al conjunt de dades originals, a causa del problema de desequilibri de classes, s'obtenen valors molt elevats en termes d'Exactitud, Precisió i Sensibilitat, la qual cosa l'exactitud i la precisió quedaran invalidades i com a alternativa, s'avaluaran la Sensibilitat i l'Especificitat. El valor del primer d'ells, es troba proper a 1, mentre que el segon, a zero. Per tant, el model és perfectament capaç d'identificar els verdaders positius, però no és capaç d'identificar els veritables negatius.

Per altra banda, pel conjunt de dades obtingut del sobre mostreig, s'obtenen uns valors elevats d'Exactitud, Precisió, Sensibilitat i Especificitat. A diferència del cas anterior, com que no presenta un problema de desequilibri de classes, es poden utilitzar totes les mètriques de la taula. Es pot dir que el model és capaç de classificar la variable resposta correctament i de predir de manera correcta tant les observacions positives com negatives, és a dir, aquelles llars que consumeixen carn i les que no en consumeixen.

Els valors obtinguts pel conjunt de dades emprant el submostreig, en termes d'Exactitud, Precisió i Sensibilitat són molt similars als obtinguts amb el sobre mostreig. Però l'Especificitat és molt menor. És a dir, que no és capaç de detectar els verdaders negatius i, per tant, aquest no serà un bon model per dur a terme prediccions.

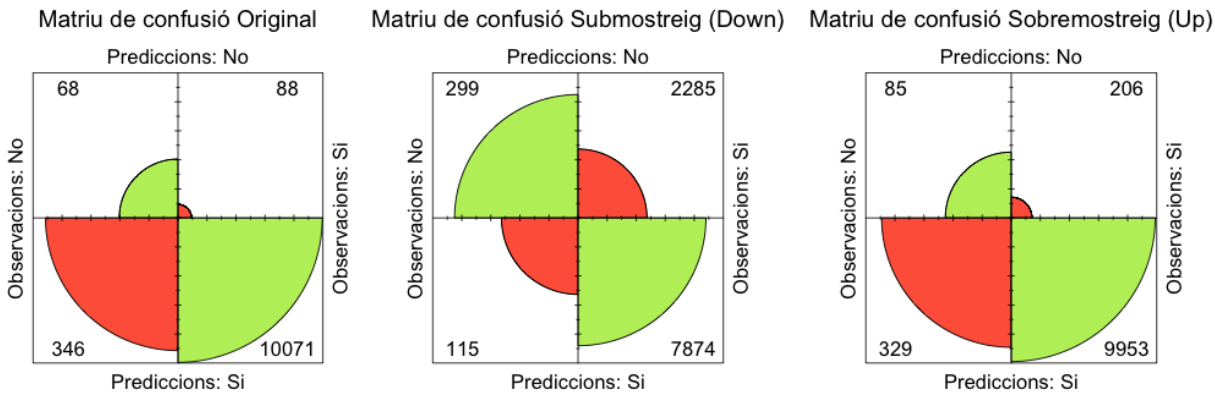
3.2.3.3.Bagging

Els arguments per definir la funció *train* són el mètode especificat com *treebag* i l'hiperparàmetre corresponent al nombre de rèpliques bootstrap on *nbagg=25*. Les respectives corbes ROC i matrius de confusió es mostren a continuació.



Gràfic 40. *Corbes ROC de Bagging.*

La corba ROC que es troba més allunyada de la recta de quaranta-cinc graus correspon amb el model utilitzant el submostreig. El model amb menor capacitat predictiva de tots tres, és el model emprant el sobre mostreig ja que la corba és la més propera a la recta.



Gràfic 41. *Matrius de Confusió Bagging.*

La interpretació de les corresponents matrius de confusió obtingudes d'aplicar l'algoritme *Bagging* és la següent; la matriu de confusió original la qual presenta el desequilibri de classes mostra que del conjunt test hi ha 68 llars que el model prediu que no consumeix carn quan realment no consumeix, 88 llars que el model prediu que no consumeix carn quan realment sí que consumeix, 346 llars que el model prediu que sí que consumeixen,

però realment no en consumeixen i 10.071 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen. Amb relació a la matriu de confusió de submostreig, hi ha 299 llars que el model prediu que no consumeixen carn quan realment no consumeixen, 2.285 llars que el model prediu que no consumeix carn quan realment sí que consumeixen, 115 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 7.874 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen. Pel que fa a la matriu de confusió de sobre mostreig, hi ha 85 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 206 llars que el model prediu que no consumeixen carn quan realment sí que consumeixen, 329 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 9.953 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen.

Analíticament, els resultats obtinguts han estat els següents:

	Exactitud	Precisió	Sensibilitat	Especificitat
Original	0.9590	0.9668	0.9918	0.1643
Submostreig	0.7730	0.9856	0.7751	0.7222
Sobre mostreig	0.9494	0.9680	0.9797	0.2053

Taula 7. *Mètriques de Validació de Bagging.*

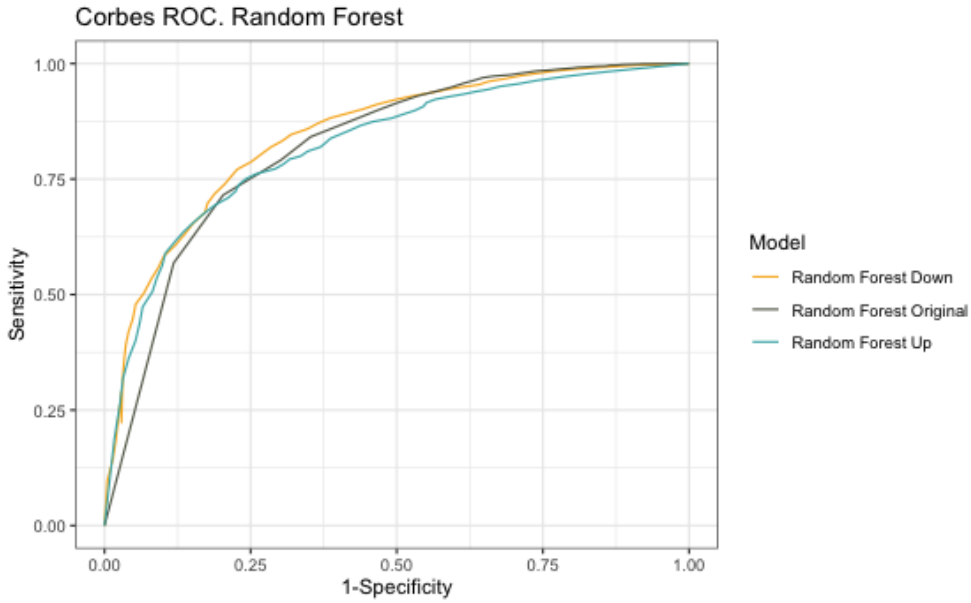
A conseqüència del desequilibri de classes, en el conjunt de dades originals s'observen valors molt elevats d'Exactitud, Precisió i Sensibilitat, la qual cosa les dues primeres no seran útils per la validació del model. En el seu lloc, s'avaluaran la Sensibilitat i l'Especificitat. El valor del primer d'ells és proper a 1, mentre que el segon, a zero, per tant, el model és perfectament capaç d'identificar els verdaders positius, però no és capaç d'identificar els veritables negatius.

Referent al conjunt de dades utilitzant al sobre mostreig, l'Exactitud, la Precisió, la Sensibilitat i l'Especificitat prenen valors propers a un. A diferència del cas anterior, com que no presenta un problema de desequilibri de classes, podem fer servir totes les mètriques de la taula. Globalment, es pot dir que el model és capaç tant de classificar la variable resposta com de realitzar prediccions correctament.

Igual que al model anterior, el valor de l'Especificitat pren un valor inferior en el submostreig. És a dir, que tampoc no és capaç de detectar els verdaders negatius i, per tant, no serà un bon model per dur a terme prediccions.

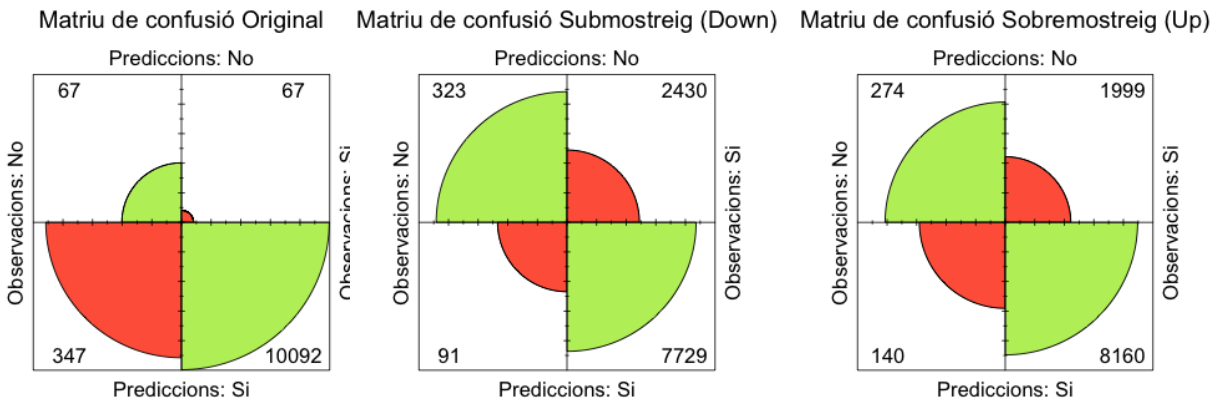
3.2.3.4.Random Forest

Per estimar el model amb la funció *train* de *caret*, caldrà especificar el mètode que serà *rf* i l'hiperparàmetre *ntree* que correspondrà al nombre d'arbres i prendrà valor igual a 50.



Gràfic 42. *Corbes ROC de Random Forest.*

A simple vista, la corba ROC del model emprant el submostreig, és la que es troba més allunyada de la recta de quaranta-cinc graus tot i que en general, totes les corbes es troben a la mateixa distància.



Gràfic 43*. Matrius de Confusió Random Forest.*

La descripció de les matrius de confusió obtingudes d'aplicar l'algoritme *Random Forest* és la següent; la matriu de confusió original la qual presenta el desequilibri de classes mostra que del conjunt test hi ha 67 llars que el model prediu que no consumeix carn quan realment no consumeix, 67 llars que el model prediu que no consumeix carn quan realment sí que consumeix, 347 llars que el model prediu que sí que consumeixen, però realment no en consumeixen i 10.092 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen. Amb relació a la matriu de confusió de submostreig, hi ha 323 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 2.430 llars que el model prediu que no consumeixen carn quan realment sí que en consumeixen, 91 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 7.729 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen. Pel que fa a la matriu de confusió de sobre mostreig, hi ha 274 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 1.999 llars que el model prediu que no consumeixen carn quan realment sí que consumeixen, 140 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 8.160 llars que el model prediu que sí que consumeixen i realment sí que en consumeixen.

Analíticament, les mètriques d'avaluació de capacitat predictiva es mostren a la ***Taula 8.***

	Exactitud	**Precisió**	**Sensibilitat**	**Especificitat**
Original	0.9608	0.9668	0.9934	0.1618
Submostreig	0.7616	0.9884	0.7608	0.7802
Sobre mostreig	0.7977	0.9831	0.8032	0.6618

Taula 8. *Mètriques de Validació de Random Forest.*

El conjunt de dades original presenta valors molt elevats en termes d'Exactitud, Precisió i Sensibilitat. Això és a causa del problema del desequilibri de classes i, per tant, l'exactitud i la precisió no seran adequades. En el seu lloc, s'avaluaran la Sensibilitat i l'Especificitat. El primer d'ells pren un valor molt proper a un i, el segon, a zero. Com en la resta dels algoritmes anteriors, el model és perfectament capaç d'identificar els verdaders positius, però no els veritables negatius.

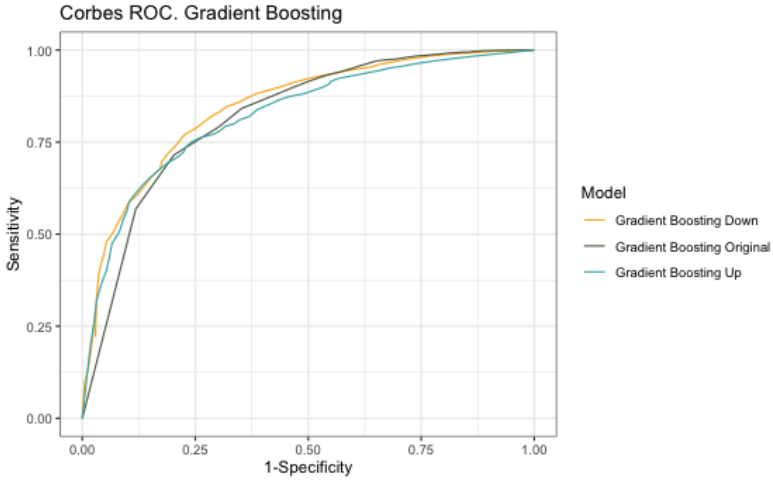
Del conjunt de dades emprant sobre mostreig, s'obtenen uns valors elevats d'Exactitud, Precisió, Sensibilitat i Especificitat. Ara sí que podem utilitzar totes les mètriques, ja que no existeix el problema de desequilibri de classes. De manera general, el model és capaç

de classificar correctament la variable resposta, consum de carn, i de predir de manera adequada tant les observacions positives com negatives, és a dir, aquelles llars que consumeixen carn i les que no en consumeixen.

Si comparem els dos últims, s'obté que el mètode de sobre mostreig és millor en termes d'Exactitud i Sensibilitat, mentre que el submostreig és millor en termes de Precisió i Especificitat.

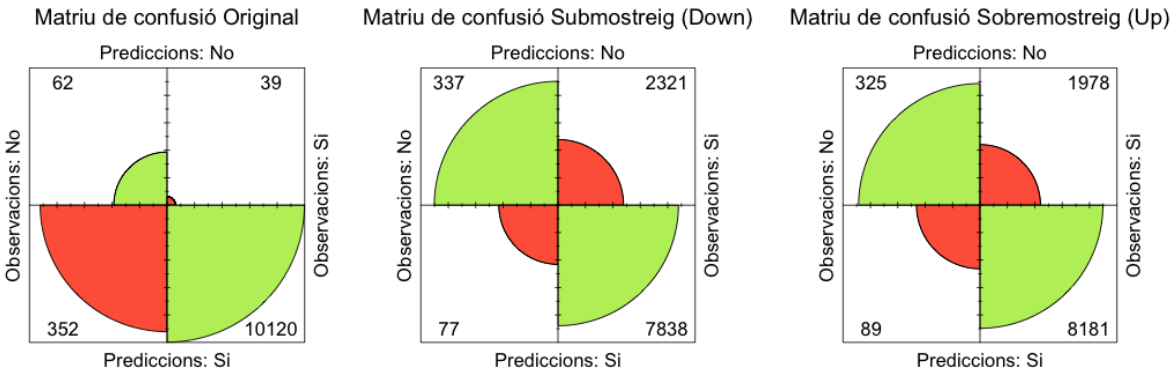
3.2.3.5.Gradient Boosting

Per definir el model *Gradient Boosting*, serà convenient especificar el mètode, que serà *gmb*, i l'hiperparàmetre corresponent amb la profunditat de l'arbre. Concretament, en anàlisi s'utilitzarà *tuneLength* = 5. Seguidament, es mostrarà la representació gràfica de les corbes ROC i les respectives matrius de confusions.



Gràfic 44. *Corbes ROC de Gradient Boosting.*

En el gràfic s'observa com la corba ROC corresponent al model del submostreig és la que es troba més allunyada de la recta de quaranta-cinc graus. Generalment, totes les corbes aconsegueixen la mateixa distància i no es veuen grans dissimilituds entre elles.



Gràfic 45. *Matrius de Confusió Gradient Boosting.*

La interpretació de les matrius de confusió obtingudes de l'algoritme *Gradient Boosting* és la següent; la matriu de confusió original la qual presenta el desequilibri de classes mostra que del conjunt test, hi ha 62 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 29 llars que el model prediu que no consumeixen carn quan realment sí que consumeixen, 352 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 10.120 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen. Amb relació a la matriu de confusió de submostreig, hi ha 337 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 2.321 llars que el model prediu que no consumeixen carn quan realment sí que en consumeixen, 77 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 7.838 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen. Pel que fa a la matriu de confusió de sobre mostreig, hi ha 325 llars que el model prediu que no consumeixen carn quan realment no en consumeixen, 1.978 llars que el model prediu que no consumeixen carn quan realment sí que en consumeixen, 89 llars que el model prediu que sí que consumeixen carn, però realment no en consumeixen i 8.181 llars que el model prediu que sí que consumeixen carn i realment sí que en consumeixen.

En termes numèrics, els resultats de les mètriques per tal d'avaluar el model es troben a la ***Taula 9***.

	Exactitud	**Precisió**	**Sensibilitat**	**Especificitat**
Original	0.9630	0.9664	0.9962	0.1498
Sub mostreig	0.7732	0.9903	0.7715	0.8140
Sobre mostreig	0.8045	0.9892	0.8053	0.7850

Taula 9. *Mètriques de Validació de Gradient Boosting.*

El conjunt de dades original presenta valors molt elevats tant d'Exactitud com de Precisió i de Sensibilitat. Això és degut al problema del desequilibri de classes i, per tant, l'exactitud i la precisió quedaran invalidades. En el seu lloc, s'avaluaran la Sensibilitat i l'Especificitat. El primer d'ells pren un valor proper a un i l'altre a zero. Com s'ha vist en la resta dels algoritmes, el model és perfectament capaç d'identificar els verdaders positius, però no els veritables negatius.

En el conjunt de dades emprant sobre mostreig, sí que podem utilitzar totes les mètriques, ja que no existeix el problema de desequilibri de classes. Es percep que les quatre mètriques presenten valors elevats, per tant, el model és capaç de classificar correctament la variable resposta, consum de carn, i de predir de manera adequada tant les observacions positives com negatives, és a dir, aquelles llars que consumeixen carn i les que no en consumeixen.

Si comparem els dos últims, s'obté que el mètode de sobre mostreig és millor en termes d'Exactitud i Sensibilitat, mentre que el sub mostreig és millor en termes de Precisió i Especificitat.

3.2.4. Comparació de models

Una vegada entrenats (amb el conjunt d'entrenament) i validats (amb el conjunt test) els diferents models, caldrà identificar quin de tots prediu millor la variable resposta. En la modelització s'ha vist que, quan s'introdueix el submostreig o sobre mostreig respecte al conjunt de dades originals, la matriu de confusió millora considerablement, permetent així al model predir correctament tant les llars que no consumeixen carn com les que sí. Per aquesta raó, serà necessari escollir amb quin tipus de remostreig treballarem; si amb sobre mostreig o bé amb el submostreig. Per escollir-lo, es compararan diferents mètriques de cada model per tal d'analitzar si hi ha diferències significatives entre tots dos. A continuació, es mostra una taula resumida amb els resultats obtinguts de l'apartat anterior.

		Regressió Logística	**Arbre de decisió**	***Bagging***	***Random Forest***	***Gradient Boosting***
Sub-mostreig	**Exactitud**	0.7235	0.8032	0.7730	0.7616	0.7732
	Precisió	0.9887	0.9882	0.9856	0.9884	0.9903
	Sensibilitat	0.7204	0.8048	0.7751	0.7608	0.7715
	Especificitat	0.7995	**0.7633**	**0.7222**	0.7802	0.8140
Sobre mostreig	**Exactitud**	0.7141	0.9176	0.9494	0.7977	0.8045
	Precisió	0.9889	0.9727	0.9680	0.9831	0.9892
	Sensibilitat	0.7104	0.9406	0.9797	0.8032	0.8053
	Especificitat	0.8043	**0.3527**	**0.2053**	0.6618	0.7850

Taula 10. *Mètriques de Validació de tots els models (test).*

Tal com s'observa a la ***Taula 10***, totes les mètriques prenen valors similars excepte l'Especificitat dels models Arbres de Decisió i *Bagging*, on el valor d'aquests és molt

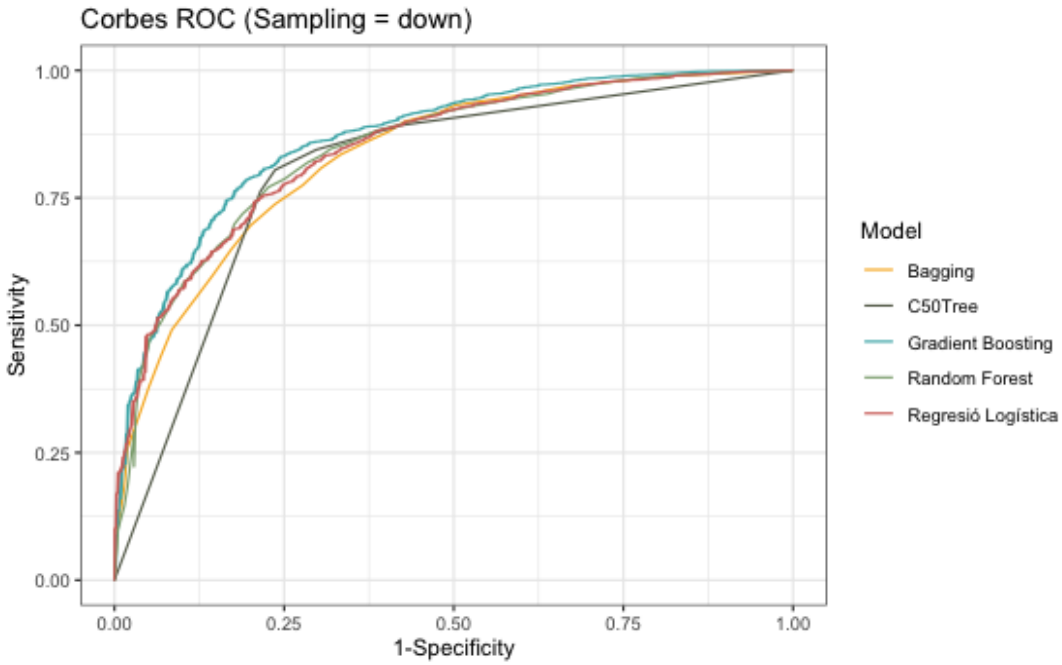
menor en el sobre mostreig que no pas en el submostreig, per tant, els models referents al sobre mostreig no seran capaços de predir correctament els verdaders negatius. Per aquesta raó, **es decidirà treballar amb els models corresponents al submostreig.**

La comparació de models es realitzarà emprant dos mètodes diferents; el primer d'ells, consistirà a comparar les diferents mètriques de validació respecte al conjunt d'entrenament i s'escollirà aquell model que tingui una major capacitat predictiva. El segon, correspondrà en comprar les diferents mètriques dels models en el conjunt test i s'escollirà aquell model que millor ajusti les dades.

a. Conjunt Entrenament

	Gradient Boosting	***Random Forest***	**Regressió Logística**	***Bagging***	**Arbre de decisió**
AUC	0.8540	0.8351	0.8298	0.8250	0.8095

Taula 11. *AUC de tots els models.*

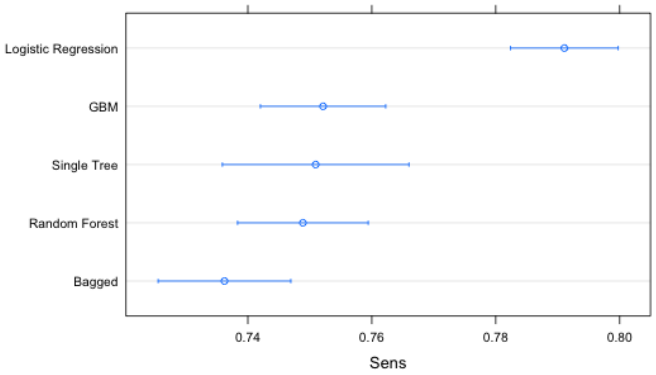


Gràfic 46. *Corba ROC de tots els models.*

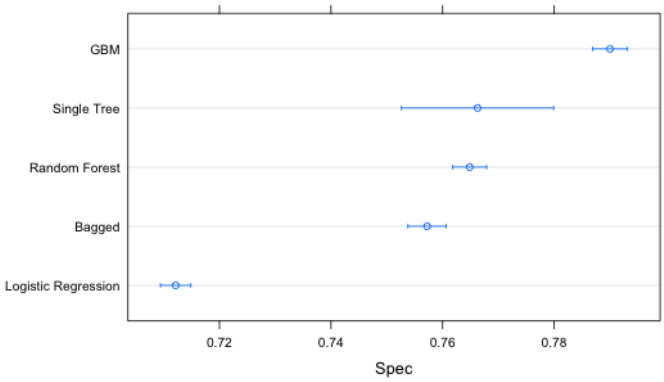
A partir del conjunt d'entrenament, observant els valors AUC de la ***Taula 11*** i les corbes ROC del ***Gràfic 46***, el model *Gradient Boosting* és aquell que presenta un major valor AUC i, per tant, el que es troba més allunyat de la recta de quaranta-cinc graus, seguit del model *Random Forest* i la Regressió Logística. En canvi, l'Arbre de decisió pren el valor

AUC més baix i com a conseqüència, és el que s'apropa més a la diagonal. Com a resultat, els algoritmes *Gradient Boosting* i *Random Forest* es podrien considerar els millors models.

El ***Gràfic 47*** mostra la Sensibilitat dels diferents models i, s'observa que generalment tots ells es troben al voltant del 0,7 i el model de Regressió Logística, és el que pren un valor més elevat entre la resta. Pel que fa a l'Especificitat del ***Gràfic 48***, tots els models presenten una menor variabilitat respecte a la sensibilitat (menys l'Arbre de decisió) i els valors més alts corresponen als models *Gradient Boosting* i l'Arbre de decisió, però aquest últim presenta un rang de variància major que el *Gradient Boosting*.



Gràfic 47. *Sensibilitat de tots els models.*



Gràfic 48*. Especificitat de tots els models.*

A continuació, mitjançant el test de Friedman, es compararan si existeixen diferències significatives entre les mètriques ROC, Sensitivitat i Especificitat dels diferents models.

	P valor
ROC	2,2e-16
Sensitivitat	7,724e-14
Especificitat	2,2e-16

Taula 12. *Test de Friedman*

Sota la hipòtesi nul·la que no hi ha diferències significatives entre els cinc models i sota un nivell de significació del 0,05, s'obté un p valor inferior a 0,05 en tots els casos. Com a resultat, es pot afirmar que hi ha evidències significatives per rebutjar la hipòtesi nul·la i concloure que tots els models no són equivalents. El test no és capaç d'indicar entre quins models existeixen diferències, és per això que, a partir del test de suma de rangs de Wicoxon s'hauran de realitzar comparacions múltiples per tal d'identificar-les.

Model A	Model B	P valor
Gradient Boosting	Bagging	7.790492e-09
Regressió Logística	Bagging	3.884721e-02
Regressió Logística	Gradient Boosting	7.790492e-09
Random Forest	Bagging	1.068958e-06
Random Forest	Gradient Boosting	7.790492e-09
Random Forest	Regressió Logística	2.286724e-02
Arbre de decisió	Bagging	6.732061e-06
Arbre de decisió	Gradient Boosting	7.790492e-09
Arbre de decisió	Regresió Logística	4.692061e-07
Arbre de decisió	Random Forest	2.205555e-08

Taula 13. *Comparacions múltiples de la mètrica ROC (AUC).*

Model A	Model B	P valor
Gradient Boosting	Bagging	4.913309e-03
Regressió Logística	Bagging	1.577600e-08
Regressió Logística	Gradient Boosting	1.317263e-08
Random Forest	Bagging	4.389818e-02
Random Forest	**Gradient Boosting**	**1.000000e+00**
Random Forest	Regressió Logística	1.188766e-07
Arbre de decisió	**Bagging**	**1.472794e-01**
Arbre de decisió	**Gradient Boosting**	**1.000000e+00**
Arbre de decisió	Regresió Logística	1.099561e-04
Arbre de decisió	**Random Forest**	**1.000000e+00**

Taula 14. *Comparacions múltiples de la mètrica Sensibilitat.*

Model A	Model B	P valor
Gradient Boosting	Bagging	7.783460e-09
Regressió Logística	Bagging	7.783460e-09
Regressió Logística	Gradient Boosting	7.783460e-09
Random Forest	Bagging	3.920997e-05
Random Forest	Gradient Boosting	7.783460e-09
Random Forest	Regressió Logística	7.783460e-09
Arbre de decisió	**Bagging**	**1.546038e-01**
Arbre de decisió	Gradient Boosting	8.701074e-03
Arbre de decisió	Regresió Logística	2.460601e-07
Arbre de decisió	**Random Forest**	**6.293344e-01**

Taula 15. *Comparacions múltiples de la mètrica Especificitat.*

A partir de les taules anteriors, es pot extreure la següent informació: referent a la variable AUC (***Taula 13***), totes les parelles de models són significativament diferents. Pel que fa a la Sensibilitat (***Taula 14***), tots els models presenten diferències significatives menys els models *Random Forest* – *Gradient Boosting*, Arbre de decisió – *Bagging*, Arbre de decisió – *Gradient Boosting* i Arbre de decisió – *Random Forest*. Finalment, referent a

l'Especificitat (***Taula 15***), totes les parelles presenten diferències significatives menys els models Arbre de decisió – *Bagging* i Arbre de decisió – *Random Forest*.

b. Conjunt Test

Tot i que mitjançant el mètode de validació *Cross-Validation* del conjunt d'entrenament (apartat a.) s'aconsegueixen estimacions molt bones de l'error, és convenient comprovar que no existeix sobre ajust o *overfitting*. Per tant, serà necessari avaluar el conjunt test amb els models que hem entrenat. Aquesta és la raó per la qual inicialment hem separat la base de dades resultant en el conjunt entrenament i test, mantenint-se aquest últim, aïllat de tot el procés de transformacions, entrenament i optimització.

A la ***Taula 11*** s'ha vist que tots els models presenten una **Exactitud** entre 0,7 i 0,8 on l'Arbre de decisió obté el valor més gran, seguit de *Gradient Boosting*, *Bagging*, *Random Forest* i Regressió Logística. En termes de **Precisió**, el model que pren un major valor és *Gradient Boosting*, Regressió Logística, *Random Forest*, l'Arbre de decisió i *Bagging*. Referent a la **Sensibilitat**, el model corresponent a l'Arbre de decisió és el que pren major valor, seguit del *Bagging*, *Gradient Boosting*, *Random Forest* i Regressió Logística. Finalment, respecte a l'**Especificitat**, el model que pren un valor més gran és *Gradient Boosting*, seguit de la Regressió Logística, *Random Forest*, Arbre de decisió i *Bagging*.

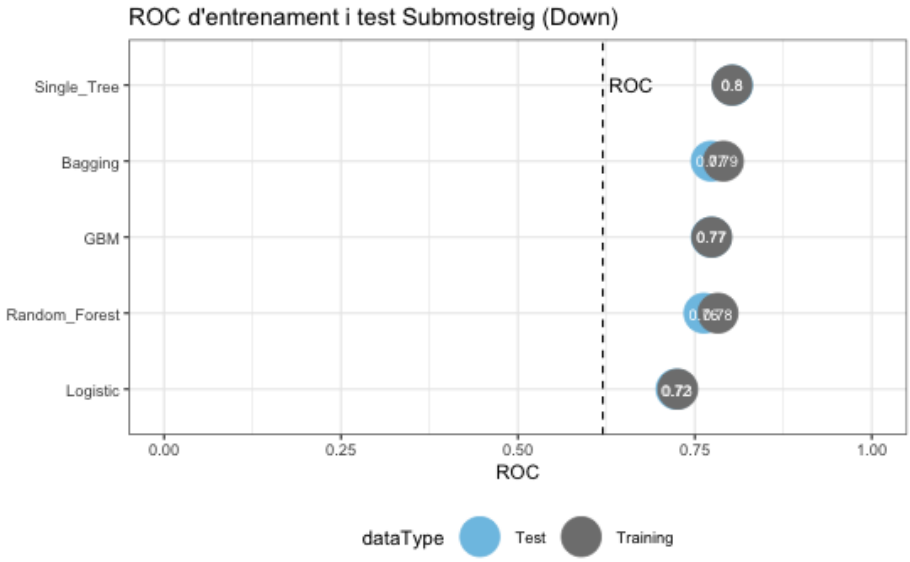
		Regressió Logística	Arbre de decisió	*Bagging*	*Random Forest*	*Gradient Boosting*
Sub mostreig	**Exactitud**	0.7235	0.8032	0.7730	0.7616	0.7732
	Precisió	0.9887	0.9882	0.9856	0.9884	0.9903
	Sensibilitat	0.7204	0.8048	0.7751	0.7608	0.7715
	Especificitat	0.7995	0.7633	0.7222	0.7802	0.8140

Taula 16. *Millors i Pitjors Mètriques de Validació de tots els models (test).*

En la ***Taula 16*** s'indiquen en verd, aquells valors que prenen un valor més alt i, en vermell, el valor més baix de cadascuna de les mètriques. Com a resultat, s'obté que els millors models són l'Arbre de decisió i *Gradient Boosting*.

	Gradient Boosting	***Random Forest***	**Regressió Logística**	***Bagging***	**Arbre de decisió**
Test	0.7732	0.7628	0.7235	0.7730	0.8032
Training	0.7740	0.7827	0.7258	0.7904	0.8026

Taula 17. *ROC test vs ROC entrenament.*



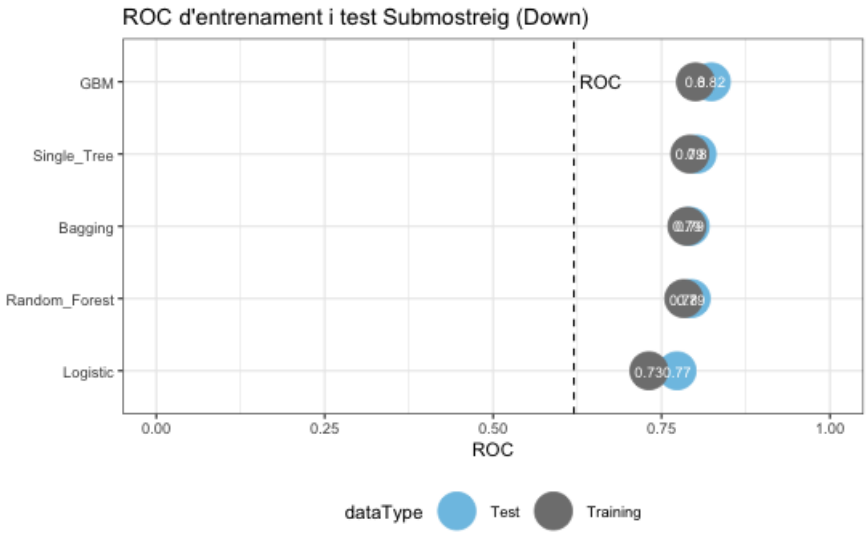
Gràfic 49. *ROC d'entrenament versus ROC test.*

Els models *Gradient Boosting*, *Bagging*, *Random Forest* i Regressió Logística, aconsegueixen majors prediccions correctes en el conjunt d'entrenament que no pas en el conjunt test, és a dir, són models optimistes. Tot i que la variància entre el conjunt d'entrenament i test és molt petita. En contra, l'Arbre de Decisió assoleix majors prediccions correctes en el conjunt test que no pas en el conjunt d'entrenament, és a dir, és un model pessimista. Anàlogament, aquesta deferència és minúscula. Finalment, el model que obté un major valor AUC en el test correspon a l'Arbre de Decisió.

Un possible problema a causa de l'agrupació de les dades referents al 2006, 2016 i 2020, és que les dades d'entrenament i test siguin molt semblants i, per tant, no permetin comprovar si el model és capaç de realitzar prediccions correctament sobre les dades d'un altre any. Com a solució, s'agafaran mostres d'un any i es comprovarà si el model d'aquest any prediu correctament les dades d'un altre any diferent. En el nostre cas, el que farem serà prendre les dades del 2006 i 2016 com a conjunt d'entrenament i, les del 2020 com a conjunt de test. D'aquesta manera es comprovarà si els millors models corresponen amb els resultats obtinguts anteriorment.

	Gradient Boosting	***Random Forest***	**Regressió Logística**	***Bagging***	**Arbre de decisió**
Test	0.8239336	0.7941223	0.7730853	0.7925366	0.8028966
Training	0.8003652	0.7920299	0.7304670	0.7877003	0.7312971

Taula 18*. ROC test (2020) vs ROC entrenament (2006, 2016).*



Gràfic 50. *ROC d'entrenament (2006, 2016) versus ROC test (2020).*

En aquest cas, tots els models aconsegueixen més prediccions correctes en el conjunt test que no pas en el conjunt d'entrenament, és a dir, que és són models pessimistes. Finalment, el model que obté un major valor AUC en el test correspon al model *Gradient Boosting.*

Amb tota aquesta informació, podem afirmar que donats els bons resultats del model ***Gradient Boosting,*** tant en el conjunt d'entrenament com en el del test, serà el millor per estudiar les prediccions de consum de carn a les llars espanyoles. Tot i que l'arbre de decisió en el conjunt test presenti bons resultats, és un algoritme molt inestable, ja que petits canvis en les dades, podrien canviar totalment l'arbre i, per tant, no ser capaç de predir adequadament.

3.2.5. Anàlisi i visualització de resultats

A partir dels models anteriors, hem de ser capaços d'extreure informació d'aquests. És per això que, en aquest apartat s'abordaran tres objectius principals; el primer d'ells serà conèixer la influència de les variables predictores sobre la variable resposta. El segon, predir el valor de la variable resposta en relació amb les característiques de les llars i finalment conèixer el perfil tant de les famílies que consumeixen carn com les que no i

també es determinarà quin és el perfil d'aquelles llars que el model no prediu correctament.

3.2.5.1.Influencia de les variables sobre la variable resposta

Per entendre la influència de les variables predictores sobre la variable resposta, que prendrà valor 1 si la llar consumeix carn i 0 en cas contrari, s'haurà d'utilitzar el model de Regressió Logística, ja que és l'únic capaç de quantificar la influència de cadascuna de les variables predictores sobre la variable resposta. La sortida del model és la següent:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6135  -0.8390  -0.1668   0.8950   2.1907

Coefficients:0
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.21214    0.04556   4.656 3.22e-06 ***
GASTOT                         1.44617    0.08242  17.546  < 2e-16 ***
SEXO                          -0.07847    0.04382  -1.791  0.07336 .
COMITOT                        0.47946    0.06033   7.947 1.91e-15 ***
EDAD_grupo_X.64..100.         -0.09365    0.06076  -1.541  0.12323
EDAD_grupo_X.0.30.            -0.15957    0.04869  -3.277  0.00105 **
DENSIDAD_X2                    0.05502    0.04529   1.215  0.22447
DENSIDAD_X3                    0.04269    0.05530   0.772  0.44017
CCAA_X2                       -0.03483    0.04848  -0.718  0.47247
CCAA_X3                        0.02730    0.04939   0.553  0.58044
NUMOCUP_X1                    -0.07995    0.06032  -1.326  0.18499
NUMOCUP_X2                     0.10011    0.06819   1.468  0.14205
NUMOCUP_X3                     0.04580    0.05592   0.819  0.41283
NUMOCUP_X4                     0.09023    0.07275   1.240  0.21488
IMPEXAC_grupo_X.1800..Inf.     0.03701    0.05418   0.683  0.49455
IMPEXAC_grupo_X.0..1100.      -0.11022    0.05198  -2.120  0.03397 *
ECIVILSP_X2                    0.23230    0.05927   3.919 8.89e-05 ***
ECIVILSP_X3                    0.09663    0.05381   1.796  0.07251 .
ECIVILSP_X4                    0.12283    0.04620   2.659  0.00784 **
ZONARES_X2                    -0.12626    0.05112  -2.470  0.01351 *
ESTUDIOS2_grupo_X.20...75..    0.22012    0.15537   1.417  0.15655
ESTUDIOS2_grupo_X.75...100..   0.29946    0.14082   2.127  0.03345 *
ESTUDIOS3_grupo_X.20...75..    0.10632    0.12080   0.880  0.37882
ESTUDIOS3_grupo_X.75...10..    0.21624    0.15343   1.409  0.15872
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4599.7  on 3317  degrees of freedom
Residual deviance: 3489.8  on 3294  degrees of freedom
AIC: 3537.8

Number of Fisher Scoring iterations: 5
```

Sempre suposant que la resta de variables romanen constant, la interpretació dels coeficients és la següent:

- **Interpretació del coeficient de la variable GASTOT, :** Observem que el valor del coeficient és positiu, per tant, com major sigui el valor de la despesa total, major serà la probabilitat de què la llar consumeixi carn. A més a més, un increment unitari del GASTOT, fa incrementar el *logit* (logaritme dels *odds*) de la variable resposta en 1,44617. Altrament dit, un augment unitari de la variable GASTOT suposa un augment dels *odds* de què la llar consumeixi carn del 4,25.
- **Interpretació del coeficient de la variable COMITOT, :** Observem que el valor del coeficient és positiu, per tant, com major sigui el nombre de menjars i sopars bisetmanals, major serà la probabilitat de què la llar consumeixi carn. A més a més, un increment unitari de COMITOT, fa incrementar el logit (logaritme del *odds*) de la variable resposta en 0,47946. Altrament dit, un augment unitari de la variable COMITOT suposa un augment dels *odds* de què la llar consumeixi carn de l'1,62.
- **Interpretació del coeficient de la variable EDAD_grupo_X.0.30., :** Sent el grup d'edat de referència l'interval (30,64] anys. Si l'edat mitjana de llar es troba entre els 0 i 30 anys, té 1,17 vegades menys probabilitats de consumir carn que si l'edat mitjana de la llar es troba entre els 31 i 64 anys.
- **Interpretació del coeficient de la variable IMPEXAC_grupo_X.0..1100, :** Sent la categoria de referència els ingressos nets totals de la llar de (1.100,1.800]. Si els ingressos nets mensuals totals percebuts per la llar, es troben entre 0 € i 1.100 €, la llar té 1,12 vegades menys probabilitats de consumir carn que si els ingressos nets mensuals totals de la llar es troben entre 1.100 € i 1.800 €.
- **Interpretació del coeficient de la variable ECIVILSP_X2, :** Sent l'estat civil del sustentador principal solter la categoria de referència. Si el sustentador principal de la llar es troba casat, té 1,26 vegades més probabilitats de consumir carn que si el sustentador principal es troba solter.
- **Interpretació del coeficient de la variable ECIVILSP_X4, :** Sent la categoria de referència cap ocupat. Si a la llar té quatre o més ocupats, aquesta té 1,13 vegades més probabilitats de consumir carn que si no hi ha cap ocupat.
- **Interpretació del coeficient de la variable ZONARES_X2, :** Sent la zona residencial urbana la categoria de referència. Si llar està situada en una zona rural, té 1,13 vegades menys probabilitats de consumir carn que si es troba en una zona urbana.

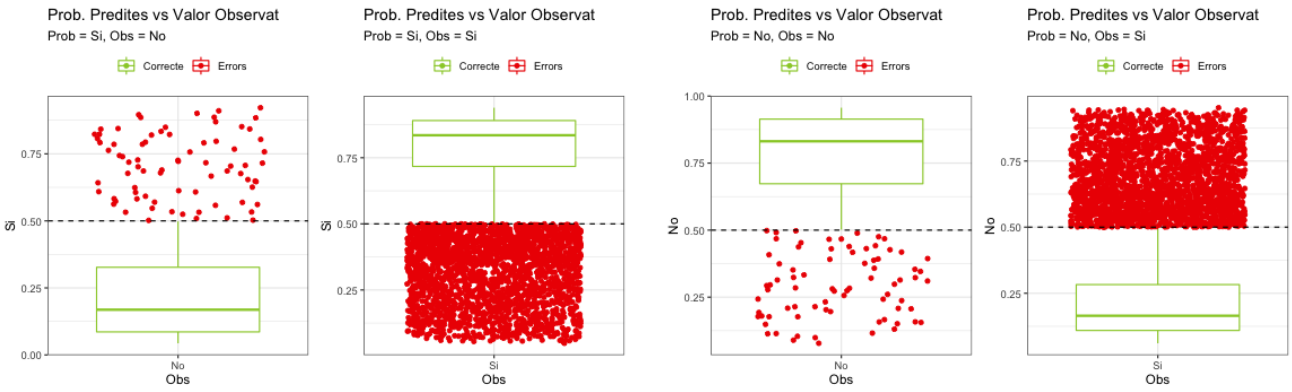
- **Interpretació del coeficient de la variable ESTUDIOS2_grupo_X.75...100., :** Sent la categoria de referència el percentatge d'individus de la llar entre (0,20] amb Estudis Obligatoris. Si llar es troba formada entre el 75% i 100% d'individus amb Estudis Obligatoris, té 1,35 vegades més probabilitats de consumir carn que si es troba entre el 0% i 20%.

Com a síntesi general, podríem dir que la probabilitat de consumir carn serà major que no consumir-ne, com major sigui la despesa, major sigui el nombre de dinars i sopars bisetmanals, el grup d'edat es trobi entre 31 i 64 anys, els ingressos nets mensuals totals siguin superiors als 1.800 €, el sustentador principal es trobi casat o bé separat o divorciat, que visqui en una zona urbana o el percentatge d'individus amb Estudis Obligatoris es trobi entre el 75% i el 100%. En l'altre extrem, considerem que la probabilitat de no consumir carn serà major com menor sigui la despesa, menor sigui el nombre de dinars i sopars bisetmanals, el grup d'edat estigui entre els 0 i 30 anys, els ingressos nets mensuals totals siguin inferiors als 1.100 €, l'estat civil del sustentador principal sigui solter, que la llar visqui en una zona rural o el percentatge d'individus amb Estudis Obligatoris estigui comprès entre el 0% i el 20%.

3.2.5.2.Predicció de la variable resposta

La importància de la interpretabilitat dels models de ML, és crucial per poder justificar i comprendre les prediccions i resultats obtinguts. En l'apartat de Comparació de Models, s'ha conclòs que el model *Gradient Boosting* era el millor model per predir si la llar consumiria carn o no. Per tant, l'anàlisi de resultats es realitzarà sobre el model escollit.

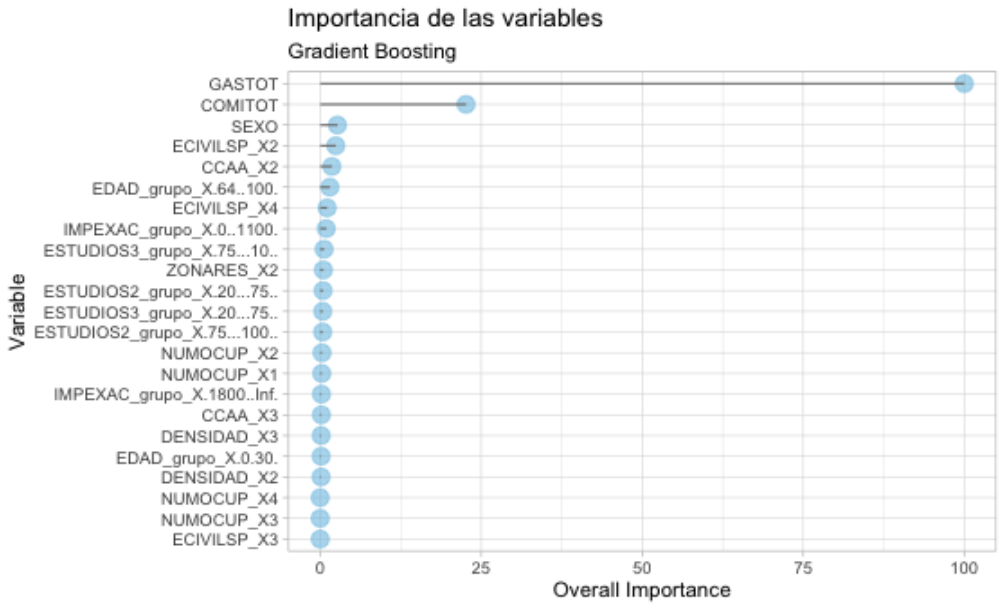
Tot seguit, a partir de les probabilitats predites de la variable resposta, es representaran gràficament quines observacions s'han classificat correctament com a 0 (no consumeixen carn), com a 1 (si consumeixen carn) i quines s'han classificat incorrectament (tant com a 0 i 1).



***Gràfic 51**. Probabilitats Predites versus Valors Observats.*

El ***Gràfic 51*** deixa veure com es distribueixen els errors i si aquests es troben molt llunyans al llindar del 0,5. Es pot percebre que aquelles llars que realment no consumeixen carn i la probabilitat de consumir carn és baixa, seran les observacions correctes i en mitjana la probabilitat de consumir carn és aproximadament de 0,15 (propera a 0). En canvi, aquelles llars que realment no consumeixen carn i la probabilitat de consumir carn és alta, seran les observacions incorrectes i es pot veure que aquestes probabilitats no segueixen cap patró en concret. Per altra banda, aquelles llars que realment sí que consumeixen carn i la probabilitat de consumir carn és elevada, seran les observacions correctes i en mitjana la probabilitat de consumir carn és d'aproximadament del 0,80. En canvi, aquelles llars que realment sí que consumeixen carn, però la probabilitat de consumir carn és baixa, seran les observacions incorrectes. Es nota que hi ha un major nombre d'observacions mal predites, però tampoc segueixen cap patró específic. Anàlogament, es pot aplicar el mateix raonament per les probabilitats predites de no consumir carn.

Encara que existeixin diferents estratègies de parametrització i optimització, la importància relativa de les variables és un concepte unificador que permet realitzar comparacions sobre les característiques entre els diferents models. La interpretació de l'indicador d'importància relativa de les variables es porta a cap en termes absoluts, és a dir, sense tenir compte ni el signe ni la direcció de l'efecte.



Gràfic 52. *Importància Relativa de les variables.*

El ***Gràfic 52*** deixa veure les variables més importants, i aquestes són: GASTOT, COMITOT, SEXO, ECIVILSP_X2, CCAA_X2, EDAD_grupo_X.64..100, ECIVILSP_X4 i IMPEXAC_grupo_X.0..1100. Seguidament, a partir del resum de les variables amb major importància, s'estudiaran les característiques de les llars que s'han classificat correctament (com a 0 i com a 1) i incorrectament (com a 0 i com a 1). Els resultats obtinguts es mostren a continuació:

1. Classificació correcte

1.1.Llars que consumeixen carn

El perfil d'aquelles llars que consumeixen carn tenen en mitjana, una despesa monetària i no monetària elevada temporalment i poblacionalment de 4.662.767 €, sent 430.842 € la despesa mínima de les llars i, 121.716.283 € la despesa màxima. En mitjana, la llar es troba composta pel 48,17% de dones i, cadascuna de les llars en mitjana realitza 75,44 dinars i sopars bisetmanals amb un nombre màxim de 280 dinars i sopars durant les dues setmanes. L'edat mitjana de la llar és d'entre 31 i 64 anys (la mitjana del grup d'edat entre 65 i 100 anys és de 0,37) i l'estat civil del sustentador principal és solter (ja que la mitjana de l'estat civil del sustentador principal separat o divorciat és de 0,08). Finalment, els ingressos mensuals nets totals de cada una de les llars oscil·la entre els 1.100 € i 1.800 € (la mitjana dels ingressos mensuals nets totals fins a 1.100 € és de 0,13).

1.2.Llars que no consumeixen carn

El perfil d'aquelles llars que no consumeixen tenen en mitjana, una despesa monetària i no monetària elevada temporalment i poblacionalment de 583.811 €, sent 2.514 € la despesa mínima de les llars i, 4.545.001 € la despesa màxima. Referent al percentatge de dones, en mitjana la llar està constituïda pel 46,07% dones i en mitjana cada llar realitza 38,57 dinars i sopars cada dues setmanes, sent 308 el màxim nombre de dinars i sopars realitzats durant dues setmanes. L'edat mitjana de la llar és d'entre 31 i 64 anys (la mitjana del grup d'edat entre 65 i 100 anys és de 0,35) i l'estat civil del sustentador principal és solter (la mitjana de l'estat civil del sustentador principal separat o divorciat és de 0,1). Finalment, els ingressos mensuals nets totals de cada una de les llars oscil·la entre els 1.100 € i 1.800 € (la mitjana dels ingressos mensuals nets totals fins a 1.100 € és de 0,23).

Si es comparen els dos perfils, respecte a la variable despesa monetària i no monetària elevada temporalment i poblacionalment en mitjana és 7,99 vegades més alta en les llars que consumeixen carn que en les que no en consumeixen. Referent al percentatge de dones de la llar, és lleugerament superior en les llars que mengen carn. A més, realitzen el doble nombre de dinars i sopars respecte a les llars que no consumeixen carn. Amb referència a l'edat mitjana, l'estat civil del sustentador principal i els ingressos mensuals nets totals és el mateix pels dos tipus de perfils. Pel que fa al grup d'edat major de seixanta-quatre anys, és quasi igual de probable que consumeixi o no carn, ja que la probabilitat de què un individu major de seixanta-quatre anys consumeixi carn és de 0,37, mentre que és de 0,35, si no en consumeix. Amb l'estat civil del sustentador principal, separat o divorciat, passa exactament el mateix. La probabilitat de què el sustentador principal es trobi separat o divorciat i consumeixi carn és de 0,08, mentre que és de 0,1, si no en consumeix. Finalment, la probabilitat de què els ingressos mensuals nets totals de la llar oscil·lin entre els 0 € i 1.100 € i consumeixi carn és de 0,13 mentre que és de 0,23, si no consumeix.

2. Classificació incorrecte

2.1.Llars que consumeixen carn

El perfil d'aquelles llars que consumeixen carn tenen en mitjana, una despesa monetària i no monetària elevada temporalment i poblacionalment és d'1.089.531 €, sent 9.525 € la despesa mínima de les llars i, 5.779.972 € la despesa màxima. En mitjana, la llar es troba composta pel 44,97% de dones i, cadascuna de les llars en mitjana realitza 36,88 dinars i

sopars bisetmanals amb un nombre màxim de 239 dinars i sopars durant les dues setmanes. L'edat mitjana de la llar és d'entre 31 i 64 anys (la mitjana del grup d'edat entre 65 i 100 anys és de 0, 35) i l'estat civil del sustentador principal és solter (ja que la mitjana de l'estat civil del sustentador principal separat o divorciat és de 0,2). Finalment, els ingressos mensuals nets totals de cada una de les llars oscil·la entre els 1.100 € i 1.800 € (la mitjana dels ingressos mensuals nets totals fins a 1.100 € és de 0,35).

2.2.Llars que no consumeixen carn

El perfil d'aquelles llars que no consumeixen tenen en mitjana, una despesa monetària i no monetària elevada temporalment i poblacionalment de 2.801.768 €, sent 499.686 € la despesa mínima de les llars i, 19.671.272 € la despesa màxima. Referent al percentatge de dones, en mitjana la llar està constituïda pel 45,32% dones i en mitjana cada llar realitza 64,75 dinars i sopars cada dues setmanes, sent 168 el màxim nombre de dinars i sopars realitzats durant dues setmanes. L'edat mitjana de la llar és d'entre 31 i 64 anys (la mitjana del grup d'edat entre 65 i 100 anys és de 0,30) i l'estat civil del sustentador principal és solter (la mitjana de l'estat civil del sustentador principal separat o divorciat és de 0,11). Finalment, els ingressos mensuals nets totals de cada una de les llars oscil·la entre els 1.100 € i 1.800 € (la mitjana dels ingressos mensuals nets totals fins a 1.100 € és de 0,23).

Si es comparen els dos perfils, respecte a la variable despesa monetària i no monetària elevada temporalment i poblacionalment en mitjana 2,6 vegades superior en les llars que no consumeixen carn que no pas les que sí que consumeixen. Referent al percentatge de dones de la llar, és lleugerament superior en les llars que no consumeixen carn. A més, realitzen 1,8 vegades més dinars i sopars respecte a les llars que consumeixen carn. Amb referència a l'edat mitjana, l'estat civil del sustentador principal i els ingressos mensuals nets totals és el mateix pels dos tipus de perfils. Pel que fa al grup d'edat major de seixanta-quatre anys, és quasi igual de probable que consumeixi o no carn, ja que la probabilitat de què un individu major de seixanta-quatre anys consumeixi carn és de 0,35, mentre que és de 0,30, si no en consumeix. Amb l'estat civil del sustentador principal, separat o divorciat, la probabilitat de què el sustentador principal es trobi separat o divorciat i consumeixi carn és de 0,2, mentre que és de 0,1, si no en consumeix. Finalment, la probabilitat de què els ingressos mensuals nets totals de la llar oscil·lin entre els 0 € i 1.100 € i consumeixi carn és de 0,35 mentre que és de 0,23, si no consumeix.

3.3. Problema de regressió: modelització dels costos ocults de les llars espanyoles

L' objectiu d'aquest problema consisteix en determinar quins són els factors que expliquen el cost ocult de les llars espanyoles consumidores de carn i, quina influència tenen sobre aquest. Per tant, la **variable resposta** és defineix com el cost medi ambiental en euros per aquelles llars que consumeixen carn elevat temporalment i poblacionalment i les **variables explicatives** són les mateixes que les de l'apartat anterior sense tenir en compte la variable GASTOT ja que la variable cost ocult ha estat calculada amb aquesta i per tant és evident que a més GASTOT, més cost ocult/medi ambiental.

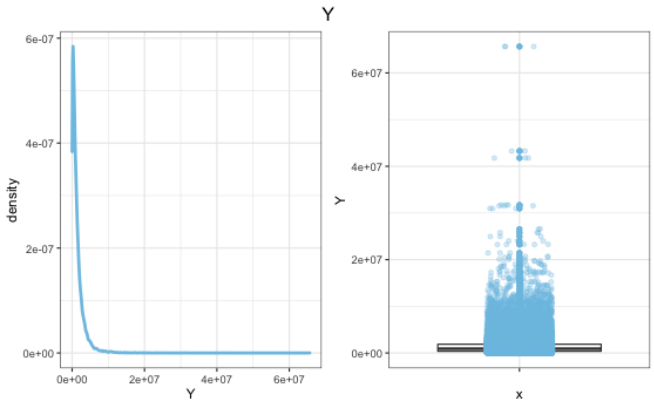
Per calcular la variable resposta s'han utilitzat tant la Base de Dades I com la Base de Dades II i, aquesta es defineix com:

$$\boldsymbol{Y = \text{PREU\_KG\_MA} * \text{PES}}$$

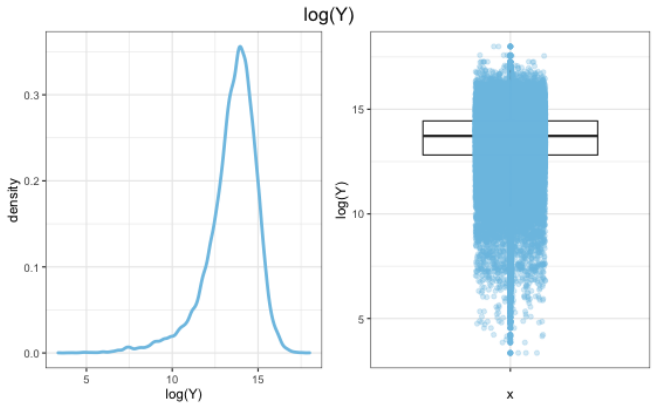
$$On\ PES = \frac{GASTOT}{\text{PREU\_KG\_CARN}}$$

3.3.1. Anàlisi exploratòria de les dades

Les metodologies aplicades en l'àmbit exploratori de dades són el primer pas abans d'aplicar les tècniques estadístiques avançades, com poden ser l'estadística inferencial o l'aprenentatge automàtic. Per aquesta raó, s'utilitzaran tècniques gràfiques per conèixer el comportament tant de la variable resposta com el de les variables explicatives.



Gràfic 53. *Corba de densitat i boxplot de la variable Y.*



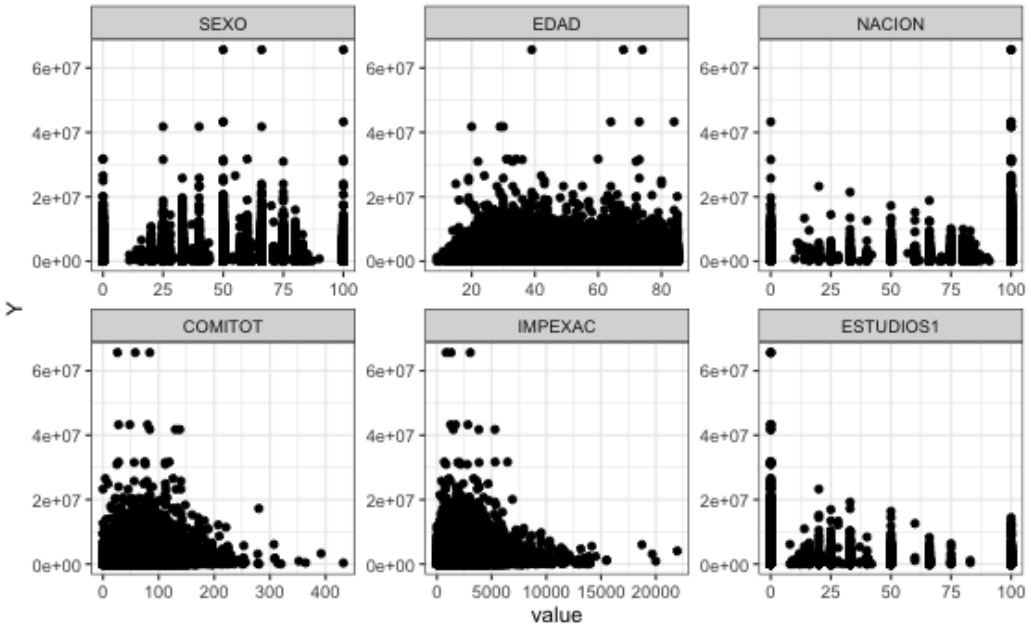
Gràfic 54. *Corba de densitat i boxplot de la variable log(Y).*

El ***Gràfic 53*** mostra que la variable resposta té una distribució asimètrica, on moltes llars tenen un cost relativament baix i unes poques tenen un cost extremadament elevat. Aquest tipus de distribució se sol visualitzar millor després de realitzar una transformació logarítmica tal com es mostra al ***Gràfic 54,*** en el qual es pot observar que majoritàriament la variable pren valors entre 22.026,47 € i 3.269.017 €.

El comportament de les variables explicatives ja ha estat estudiat en el problema de classificació. A diferència de la qüestió tractada en l'apartat anterior les variables contínues no es categoritzaran en intervals, sinó que es tractaran com a variables numèriques. Per tant, pel que fa a la correlació entre les variables EDAD:NMIEMBR3, ESTUDIOS3:ESTUDIOS2, NMIEMBR2:NMIEMBR3, EDAD:NMIEMBR1, no es tindran en compte les variables NMIEMBR i referent a la correlació entre la variable ESTUDIOS2 i ESTUDIOS3, es crearà una nova variable ESTUDIOS que recollirà la informació de totes dues i prendrà els següents valors:

$$ESTUDIOS = \begin{cases} 1 & Si\ majoritàriament\ la\ llar\ té\ estudis\ obligatoris \\ 2 & Si\ majoritàriament\ la\ llar\ té\ estudis\ superiors \\ 3 & Mateixa\ proporció\ d'individus\ a\ la\ llar\ amb\ estudis\ obligatoris\ i\ superiors \end{cases}$$

Seguidament, es creuaran les variables continues amb la variable resposta per estudiar el seu comportament.



Gràfic 55. *Gràfics de dispersió de les variables contínues versus la variable resposta.*

Gràficament, la variable resposta no presenta una relació lineal amb cap de les variables predictores. Referent a les variables COMITOT i IMPEXAC sembla que presenten un comportament molt similar. També s'observa que hi ha certs valors de les variables els quals es troben molt allunyats de la resta de dades.

3.3.2. Preparació de les dades

En estadística, els valors atípics o *outliers*, són observacions que disten molt de la resta del conjunt de dades. És a dir, un valor atípic és valor anòmal i és extremadament diferent de la resta de valors. És important identificar els valors atípics (o *outliers*) d'una mostra, ja que poden afectar considerablement el càlcul de les mesures estadístiques. Per aquesta raó, gràcies a la gran dimensió de dades, decidirem eliminar els *outliers*.

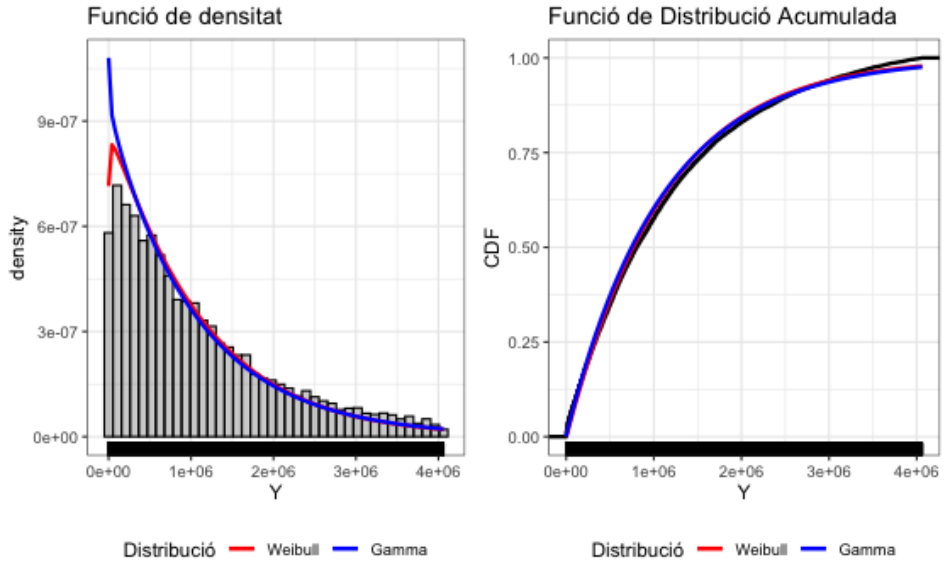
Anteriorment, s'ha comprovat que les variables NACION i ESTUDIOS1 presentaven una variància propera a zero, per tant, en aquest cas tampoc les introduirem en el model perquè no aporten informació rellevant.

Per acabar, tal com s'ha esmentat en el problema anterior, separarem la base de dades en el conjunt d'entrenament i test, on la partició serà del 80%-20% respectivament.

3.3.3. Distribució de la variable resposta

Ajustar una distribució paramètrica a partir d'un conjunt de dades, consisteix a trobar el valor dels paràmetres amb els quals, amb major probabilitat, la distribució pot haver generat les dades. Per exemple, la distribució normal té dos paràmetres (mitjana i variància), una vegada coneguts aquests dos, es coneix tota la distribució. Per tant, per escollir la família a la qual pertany la variable resposta serà necessari saber quina distribució segueix. Per això, l'ajustarem a alguna distribució coneguda fent ús del criteri d'AIC i BIC. Els resultats obtinguts sobre la distribució de la variable resposta estan referenciats a *l'Annex 4.*

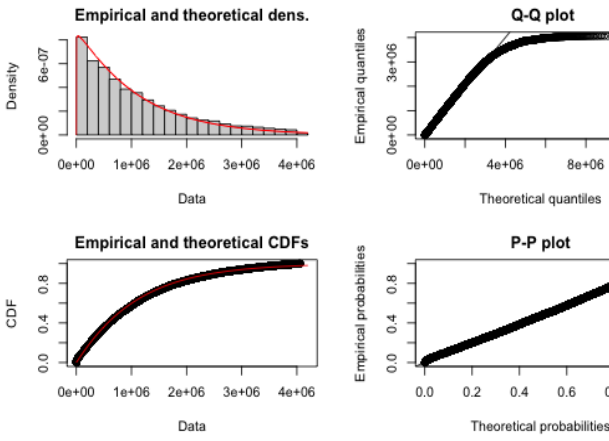
Com el nostre objectiu és explicar i no predir el cost ocult de les famílies que consumeixen carn, es farà servir el criteri d'AIC. La distribució escollida serà aquella que tingui menor valor d'AIC. Les dues distribucions que millor s'ajusten a variable resposta són les distribucions **Weibull i Gamma.**



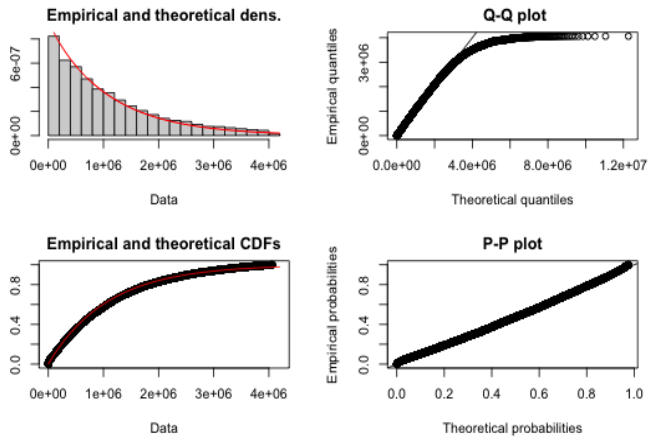
Gràfic 56. *Funció de densitat i Funció de Distribució Acumulada de la variable resposta.*

El ***Gràfic 56*** mostra l'ajust de les dues distribucions sobre el conjunt de dades i tal com s'ha comprovat analíticament, la distribució Weibull s'ajusta lleugerament millor que no pas la distribució Gamma.

A continuació, es procedirà a realitzar la validació de la distribució mitjançant la funció *fitdist* del paquet *fitdistrplus*, el qual estimarà els paràmetres de forma i escala pel criteri de Màxima Versemblança (*mle*) de la distribució Weibull i Gamma obtenint així els següents resultats:



Gràfic 57. *Validació de la distribució Weibull.*



Gràfic 58. *Validació de la distribució Gamma.*

Gràficament, s'observa que totes dues distribucions s'ajusten bastant bé a la variable resposta, concretament la cua de la dreta no s'ajusta del tot bé, això és degut al fet que la variable resposta pren pocs valors molt elevats, el que fa que les dades no s'acabin

d'ajustar perfectament. Com a conseqüència, quan es realitza el test Kolmogorov-Smirnov**,** ens porta a rebutjar la hipòtesi nul·la sobre que les dades provenen d'una distribució Weibull o Gamma. Per tant, només ens centrarem en la validació gràfica i donarem per vàlides totes dues distribucions.

3.3.4. Modelització

Els models GAMLSS assumeixen que una funció de densitat de la variable resposta pot estat definida fins a quatre paràmetres, els quals determinen la seva posició, escala i forma. Cada un d'ells pot variar independentment dels altres en funció dels predictors. Aquests tipus de models permeten establir fins a 4 funcions, on cada una, estableix la relació entre un dels paràmetres i les variables predictores. Els models GAMLSS són capaços de caracteritzar la distribució completa. A més, permeten obtenir intervals probabilístics i prediccions sense assumir que la variància és constant ni que les relacions són únicament lineals.

La implementació en *RStudio* dels models GAMLSS es troba disponible en el paquet *gamlss*. Per parametritzar la distribució Weibull o Gamma només és necessari estimar els paràmetres mitjana i escala. Els primers dos arguments de la funció corresponen amb l'estimació dels paràmetres en funció de les variables predictores i, el tercer argument correspon amb la família de la distribució, que en aquest cas serà Weibull (WEI) o Gamma (GA). Totes dues amb funció d'enllaç *log*, tant en el paràmetre mitjana com en l'escala.

La funció *gamlss* permet introduir *p-splines* (*Penalized Smoothing Splines*) als predictors continus del model. Aquests permeten modelar de forma lliure amb transicions suaus i contínues aportant una major flexibilitat al model. Com a conseqüència, la interpretació dels coeficients queda invalidada, ja que la relació entre la variable explicativa i la variable resposta no és lineal. Com el nostre objectiu principal no és realitzar prediccions sinó que, és determinar l'efecte de les variables explicatives sobre la variable resposta, només ens centrarem en el model lineal generalitzat simple.

La validació de regressió implica comprovar si els resultats que quantifiquen les relacions hipotètiques entre variables obtingudes de l'anàlisi de regressió, són acceptables. Aquest procés comporta analitzar si els residus de la regressió són aleatoris o bé segueixen algun

patró determinat, analitzar la bondat de l'ajust de la regressió i verificar si el rendiment predictiu del model es deteriora quan s'apliquen les dades que no s'han utilitzat per a l'estimació d'aquest, és a dir, les dades del conjunt test.

1. Bondat d'ajust

Una mesura de bondat d'ajust és el coeficient de determinació ajustat o R^2 ajustat el qual permet determinar que tan bé s'ajusten les dades a la regressió. Els resultats obtinguts es mostren a continuació:

	R^2 - ajustat
Weibull	0.01039332
Gamma	0.01039559

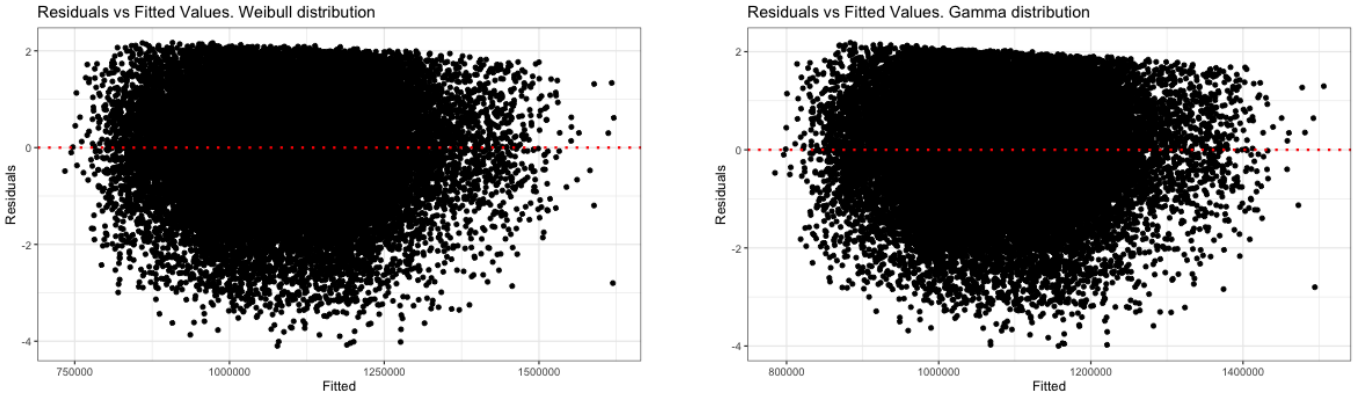
Taula 19. *R2 – ajustat.*

Com s'observa, tots dos models prenen valors molt baixos i similars entre ells. Per tant, les variables explicatives utilitzades són capaces d'explicar aproximadament un 1% del comportament de la variable resposta.

2. Anàlisi dels residus

Els residus d'un model ajustat es defineixen com les diferències entre les prediccions de la variable resposta utilitzant la funció de regressió i les respostes observades en cada combinació de valors de les variables explicatives. Si el model ajustes correctament les dades, el comportament dels residus seria totalment aleatori. En canvi, si no fos capaç d'ajustar les dades, la distribució dels residus no seria aleatòria, sinó que seguiria un patró.

Una manera de comprovar-ho és mitjançant un examen visual dels residus predits versus els valors ajustats, tal com es mostra al ***Gràfic 59.***



Gràfic 59*. Residus predits versus valor ajustats.*

Gràficament, s'observa que tant els residus de la distribució Weibull com Gamma es comporten de manera aleatòria, però a la part superior dreta es pot veure una lleugera tendència decreixent. Per aquesta raó, caldrà analitzar l'heteroscedasticitat dels residus mitjançant l'estadístic Durbin-Watson o Breuch Pagan, on sota la hipòtesi nul·la que no existeix autocorrelació dels residus, l'estadístic de prova pren un valor de DW = 1,9036 en el cas de la distribució Weibull i Gamma. Aquest valor és molt proper a 2, la qual cosa ens porta a no rebutjar hipòtesi nul·la. Per tant, tenim proves suficients per afirmar que els residus no estan autocorrelacionats.

3. Avaluació del model fora de la mostra i bondat d'ajust

Per avaluar el model, es faran ús de les mètriques de regressió sobre el conjunt d'entrenament i test, permetent detectar si el model és deficient o no.

		R^2	RMSE	MAE
Training	**Model Weibull**	0.0117342	946349.8	759015.2
	Model Gamma	0.0117117	945924.1	756305.6
Test	**Model Weibull**	0.0100631	942818.2	756216.4
	Model Gamma	0.0099640	942100.5	753478.9

Taula 20. *Mètriques de regressió.*

En ambdós casos, el RMSE del conjunt d'entrenament pren un valor superior al del conjunt test i, per tant, el model és eficient, però si s'analitza el R^2, s'observa que, per una banda, aquest valor és extremadament baix i per l'altra banda, que el valor del conjunt d'entrenament és lleugerament superior al del conjunt test, però la diferència entre tots dos és molt petita.

Com s'ha mencionat anteriorment, el nostre objectiu no es basa a predir ni obtenir un coeficient d'ajust elevat, sinó en la interpretació del model, és a dir, veure de quina manera influeixen les variables explicatives sobre la variable resposta. És per aquesta raó, que posarem un major èmfasi en la interpretació dels coeficients del model, en comptes d'avaluar la capacitat predictiva i el bon ajust del model.

3.3.5. Interpretació de resultats

A partir dels models anteriors, s'ha comprovat que el **model Weibull** s'ajusta millor a les dades i, per tant, serà l'escollit per explicar com afecten les variables predictores a la

variable resposta i, d'aquesta manera trobar el perfil d'aquelles llars que tenen uns majors costos ocults.

La sortida del programa referent al **model Weibull** es mostra a continuació:

```
******************************************************************
Family:  c("WEI", "Weibull")

Call:  gamlss(formula = Y ~ SEXO + EDAD + COMITOT + IMPEXAC + DENSIDAD
+
    CCAA + NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS, sigma.formula = Y
~
    SEXO + EDAD + COMITOT + IMPEXAC + DENSIDAD + CCAA + NUMOCUP +
        ECIVILSP + ZONARES + ESTUDIOS, family = WEI, data =
datos_train_prep,      trace = FALSE)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.363e+01  3.757e-02 362.691  < 2e-16 ***
SEXO         4.976e-04  2.069e-04   2.405 0.016195 *
EDAD         1.933e-03  4.352e-04   4.442 8.93e-06 ***
COMITOT      2.115e-03  2.092e-04  10.108  < 2e-16 ***
IMPEXAC      2.312e-05  6.536e-06   3.538 0.000404 ***
DENSIDAD2   -1.789e-02  1.312e-02  -1.363 0.172752
DENSIDAD3   -4.502e-02  1.476e-02  -3.050 0.002290 **
CCAA2       -5.733e-03  1.236e-02  -0.464 0.642734
CCAA3       -1.451e-01  1.334e-02 -10.876  < 2e-16 ***
NUMOCUP1     2.887e-02  1.581e-02   1.826 0.067809 .
NUMOCUP2     4.339e-02  1.964e-02   2.210 0.027131 *
NUMOCUP3     1.723e-01  3.244e-02   5.312 1.09e-07 ***
NUMOCUP4     1.181e-01  7.179e-02   1.645 0.099908 .
ECIVILSP2    4.199e-02  1.599e-02   2.625 0.008661 **
ECIVILSP3    4.148e-02  2.167e-02   1.915 0.055564 .
ECIVILSP4   -1.575e-02  2.196e-02  -0.717 0.473201
ZONARES2    -2.833e-02  1.564e-02  -1.811 0.070137 .
ESTUDIOS2   -2.501e-02  1.335e-02  -1.873 0.061080 .
ESTUDIOS3    6.099e-04  1.587e-02   0.038 0.969349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.471e-01  2.821e-02  -5.212 1.88e-07 ***
SEXO         4.393e-04  1.541e-04   2.850  0.00437 **
EDAD         1.342e-03  3.308e-04   4.058 4.97e-05 ***
COMITOT      1.255e-03  1.612e-04   7.781 7.40e-15 ***
IMPEXAC      1.245e-06  5.071e-06   0.245  0.80614
DENSIDAD2   -1.011e-02  1.014e-02  -0.997  0.31855
DENSIDAD3   -2.576e-02  1.121e-02  -2.298  0.02156 *
CCAA2        1.502e-02  9.743e-03   1.541  0.12325
CCAA3       -8.252e-02  1.000e-02  -8.251  < 2e-16 ***
NUMOCUP1     1.342e-02  1.201e-02   1.118  0.26354
```

```
NUMOCUP2      3.349e-02  1.510e-02   2.218  0.02654 *
NUMOCUP3      7.554e-02  2.660e-02   2.840  0.00451 **
NUMOCUP4      3.107e-02  5.765e-02   0.539  0.58990
ECIVILSP2     3.872e-02  1.197e-02   3.234  0.00122 **
ECIVILSP3     3.381e-02  1.623e-02   2.083  0.03725 *
ECIVILSP4     1.014e-02  1.605e-02   0.631  0.52779
ZONARES2     -1.602e-03  1.190e-02  -0.135  0.89294
ESTUDIOS2    -7.974e-03  1.025e-02  -0.778  0.43652
ESTUDIOS3    -5.125e-03  1.219e-02  -0.420  0.67432
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

----------------------------------------------------------------
No. of observations in the fit:  35398
Degrees of Freedom for the fit:  38
      Residual Deg. of Freedom:  35360
                      at cycle:  5

Global Deviance:     1054004
            AIC:     1054080
            SBC:     1054402
****************************************************************
```

Sempre suposant que la resta de variables romanen constants, la interpretació dels coeficients sobre la mitjana és la següent:

- **Interpretació del coeficient de la variable SEXO,** . S'observa que el valor del coeficient és positiu, per tant, com major sigui el percentatge de dones a la llar, major serà el cost mediambiental. Altrament dit, davant d'un augment de l'1% de dones a la llar, suposa un augment del cost mediambiental de 1,0005 €.
- **Interpretació del coeficient de la variable EDAD,** S'observa que el valor del coeficient és positiu, per tant, com major sigui l'edat mitjana de la llar, major serà el cost mediambiental. Altrament dit, per cada any mitjà addicional de la llar, suposa un augment del cost mediambiental de 1,0012 €.
- **Interpretació del coeficient de la variable COMITOT,**. S'observa que el valor del coeficient és positiu, per tant, com major sigui el nombre de dinars i sopars bisetmanals, major serà el cost mediambiental. Altrament dit, per un augment unitari del nombre de dinars i sopars bisetmanal de la llar, suposa un increment del cost mediambiental de 1,0020 €.
- **Interpretació del coeficient de la variable IMPEXAC,** S'observa que el valor del coeficient és positiu, per tant, com major siguin els ingressos totals nets de la llar, major serà el cost mediambiental. Altrament dit, per cada euro de més que ingressi la llar mensualment, suposa un augment del cost mediambiental d'un 1,0000 €.
- **Interpretació del coeficient de la variable DENSIDAD3 ,** Les llars que viuen en aquelles zones disseminades tenen un diferencial de -4,520e-02, és a dir, en mitjana

el cost medi ambiental serà 1,05 vegades inferior respecte les llars que viuen en zones densament poblades.

- **Interpretació del coeficient de la variable CCAA3 ,** Les llars que viuen en aquelles CCAA amb un PIB per càpita menor de 20.000 € tenen un diferencial de -1,396e-01, és a dir, en mitjana el cost mediambiental serà 1,16 vegades inferior respecte les llars que viuen en CCAA amb un PIB per càpita superior als 25.000 €.
- **Interpretació del coeficient de la variable NUMOCUP2,** Les llars en què hi ha 2 ocupats, tenen un diferencial de 4,339e-02, és a dir, en mitjana el cost mediambiental serà 1,04 vegades superior respecte les llars on no hi ha cap ocupat.
- **Interpretació del coeficient de la variable NUMOCUP3,** Les llars en què hi ha 3 ocupats, tenen un diferencial de 1,723e-01, és a dir, en mitjana el cost mediambiental serà 1,19 vegades superior respecte les llars on no hi ha cap ocupat.
- **Interpretació del coeficient de la variable ECIVILSP2 ,** Les llars on el sustentador principal està casat tenen un diferencial de 4,199e-02, és a dir, en mitjana el cost mediambiental serà 1,04 vegades superior respecte les llars on el sustentador principal està solter..

Per tant, a major percentatge de dones a la llar, edat mitjana, nombre de dinars i sopars bisetmanals o ingressos nets, si les llars viuen en zones densament poblades, en CCAA on el PIB per càpita és superior als 25.000 €, hi ha 2 o 3 ocupats a la llar o el sustentador principal es troba casat, majors seran els costos ocults associats a la llar.

3.3.6. Utilitat i aplicació de resultats

A partir dels resultats obtinguts durant l'estudi, s'establiria el plantejament d'unes polítiques destinades a pal·liar el canvi climàtic. Per dur-ho a terme, s'han de tenir en compte els perfils de llars als quals van dirigides aquestes polítiques i, els factors que determinen uns majors costos ocults. Les iniciatives proposades es mostren a continuació.

- **Canviar l'enfocament de les guies nutricionals i formar al personal sanitari**: Actualment, les guies nutricionals es basen principalment en el consum de fruites i verdures, hidrats i proteïna (normalment d'origen animal), l'alternativa correspondria en prioritzar les fonts de proteïna vegetal en comptes de l'animal. Referent al personal sanitari, s'establirien estratègies per proporcionar-los informació sobre dietes basades en aliments vegetals, amb l'objectiu de què s'introdueixin les llars espanyoles i així

aconseguir reduir el consum de carn i les conseqüències que en deriven. Així mateix, el personal sanitari és el mitjà de comunicació més ampli, ja que és capaç d'accedir a totes les edats, des de nens petits fins a generacions d'edat avançada.

- **Subvencionar a aquelles empreses disposades a innovar en la producció del sector carni:** una subvenció per part de l'Estat a aquelles empreses del sector, que estiguin disposades a implementar alternatives que promoguin una producció més sostenible, com és la reutilització de les aigües o dels gasos d'efecte hivernacle, afavorint així noves innovacions i a impulsar l'emprenedoria. Com a resultat, donaria lloc a una reducció dels costos ocults i, per tant, a un menor impacte mediambiental.

- **Introduir un nou impost sobre els productes carnis o bé rebaixar el preu dels productes vegetals:** la idea d'introduir un nou impost que gravi els productes d'origen animal, recau principalment en destinar aquesta recaptació en la investigació de noves alternatives per pal·liar el canvi climàtic en el sector ramader i, conscienciar a la població de l'impacte mediambiental generat pels productes carnis. Més concretament, l'impost gravat podria ser diferent segons la petjada ambiental del tipus de carn, ja que aquesta és major en la carn de boví que no pas en els altres tipus d'espècies. D'aquesta manera, el consum de carn en les llars espanyoles es veuria reduït, contribuint així a frenar el canvi climàtic. Per altra banda, reduir els preus de productes vegetals, incentivaria a les llars a substituir els productes d'origen animal pels d'origen vegetals, donant lloc a una reducció dels costos ocults.

- **Realitzar campanyes publicitàries, enfocades a aquelles famílies que consumeixin carn i tinguin majors costos ocults, les quals promoguin productes substitutius de la carn:** la introducció de nous productes com el tofu, heura, seitan o hamburgueses vegetals, d'entre altres, són una molt bona alternativa com a substitut de la carn. Si les campanyes publicitàries que promoguessin aquests tipus de productes, estiguessin enfocades a famílies que consumissin carn, aquestes s'animarien a introduir-los a la seva dieta. D'aquesta manera, reduirien o erradicarien el seu consum de carn i, com a conseqüència, els costos ocults es veurien disminuïts.

Amb l'aplicació de les alternatives destinades a un menor consum de carn, no només es reduiria la contaminació lligada al sector ramader, sinó que també tindria efectes positius sobre l'animal i la salut d'aquelles llars que tinguessin un consum excessiu de carn.

Actualment, amb el pas dels sistemes productius extensius als intensius, ha donat lloc a la substitució de pastures per sistemes de tancament, on els espais són molt reduïts i no permeten a l'animal moure's lliurement. A més, se'ls subministren una elevada quantitat d'hormones sintètiques amb l'objectiu d'obtenir un ràpid creixement. Com a conseqüència, donen lloc a malalties i malformacions. Per aquesta raó, va néixer la zootècnica, la qual s'encarrega de la producció d'animals tenint en compte el seu benestar. Per tant, una reducció del consum de carn, implicaria, per una banda, que menys animals haurien de viure en condicions extremes, i per l'altre, que un menor nombre d'animals hauria de ser sacrificat. Això, podria afavorir a un retorn dels sistemes productius extensius on l'animal fos capaç de desenvolupar-se en un ecosistema natural.

Per altra banda, tot i que la carn té efectes positius sobre la salut, un consum excessiu d'aquesta, podria induir a malalties cardiovasculars. Amb la reducció en el consum de carn, podria millorar la salut d'aquelles llars que consumeixen productes carnis en excés.

4. CONCLUSIONS

Al llarg de tot el treball hem arribat a diferents conclusions a través de diversos models i de l'anàlisi gràfica. A continuació, farem una breu síntesi dels resultats més rellevants que hem obtingut, per tal d'identificar el perfil de les llars espanyoles segons si consumeixen carn o no i, d'estudiar l'efecte de diverses variables sobre els costos ocults d'aquelles llars que consumeixen carn. Amb tota aquesta informació, s'il·lustraran algunes alternatives proposades per pal·liar els efectes mediambientals.

Gràcies a l'anàlisi gràfica de la variable resposta binària, consum de carn, ens ha permès detectar el problema de desequilibri de classes, on el 4% de les llars no consumien carn mentre que el 96% sí. Com a resposta, per tal d'equilibrar les proporcions, ha estat necessari modificar la distribució de les classes i, d'aquesta manera fer que el classificador de l'algoritme no tendeixi a estar esbiaixat cap a la classe majoritària. Al llarg del treball, s'ha conclòs que la millor estratègia era el submostreig el qual consistia a eliminar aleatòriament les observacions de la classe més nombrosa, és a dir, aquelles les llars que sí que consumien carn.

Del preprocessament de la base de dades, s'ha obtingut que les variables relacionades amb la nacionalitat del sustentador principal i el percentatge d'individus sense estudis primaris, presentaven una variància propera a zero. Com a conseqüència, han estat eliminades del model, ja que afegien més soroll que no pas informació.

Amb el **model logístic**, hem conclòs que la probabilitat de què la llar consumeixi carn serà major, com major sigui la despesa, més gran sigui el nombre de dinars i sopars bisetmanals, l'edat mitjana es trobi entre els 31 i 64 anys, els ingressos nets totals percebuts per la llar siguin superiors als 1.800 € mensuals, que el sustentador principal es trobi casat o bé separat o divorciat, que visqui en una zona urbana i que el percentatge d'individus de la llar amb estudis obligatoris es trobi entre el 75% i el 100%. Contràriament, la probabilitat de consumir carn serà menor, com més baixa sigui la despesa, més petit sigui el nombre de dinars i sopars bisetmanals, l'edat mitjana estigui compresa entre els 0 i 30 anys, els ingressos nets totals percebuts per les llars siguin inferiors als 1.100 € mensuals, que el sustentador principal es trobi solter, que la llar visqui en una zona rural i que el percentatge d'individus de la llar amb estudis obligatoris oscil·li entre el 0% i el 20%.

El model ***Gradient Boosting***, gràcies a la seva estabilitat entre el biaix i variància i a la seva elevada capacitat predictiva, serà el millor d'entre tots els possibles models aplicats, per predir si la llar consumirà carn o no depenent de les seves característiques. A partir de les variables explicatives més rellevants, hem pogut comparar el perfil d'aquelles llars que consumien carn i les que no. Els resultats que hem obtingut han estat els següents. El perfil de les llars que consumien carn presenten una despesa anual vuit vegades superior a les llars que no en consumien, així com el percentatge de dones era lleugerament superior en aquelles llars que consumien carn i, aquestes realitzaven el doble de dinars i sopars bisetmanals respecte a les llars que no en consumien. Per altra banda, ens ha permès comparar els perfils de llars que el model no era capaç de predir correctament, on hem pogut concloure que la despesa total de les llars que no consumien carn era el doble respecte les que sí que en consumien, que el percentatge de dones era lleugerament superior en les que no menjaven carn i que les llars que no en consumien efectuaven el doble de dinars i sopars respecte a aquelles que sí que en consumien.

Del **model Weibull,** s'ha deduït que les variables explicatives que convergien a un major cost ocult de la llar veien determinades per, un major percentatge de dones, una major edat mitjana i un major nombre de dinars i sopars bisetmanals. A més, que visquessin en zones densament poblades, en comunitats autònomes on el PIB per càpita era superior als 25.000 €, hi hagués dos o tres ocupats per llar i el sustentador principal es trobés casat.

Finalment, amb la informació obtinguda del problema de classificació i regressió, s'han posat en comú algunes iniciatives per pal·liar els costos ocults, ergo el canvi climàtic. Algunes de les alternatives proposades han estat la formació del personal sanitari per tal de canviar l'enfoc de les guies nutricionals, atorgar ajudes econòmiques que promoguin la innovació de noves tècniques sostenibles enfocades a la ramaderia, la introducció d'impostos o bé la reducció dels preus dels productes vegetals i la realització de campanyes publicitàries dirigides a aquelles famílies amb uns majors costos ocults. Totes aquestes propostes indueixen a un menor consum de carn i, per tant, a un menor cost ocult.

Referent a l'enfoc de les companyes publicitàries, s'ha conclòs que aquestes han d'anar dirigides a aquelles famílies que consumeixin carn i que en mitjana, la despesa monetària i no monetària elevada temporalment i poblacionalment sigui de 4.662.767 €, que les llars

estiguin compostes pel 48,17% de dones, que realitzin 75,44 dinars i sopars bisetmanals, que la mitjana d'edat de la llar estigui compresa entre els 31 i 64 anys, que l'estat civil del sustentador principal sigui solter, que els ingressos mensuals nets totals de cadascuna de les llars oscil·li entre els 1.100 € i 1.800 € i que visquin en comunitats autònomes amb un PIB per càpita superior als 25.000 €.

Finalment, l'anàlisi estadística del consum de carn de les llars espanyoles ens ha permès quantificar les seves conseqüències mediambientals, aportant així, propostes i noves línies d'investigació que beneficiarien tant a la població com al medi ambient.

5. BIBLIOGRAFIA

1. Memoria 2021. Todo un futuro juntos. ANICE. [Online].; 2021. Available from: https://www.anice.es/industrias/el-sector/el-sector-carnico-espanol_171_1_ap.html.
2. Ministerio de Agricultura, Pesca i Alimentación. MAPA. [Online].; 2022. Available from: https://www.mapa.gob.es/es/ganaderia/temas/produccion-y-mercados-ganaderos/sectores-ganaderos/Default.aspx.
3. SEPEAP Sociedad Española de Pediatría Extrahospitalaria y Atención Primaria. Plan estratégico y de acción para reducir el riesgo de selección y diseminación de resistencias a los antimicrobianos (PRAM). PEDIATRÍA INTEGRAL (Programa de Formación Continuada en Pediatría Extrahospitalaria). 2015 Junio;: p. 313-323.
4. Organización de las Naciones Unidas para la Agricultura y la Alimentación. Ingeniería Económica Aplicada a la Industria Pesquera. 1998..
5. Lorca MP. Población, cambio climático y huella ambiental//Population, Climate Change and Environmental Footprint. In Ecozon@: European Journal of Literature, Culture and Environment.; 2018. p. 11-36.
6. Costantini AO,PMG,BM,GFA,CVRN,RRI,&TMA. Emisiones de gases de efecto invernadero en la producción ganadera. Asociación Argentina para el Progreso de las Ciencias. 2018.
7. CLIMATE CHANGE CONNECTION. CO2 equivalents. [Online].; 2020. Available from: https://climatechangeconnection.org/emissions/co2-equivalents/.
8. United Nations Climate Change. What is the Kyoto Protocol? [Online]. Available from: https://unfccc.int/es/kyoto_protocol.
9. Acciona. ¿EN QUÉ CONSISTE EL MERCADO DE CARBONO? [Online].; 2019. Available from: https://www.sostenibilidad.com/energias-renovables/en-que-consiste-el-mercado-de-carbono/?_adin=01874690616.
10. Organización de las Naciones Unidas para la Alimentación y la Agricultura. Modelo de Evaluación Ambiental de la Ganadería Mundial (GLEAM). [Online].; 2010. Available from: https://www.fao.org/gleam/results/es/.
11. Hoekstra AY. The hidden water resource use behind meat and dairy. Animal Frontiers. 2012.
12. Iberdrola. Descubre los principales beneficios del 'Machine Learning'. [Online]. Available from: https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico.
13. Liakos KG,BP,MD,PS,&BD. Machine learning in agriculture: A review. Sensors, 18(8), 2674.:; 2018.
14. Santos PRdl. Telefónica. [Online].; 2021. Available from: https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/.
15. Romero Rojas B. Una introducción a los modelos de Machine Learning (Bachelor's thesis). 2020 Oct 26..
16. GRUPO ATICO34. ¿Qué es el underfitting y cómo prevenirlo? [Online]. Available from: https://protecciondatos-lopd.com/empresas/underfitting/#Cuales_son_sus_causas.
17. GRUPO ATICO34. Overfitting. Qué es, causas, consecuencias y cómo solucionarlo. [Online]. Available from: https://protecciondatos-lopd.com/empresas/overfitting/.
18. V7. The Data Engine for AI. [Online].; 2022. Available from: https://www.v7labs.com/blog/overfitting-vs-underfitting.
19. Rodrigo JA. Ciencia de Datos. [Online].; 2020. Available from: https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_C50.html.

20. Martin DP. Wicked Good Data. [Online].; 2016 [cited 2022 Maig. Available from: https://dpmartin42.github.io/posts/r/imbalanced-classes-part-1.
21. Martin DP. Wicked Good Data. [Online].; 2017 [cited 2022 Maig. Available from: http://dpmartin42.github.io/posts/r/imbalanced-classes-part-2.
22. Santos PRdl. Machine Learning a tu alcance: La matriz de confusión. [Online].; 2018. Available from: https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/.
23. Sitio Big Data. Aprendizaje automatico y las Metricas de regresión. [Online].; 2018. Available from: https://sitiobigdata.com/2018/08/27/machine-learning-metricas-regresion-mse/.
24. Mathworks. [Online]. Available from: https://la.mathworks.com/discovery/linear-regression.html.
25. IBM. Regresión logística. [Online]. Available from: https://www.ibm.com/es-es/topics/logistic-regression.
26. Rodrigo JA. Regresión logística simple y múltiple. [Online].; 2016. Available from: https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple.html.
27. Berlanga V,RHMJ,&VBR. Cómo aplicar árboles de decisión en SPSS. REIRE. Revista d'Innovació i Recerca en Educació, 2013, vol. 6, num. 1. 2013;: p. 65-79.
28. Greenwell BB&B. Hands-On Machine Learning with R. [Online].; 2020. Available from: https://bradleyboehmke.github.io/HOML/.
29. Fernández Casal , Costa Bouzas , Oviedo de la Fuente. Aprendizaje Estadístico. [Online].; 2021. Available from: https://rubenfcasal.github.io/aprendizaje_estadistico/.
30. Instituto Nacional de Estadística (INE). INE. [Online].; 2016. Available from: https://www.ine.es/metodologia/t25/t2530p45816.pdf.
31. Organización de las Naciones Unidas para la Alimentación y Agricultura. Modelo de Evaluación Ambiental de la Ganadería Mundial (GLEAM). [Online].; 2010. Available from: https://www.fao.org/gleam/es/.
32. Ministerio de Agricultura, Alimentación y Medio Ambiente. JORNADA GANADERÍA Y MEDIO AMBIENTE. [Online].; 2016. Available from: https://www.mapa.gob.es/es/ganaderia/temas/ganaderia-y-medio-ambiente/2-presentacionodonsobrino_tcm30-108183.pdf.
33. EpData. EpData.es. [Online].; 2021. Available from: https://www.epdata.es/datos/carne-consume-ano-espana-graficos-evolucion/605.
34. CARACTERIZACIÓN DEL SECTOR VACUNO DE CARNE EN ESPAÑA. Subdirección General de Producciones Ganaderas y Cinegéticas Dirección General de Producciones y Mercados Agrarios. [Online].; 2021. Available from: https://www.mapa.gob.es/es/ganaderia/temas/produccion-y-mercados-ganaderos/caracterizacionsectorvacunodecarne_julio2021_pub_tcm30-553721.pdf.
35. Organització de les Nacions Unides per l'Alimentació i l'Agricultura. Departament d'Agricultura i Protecció al Consumidor. Producció i Sanitat Animal. [Online].; 2014. Available from: https://www.fao.org/ag/againfo/themes/es/meat/background.html.
36. El Economista. [Online].; 2021. Available from: https://www.eleconomista.es/actualidad/noticias/11338538/07/21/Cuanta-carne-y-de-que-tipo-puede-comer-una-persona-a-la-semana-.html.
37. Roberto Rodríguez Casado PNyAG. LA HUELLA HÍDRICA DE LA GANADERÍA ESPAÑOLA. 2009 junio..
38. Azure. ¿Qué es el aprendizaje automático? [Online]. Available from: https://azure.microsoft.com/es-es/overview/what-is-machine-learning-platform/.

39. IBM. Opciones de modelo para el nodo C5.0. [Online]. Available from: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=node-c50-model-options.
40. IBM. Machine Learning. [Online].; 2020. Available from: https://www.ibm.com/es-es/cloud/learn/machine-learning.
41. Reddy S. Understanding the log loss function. [Online].; 2020. Available from: https://medium.com/analytics-vidhya/understanding-the-loss-function-of-logistic-regression-ac1eec2838ce.
42. EpData. ¿Cuánta carne se consume cada año en España? Gráficos y evolución. [Online].; 2021. Available from: https://www.epdata.es/datos/carne-consume-ano-espana-graficos-evolucion/605.

6. ANNEXOS

Annex 1. Classificació dels dotze grans grups i subgrups del grup d'Aliments .

1. Aliments i begudes no alcohòliques
   1.1. Aliments
      1.1.1. Pa i cereals
         1.1.1.1. Arròs
         1.1.1.2. Farines i altres cereals
         1.1.1.3. Pa
         1.1.1.4. Altres productes de fleca
         1.1.1.5. Pizza i quiche
         1.1.1.6. Pastes alimentàries i cuscús
         1.1.1.7. Cereals d'esmorzar
         1.1.1.8. Altres productes a base de cereals
      1.1.2. Carn
         1.1.2.1. Carn de porcí
         1.1.2.2. Carn d'oví i caprí
         1.1.2.3. Carn d'au
         1.1.2.4. Altres carns
         1.1.2.5. Despulles comestibles
         1.1.2.6. Carn seca, salada o fumada
         1.1.2.7. Altres preparats de carn
      1.1.3. Peix i marisc
         1.1.3.1. Peix fresc o refrigerat
         1.1.3.2. Peix congelat
         1.1.3.3. Marisc fresc o refrigerat
         1.1.3.4. Marisc congelat
         1.1.3.5. Peix. i marisc sec, fumat o salat
         1.1.3.6. Altres preparats de peix i marisc conservats o processats
      1.1.4. Llet, formatge i ous
         1.1.4.1. Llet fresca sencera
         1.1.4.2. Llet fresca desnatada
         1.1.4.3. Llet en conserva
         1.1.4.4. Iogurt
         1.1.4.5. Formatge i quallada
         1.1.4.6. Altres productes lactis
         1.1.4.7. Ous
      1.1.5. Olis i greixos

1.1.5.1. Mantega
1.1.5.2. Margarina i altres greixos vegetals
1.1.5.3. Oli d'oliva
1.1.5.4. Altres olis comestibles
1.1.5.5. Altres greixos animals comestibles

1.1.6. Fruits

1.1.6.1. Fruites fresques o refrigerades
1.1.6.2. Fruites congelades
1.1.6.3. Fruits secs i fruits de closca
1.1.6.4. Fruites en conserva i productes a base de fruites

1.1.7. Llegums i hortalisses

1.1.7.1. Llegums i hortalisses fresques o refrigerades, excepte patates i altres tubercles
1.1.7.2. Llegums i hortalisses congelades, excepte patates i altres tubercles
1.1.7.3. Llegums i hortalisses seques o conservades d'una altra forma o processades
1.1.7.4. Patates
1.1.7.5. Patates xips
1.1.7.6. Altres tubercles i els seus productes

1.1.8. Sucre, confitura, mel, xocolata i confiteria

1.1.8.1. Sucre
1.1.8.2. Confitures, melmelades i mel
1.1.8.3. Xocolata
1.1.8.4. Productes de confiteria
1.1.8.5. Gelats
1.1.8.6. Succedanis artificials del sucre

1.1.9. Productes alimentaris n.c.o.p

1.1.9.1. Salses i condiments
1.1.9.2. Sal, espècies i herbes culinàries
1.1.9.3. Aliments per a nadons
1.1.9.4. Plats preparats
1.1.9.5. Altres productes alimentaris n.c.o.p

2. Begudes alcohòliques, tabac i narcòtics
3. Habitatge, aigua, electricitat, gas i altres combustibles
4. Vestimenta i calçat
5. Mobiliari, equipament de la llar i despeses corrents de conservació de l'habitatge
6. Salut
7. Transports
8. Comunicacions
9. Oci, espectacles i cultura

10. Ensenyança
11. Hotels
12. Cafès i restaurants
13. Altres bens i serveis

Annex 2. Estructura de les variables de l'EPF.

- **Fitxers de la llar**

	Nom de la variable	Descripció
Informació general	**ANOENC**	Any de l'enquesta
	NUMERO	Número seqüencial que indica l'ordre de la llar en el fitxer
	CCAA	Comunitat autònoma de residència
	NUTS1	Regió
	CAPROV	Capital de província
	TAMAMU	Grandària del municipi
	DENSIDAD	Densitat de població
	CLAVE	Clau de col·laboració efectiva de la llar
	CLATEO	Clau de col·laboració teòrica de la llar
	FACTOR	Factor poblacional

	Nom de la variable	Descripció
Característiques respectives a la llar	**NMIEMB**	Nombre de membres de la llar)
	TAMANO	Grandària de la llar
	NMIEMSD	Nombre de membres de la llar que pertanyen al servei domèstic
	NMIEMHU	Nombre de membres de la llar que són hostes
	NMIEMIN	Nombre de membres de la llar que són convidats
	NMIEM1	Nombre de membres de la llar de 14 o més anys
	NMIEM2	Nombre de membres de la llar menors de 14 anys
	NMIEM3	Nombre de membres de la llar menors de 16 anys
	NMIEM4	Nombre de membres de la llar de 16 o més anys
	NMIEM5	Nombre de membres de la llar menors de 18 anys

NMIEM6	Nombre de membres de la llar de 18 o més anys
NMIEM7	Nombre de membres de la llar de 0 a 4 anys
NMIEM8	Nombre de membres de la llar de 5 a 15 anys
NMIEM9	Nombre de membres de la llar de 16 a 24 anys
NMIEM10	Nombre de membres de la llar de 25 a 34 anys
NMIEM11	Nombre de membres de la llar de 35 a 64 anys
NMIEM12	Nombre de membres de la llar de 65 a 84 anys
NMIEM13	Nombre de membres de la llar de 85 o més anys
NUMACTI	Nombre de membres actius en la llar
NUMINACTI	Nombre de membres no actius en la llar
NUMOCU	Nombre de membres ocupats en la llar
NUMNOCU	Nombre de membres no ocupats en la llar
NUMESTU	Nombre d'estudiants en la llar
NUMNOESTU	Número de no estudiants en la llar
NNINOSD	Nombre de nens dependents
NHIJOSD	Nombre de fills dependents
UC1	Grandària equivalent de la llar. Escala OCDE 1 + 0,7 * (NMIEM1 - 1) + 0,5 * NMIEM2
UC2	Grandària equivalent de la llar. Escala OCDE modificada 1 + 0,5 * (NMIEM1- 1) + 0,3 * NMIEM2
PF2TEO	Nombre de llibretes de comptes individuals teòrics
PF2RECO	Nombre de llibretes de comptes individuals recollides
TIPHOGAR1	Tipus de llar (primera classificació)
TIPHOGAR2	Tipus de llar (segona classificació)
TIPHOGAR3	Tipus de llar (tercera classificació)
TIPHOGAR4	Tipus de llar (quarta classificació)
TIPHOGAR5	Tipus de llar (cinquena classificació)
TIPHOGAR6	Tipus de llar (sisena classificació)
TIPHOGAR7	Tipus de llar (setena classificació)
TIPHOGAR8	Tipus de llar (vuitena classificació) derivada de TIPHOGAR1
TIPHOGAR9	Tipus de llar (novena classificació) derivada de TIPHOGAR4
TIPHOGAR10	Tipus de llar (desena classificació)
TIPHOGAR11	Tipus de llar (onzena classificació)
SITUOCUHOG	Situació de la llar respecte a l'ocupació
SITUACTHOG	Situació de la llar respecte a l'activitat

	Nom de la variable	Descripció
Dades del sustentador principal	**NORDENSP**	Número d'ordre
	EDADSP	Edat (calculada a data de compliment d ela fitxa de la llar)
	SEXOSP	Sexe
	PAISNACSP	País naixement
	NACIONASP	Nacionalitat
	PAISSP	País del que té nacionalitat
	SITURESSP	Situació de residencia
	ECIVILLEGASP	Estat civil legal
	NORDENCOSP	Número d'ordre del cònjuge o parella
	UNIONSP	Tipo d'unió mb el cònjuge o parella
	CONVIVENCIASP	Convivència en parella del sustentador
	NORDENPASP	Número d'ordre del pare
	PAISPADRESP	País de naixement del pare
	NORDENMASP	Número d'ordre de la mare
	PAISMADRESP	País de naixement de la mare
	ESTUDIOSSP	Estudis completats
	ESTUDREDSP	Estudis completats reduïts
	SITUACTSP	Situació en l'activitat la setmana
	SITUREDSP	Situació en l'activitat reduïda
	OCUSP	Estava el sustentador principal ocupat en la setmana anterior de la entrevista?
	JORNADASP	Tipus de jornada de treball
	PERCEPSP	És el sustentador principal perceptor d'ingressos monetaris regulars durant el mes anterior al de l'entrevista?
	IMPEXACPSP	Import exacte dels ingressos mensuals nets del. Sustentador Principal
	INTERINPSP	Interval d'ingressos mensuals nets
	TRABAJO	Ha treballat algun cop a la seva vida?
	OCUPA	Ocupació que exerceix o va exercir
	OCUPARED	Ocupació que exerceix o va exercir reduïda
	ACTESTB	Activitat de l'establiment en el que treballa o va treballar
	ACTESTBRED	Activitat de l'establiment en el que treballa o va treballar reduïda

	SITPROF	Situació professional
	SECTOR	Sector de l'activitat
	CONTRATO	Contracte laboral
	TIPOCONT	Tipus de contracte
	SITSOCI	Situació socioeconòmica del sustentador principal
	SITSOCIRE	Situació socioeconòmica del sustentador principal (Classificació Reduïda)

	Nom de la variable	**Descripció**
Característiques de l'habitatge principal	**REGTEN**	Règim de tinença
	TIPOEDIF	Mena d'edifici en el qual està situada l'habitatge
	ZONARES	Tipus de zona de residència
	TIPOCASA	Mena de casa
	NHABIT	Nombre d'habitacions
	ANNOCON	Data construcció edifici
	SUPERF	Superfície útil de l'habitatge
	AGUACALI	Disposició d'aigua calenta
	FUENAGUA	Font d'energia per a aigua calenta
	CALEF	Disposició de calefacció
	FUENCALE	Font d'energia per a calefacció

	Nom de la variable	**Descripció**
Altres habitatges a disposició de la llar	**DISPOSIOV**	Disposició dels habitatges en els últims 12 mesos
	NUMOVD	Número d'altres habitatges a disposició de la llar
	RENTENV1	Règim de tinença de l'habitatge 1
	MESESV1	Número de mesos a disposició de la llar de l'habitatge 1
	DIASV1	Número de dies a disposició de la llar de l'habitatge 1
	AGUAV1	Disposició d'aigua calenta en l'habitatge 1
	FUENACV1	Font d'energia per l'aigua calenta en l'habitatge 1
	CALEFV1	Disposició de calefacció en l'habitatge 1
	FUENCAV1	Font d'energia per calefacció en l'habitatge 1
	REGTENV2	Règim de tinença de l'habitatge 2
	MESESV2	Número de mesos a disposició de la llar de l'habitatge 2

DIASV2	Número de dies a disposició de la llar de l'habitatge 2
AGUAV2	Disposició d'aigua calenta en l'habitatge 2
FUENACV2	Font d'energia per l'aigua calenta en l'habitatge 2
CALEFV2	Disposició de calefacció en l'habitatge 2
FUENCAV2	Font d'energia per calefacció en l'habitatge 2
REGTENV3	Règim de tinença de l'habitatge 3
MESESV3	Número de mesos a disposició de la llar de l'habitatge 3
DIASV3	Número de dies a disposició de la llar de l'habitatge 3 (0-30)
AGUAV3	Disposició d'aigua calenta en l'habitatge 3
FUENACV3	Font d'energia per l'aigua calenta en l'habitatge 3
CALEFV3	Disposició de calefacció en l'habitatge 3
FUENCAV3	Font d'energia per calefacció en l'habitatge 3
REGTENV4	Règim de tinença de l'habitatge 4
MESESV4	Número de mesos a disposició de la llar de l'habitatge 4
DIASV4	Número de dies a disposició de la llar de l'habitatge 4
AGUAV4	Disposició d'aigua calenta en l'habitatge 4
FUENACV4	Font d'energia per l'aigua calenta en l'habitatge 4
CALEFV4	Disposició de calefacció en l'habitatge 4
FUENCAV4	Font d'energia per calefacció en l'habitatge 4
REGTENV5	Règim de tinença de l'habitatge 5
MESESV5	Número de mesos a disposició de la llar de l'habitatge 5
DIASV5	Número de dies a disposició de la llar de l'habitatge 5
AGUAV5	Disposició d'aigua calenta en l'habitatge 5
FUENACV5	Font d'energia per l'aigua calenta en l'habitatge 5
CALEFV5	Disposició de calefacció en l'habitatge 5
FUENCAV5	Font d'energia per calefacció en l'habitatge 5
REGTENV6	Règim de tinença de l'habitatge 6
MESESV6	Número de mesos a disposició de la llar de l'habitatge 6

DIASV6	Número de dies a disposició de la llar de l'habitatge 6
AGUAV6	Disposició d'aigua calenta en l'habitatge 6
FUENACV6	Font d'energia per l'aigua calenta en l'habitatge 6
CALEFV6	Disposició de calefacció en l'habitatge 6
FUENCAV6	Font d'energia per calefacció en l'habitatge 6
REGTENV7	Règim de tinença de l'habitatge 7
MESESV7	Número de mesos a disposició de la llar de l'habitatge 7
DIASV7	Número de dies a disposició de la llar de l'habitatge 7
AGUAV7	Disposició d'aigua calenta en l'habitatge 7
FUENACV7	Font d'energia per l'aigua calenta en l'habitatge 7
CALEFV7	Disposició de calefacció en l'habitatge 7
FUENCAV7	Font d'energia per calefacció en l'habitatge 7
REGTENV8	Règim de tinença de l'habitatge 8
MESESV8	Número de mesos a disposició de la llar de l'habitatge 8
DIASV8	Número de dies a disposició de la llar de l'habitatge 8
AGUAV8	Disposició d'aigua calenta en l'habitatge 8
FUENACV8	Font d'energia per l'aigua calenta en l'habitatge 8
CALEFV8	Disposició de calefacció en l'habitatge 8
FUENCAV8	Font d'energia per calefacció en l'habitatge 8
REGTENV9	Règim de tinença de l'habitatge 9
MESESV9	Número de mesos a disposició de la llar de l'habitatge 9
DIASV9	Número de dies a disposició de la llar de l'habitatge 9
AGUAV9	Disposició d'aigua calenta en l'habitatge 9
FUENACV9	Font d'energia per l'aigua calenta en l'habitatge 9
CALEFV9	Disposició de calefacció en l'habitatge 9
FUENCAV9	Font d'energia per calefacció en l'habitatge 9

Nom de la variable	**Descripció**
GASTOT	Import total de la despesa anual de la llar monetari i no monetari, elevat temporal i poblacionalment

Despeses de consum de la llar	**IMPUTGAS**	Percentatge d'imputació de la despesa total
	GASTMON	Import total de la despesa anual de la llar monetari, elevat temporal i poblacionalment
	GASTNOM1	Import total de la despesa anual de la llar no monetari procedent del autoconsum elevat temporal o poblacionalment
	GASTNOM2	Import total de la despesa anual de la llar no monetari procedent del autosuministrament elevat temporal o poblacionalment
	GASTNOM3	Import total de la despesa anual de la llar no monetari del salari en espècie elevat temporal o poblacionalment
	GASTNOM4	Import total de la despesa anual de la llar no monetari procedent del lloguer imputat als habitatges, principal i altres habitats a disposició de la llar, en propietat o cedides gratuïta o semigratuïtament per raó diferent a la feina elevat temporal o poblacionalment

	Nom de la variable	**Descripció**
Ingressos regulars mensuals de la llar	**CAPROP**	Ingressos de feina per compte pròpia
	CAJENA	Ingressos de feina per compte aliena
	PENSIO	Ingressos per pensions contributives i no contributives
	DESEM	Subsidis i presentacions d'atur
	OTRSUB	Altres subsidis i prestacions socials
	RENTAS	Rendes de la propietat i del capital
	OTROIN	Altres ingressos regulars
	FUENPRIN	Principal font d'ingressos
	FUENPRINRED	Principal font d'ingressos reduïda
	IMPEXAC	Import exacte dels ingressos mensuals nets total de la llar
	INTERIN	Interval d'ingressos mensuals nets totals de la llar
	NUMPERI	Número de membres de la llar perceptors d'ingressos

	Nom de la variable	**Descripció**

Número de menjars i sopars durant la bisetmanal mostral	**COMIMH**	Número de menjars i sopars efectuats per els membres de la llar
	COMISD	Número de menjars i sopars efectuats per el servei domèstic
	COMIHU	Número de menjars i sopars efectuats per els hostes
	COMIINV	Número de menjars i sopars efectuats per els convidats
	COMTOT	Número total de menjars i sopars

- **Fitxer dels membres de la llar**

Nom de la variable	Descripció
ANOENC	Any de l'enquesta
NUMERO	Número seqüencial que indica l'ordre de la llar en el fitxer
NORDEN	Número de l'ordre del membre de la llar
CATEGMH	Categoria del membre de la llar
SUSPRIN	Es el sustentador principal?
RELASP	Relació de parentesc amb el sustentador principal
EDAD	Edat calculada amb referencia a la data de l'entrevista de la fitxa de la llar
SEXO	Sexe
PAISNACIM	País de naixement
NACIONA	Nacionalitat
PAISNACION	País del que te la nacionalitat estrangera
SITURES	Situació de residencia
ECIVILEGAL	Estat civil legal
NORDENCO	Número d'ordre del cònjuge o parella
UNION	Tipus d'unió amb el cònjuge o parella
CONVIVENC	Convivència en parella
NORDENPA	Número d'ordre del pare
PAISPADRE	País de naixement del pare
NORDENMA	Número d'ordre de la mare
PAISMADRE	País de naixement de la mare
ESTUDIOS	Estudis completats
ESTUDRED	Estudis completats reduïda
SITUACT	Situació en l'activitat de la setmana anterior a la entrevista
SITURES	Situació de l'activitat reduïda

OCU	Estava ocupat el membre de la llar la setmana anterior a l'entrevista?
JORNADA	Tipus de jornada a la feina
PERCEP	Es perceptor d'ingressos monetaris regulars durant el mes anterior a l'entrevista?
IMPEXACP	Import exacte dels ingressos mensuals nets totals del membre de la llar
INTERINP	Interval d'ingressos mensuals nets totals dels membres de la llar
NINODEP	És nen dependent?
HIJODEP	És fill dependent?
ADULTO	És adult?
FACTOR	Factor poblacional

- **Fitxer de despeses**

Nom Variable	Descripció
ANOENC	Any de l'enquesta
NUMERO	Número seqüencial que indica l'ordre de la llar en el fitxer
CODIGO	Codi de despesa segons ECOICOP.
GASTO	Import total de la despesa monetària i no monetari elevat temporal i poblacionalment
PORCENDES	Percentatge de desglossament de la despesa total (2 decimals)
PORCENIMP	Percentatge d'imputació de la despesa total (2 decimals)
CANTIDAD	Quantitat (només per als codis que requereixen quantitat física) elevat temporal i poblacionalment.
GASTOMON	Import total de la despesa monetària (per al salari en espècie es comptabilitza només l'import del pagament realitzat per la llar) elevat temporal i poblacionalment
GASTNOM1	Import total de la despesa no monetària procedent de l'autoconsum elevat temporal i poblacionalment
GASTNOM2	Import total de la despesa no monetària procedent de l'autosubministrament elevat temporal i poblacionalment

GASTNOM3	Import total de la despesa no monetària procedent del salari en espècie, (no s'inclou lloguer imputat a l'habitatge cedit per raó de treball) elevat temporal i poblacionalment
GASTNOM4	Import total de la despesa no monetària procedent del lloguer imputat a l'habitatge principal i altres habitatges a la disposició de la llar, en propietat o cedida gratuïta o semigratuitament per raó diferent a treball elevat temporal i poblacionalment
GASTNOM5	Import total de la despesa no monetària procedent del lloguer imputat a l'habitatge, principal i altres habitatges a la disposició de la llar, cedida per raó de treball elevat temporal i poblacionalment
FACTOR	Factor poblacional

Annex 3. Càlcul de les variables auxiliars la Base de Dades I. Consum de la llar.

- **GASTOT:** Despesa total de la llar en euros.

$$\boldsymbol{GASTOT_{i,j}} = \sum_{\boldsymbol{i \in ANOENC}} \sum_{\boldsymbol{j \in CODIGO}} \boldsymbol{GASTO_{i,j}}.$$

- **SEXO:** Percentatge de dones a la llar. (0-100)

$$SEXO = \frac{n_{Dones}}{n_{Total}} \cdot 100.$$

- **EDAD:** Edat mitjana dels membres de la llar en anys. (0-85)

$$EDAD = \frac{\sum_{i \in Membres\ \ llar} EDAD_i}{n_{Total}}.$$

- **NACION:** Percentatge de membres amb nacionalitat espanyola. (0-100)

$$NACION = \frac{n_{Nacionalitat\ espanyola}}{n_{Total}} \cdot 100.$$

- **CCAA[4]:** Comunitat Autònoma segons PIB per càpita. (1: CCAA amb PIB per càpita superior a 25.000 euros; 2: CAA amb PIB per càpita entre 25.000 i 20.000 euros; 3: CCAA amb PIB per càpita inferior a 20.000 euros).

[4] Font: https://www.bankinter.com/blog/finanzas-personales/pib-per-capita-espana-comunidades-autonomas-grafico

- CCAA amb PIB per càpita superior a 25.000 euros: Aragó, Catalunya, Comunitat Foral de Navarra, País Basc, Comunitat de Madrid, La Rioja.
- CAA amb PIB per càpita entre 25.000 i 20.000 euros: Comunitat Valenciana, Principat d'Astúries, Galícia, Illes Balears, Cantabria, Castella i Lleó.
- CCAA amb PIB per càpita inferior a 20.000 euros: Regió de Múrcia, Ceuta i Melilla, Castella la Manxa, Andalusia, Canàries.

- **ESTUDIOS1:** Percentatge de membres de la llar sense Estudis Primaris. (0-100)

$$ESTUDIOS1 = \frac{n_{\ ESTUDIOS=1}}{n_{Total}} \cdot 100.$$

- **ESTUDIOS2:** Percentatge de membres de la llar amb Estudis Obligatoris. (0-100)

$$ESTUDIOS2 = \frac{n_{\ ESTUDIOS=2;\ ESTUDIOS=3}}{n_{Total}} \cdot 100.$$

- **ESTUDIOS3:** Percentatge de membres de la llar amb Estudis Superiors. (0-100)

$$ESTUDIOS3 = \frac{n_{\ ESTUDIOS=3;\ ESTUDIOS=4;\ ESTUDIOS=5;\ ESTUDIOS=6;\ ESTUDIOS=7;\ ESTUDIOS=8}}{n_{Total}} \cdot 100.$$

Annex 4. Distribució de la variable resposta

Distribució	AIC	BIC
Weibull	**1054559**	1054576
Gamma	**1054586**	1054603
Exponencial	1054591	1054600
Log Normal	1066968	1066985
Log Gamma	1072500	1072517
Rayleigh	1087647	1087655
Weibull Inversa	1090820	1090837
Primes de Beta	1120908	1120925
Gamma Inversa	1120974	1120991
Inversa Gaussiana	1132317	1132334

Annex 5. Codi R emprat

A continuació es mostra el codi emprat pel processament, la construcció de la base de dades, la normalització, la descriptiva, els gràfics i la construcció de models.

Bloc A. Problema de classificació

A.1) Base de dades

Preprocessament de l'Enquesta de Pressupostos Familiars corresponent als anys 2006,2016 i 2020.

Importació base de dades

Primerament, procedim a importar la base de dades relatives al 2006, 2016 i 2020 de l'arxiu de despeses, de la llar i dels membres de la llar.

Anotació: Del 2006 al 2015 els resultats es mostren segons la classificació COICOP I a partir de 2016 es mostren segons la classificació ECOICOP. (En el nostre anàlisi no té cap efecte)

```
setwd("~/Desktop")
# Nota: les variables referents al fitxer de 2006 es trobaven en .txt i no es podien llegir directament al R, així que les he tingut que separar a Excel i posteriorment ja les he importat al R.

library(readxl)
EPFgastos_2006 <- read_excel("EPFgastos_2006.xlsx")
EPFhogar_2006 <- read_excel("EPFhogar_2006.xlsx")
EPFmhogar_2006 <- read_excel("EPFmhogar_2006.xlsx")

setwd("~/Desktop/TFG/Dades/Fitxers")
library(readr)
EPFgastos_2016 <- read_delim( "EPFgastos_2016.csv", delim = "\t", escape_double = FALSE, trim_ws = TRUE);
EPFhogar_2016 <- read_delim("EPFhogar_2016.csv", delim = "\t", escape_double = FALSE, trim_ws = TRUE)
EPFmhogar_2016 <- read_delim("EPFmhogar_2016.csv", delim = "\t", escape_double = FALSE, trim_ws = TRUE)

setwd("~/Desktop/TFG/Dades/Fitxers")
EPFgastos_2020 <- read_delim("EPFgastos_2020.csv", delim = "\t", escape_double = FALSE, trim_ws = TRUE);
EPFhogar_2020 <- read_delim("EPFhogar_2020.csv", delim = "\t", escape_double = FALSE, trim_ws = TRUE)
EPFmhogar_2020 <- read_delim("EPFmhogar_2020.csv", delim = "\t", escape_double = FALSE, trim_ws = TRUE)
```

Selecció de variables i exploració de la base de dades.

En aquest punt el que es tractarà es de seleccionar les variables que són del nostre interès relatives als 3 arxius esmentats anteriorment.

Anotació: Les variables del 2006 quasi totes són iguals, però hi ha algunes que canvien. Es per això que les columnes son diferents respecte les del 2016 i 2020. Cal esmentar que la informació final serà la mateixa pels 3 anys.

```
### Seleccio de variables
EPFgastos_2006 <- EPFgastos_2006[,c("ANOENC", "NUMERO", "CODIGO","GASTO")]
EPFhogar_2006  <-  EPFhogar_2006[,c("ANOENC",  "NUMERO",  "CCAA",  "DENSI",
"NUMOCUP", "ZONARES", "COMITOT", "IMPEXAC", "ECIVILSP")]
EPFmhogar_2006  <-  EPFmhogar_2006[,c("ANOENC","NUMERO",  "EDAD",  "SEXO",
"NACIONA", "ESTUDIOS")]

EPFgastos_2016 <- EPFgastos_2016[,c("ANOENC", "NUMERO", "CODIGO","GASTO")]
EPFhogar_2016  <-  EPFhogar_2016[,c("ANOENC",  "NUMERO",  "CCAA",  "DENSIDAD",
"NUMOCU", "ZONARES", "COMITOT", "IMPEXAC", "ECIVILLEGALSP")]
EPFmhogar_2016  <-  EPFmhogar_2016[,c("ANOENC","NUMERO",  "EDAD",  "SEXO",
"NACIONA", "ESTUDIOS")]

EPFgastos_2020 <- EPFgastos_2020[,c("ANOENC", "NUMERO", "CODIGO","GASTO")]
EPFhogar_2020  <-  EPFhogar_2020[,c("ANOENC",  "NUMERO",  "CCAA",  "DENSIDAD",
"NUMOCU", "ZONARES", "COMITOT", "IMPEXAC", "ECIVILLEGALSP")]
EPFmhogar_2020  <-  EPFmhogar_2020[,c("ANOENC","NUMERO",  "EDAD",  "SEXO",
"NACIONA", "ESTUDIOS")]

# Definim les classes de les variables
# 2006
EPFgastos_2006$ANOENC <- as.factor(EPFgastos_2006$ANOENC)
EPFgastos_2006$NUMERO <- as.factor(EPFgastos_2006$NUMERO)
EPFgastos_2006$CODIGO <- as.factor(EPFgastos_2006$CODIGO )
EPFgastos_2006$GASTO <- as.integer(EPFgastos_2006$GASTO)

EPFhogar_2006$ANOENC <- as.factor(EPFhogar_2006$ANOENC )
EPFhogar_2006$NUMERO <- as.factor(EPFhogar_2006$NUMERO)
EPFhogar_2006$CCAA <- as.factor(EPFhogar_2006$CCAA)
EPFhogar_2006$DENSI <- as.factor(EPFhogar_2006$DENSI)
EPFhogar_2006$NUMOCUP <- as.factor(EPFhogar_2006$NUMOCUP)
EPFhogar_2006$ZONARES <- as.factor(EPFhogar_2006$ZONARES)
EPFhogar_2006$COMITOT <- as.integer(EPFhogar_2006$COMITOT)
EPFhogar_2006$IMPEXAC <- as.integer(EPFhogar_2006$IMPEXAC)
EPFhogar_2006$ECIVILSP <- as.factor(EPFhogar_2006$ECIVILSP)

EPFmhogar_2006$ANOENC <- as.factor(EPFmhogar_2006$ANOENC)
EPFmhogar_2006$NUMERO <- as.factor(EPFmhogar_2006$NUMERO)
EPFmhogar_2006$EDAD <- as.integer(EPFmhogar_2006$EDAD)
EPFmhogar_2006$SEXO <- as.factor(EPFmhogar_2006$SEXO)
EPFmhogar_2006$NACIONA <- as.factor(EPFmhogar_2006$NACIONA)
EPFmhogar_2006$ESTUDIOS <- as.factor(EPFmhogar_2006$ESTUDIOS)

# 2016
EPFgastos_2016$ANOENC <- as.factor(EPFgastos_2016$ANOENC)
EPFgastos_2016$NUMERO <- as.factor(EPFgastos_2016$NUMERO)
EPFgastos_2016$CODIGO <- as.factor(EPFgastos_2016$CODIGO )
EPFgastos_2016$GASTO <- as.integer(EPFgastos_2016$GASTO)

EPFhogar_2016$ANOENC <- as.factor(EPFhogar_2016$ANOENC )
EPFhogar_2016$NUMERO <- as.factor(EPFhogar_2016$NUMERO)
EPFhogar_2016$CCAA <- as.factor(EPFhogar_2016$CCAA)
EPFhogar_2016$DENSIDAD <- as.factor(EPFhogar_2016$DENSIDAD)
EPFhogar_2016$NUMOCU <- as.factor(EPFhogar_2016$NUMOCU)
EPFhogar_2016$ZONARES <- as.factor(EPFhogar_2016$ZONARES)
EPFhogar_2016$COMITOT <- as.integer(EPFhogar_2016$COMITOT)
EPFhogar_2016$IMPEXAC <- as.integer(EPFhogar_2016$IMPEXAC)
EPFhogar_2016$ECIVILLEGALSP <- as.factor(EPFhogar_2016$ECIVILLEGALSP)

EPFmhogar_2016$ANOENC <- as.factor(EPFmhogar_2016$ANOENC)
EPFmhogar_2016$NUMERO <- as.factor(EPFmhogar_2016$NUMERO)
EPFmhogar_2016$EDAD <- as.integer(EPFmhogar_2016$EDAD)
EPFmhogar_2016$SEXO <- as.factor(EPFmhogar_2016$SEXO)
EPFmhogar_2016$NACIONA <- as.factor(EPFmhogar_2016$NACIONA)
EPFmhogar_2016$ESTUDIOS <- as.factor(EPFmhogar_2016$ESTUDIOS)
```

```
# 2020
EPFgastos_2020$ANOENC <- as.factor(EPFgastos_2020$ANOENC)
EPFgastos_2020$NUMERO <- as.factor(EPFgastos_2020$NUMERO)
EPFgastos_2020$CODIGO <- as.factor(EPFgastos_2020$CODIGO )
EPFgastos_2020$GASTO <- as.integer(EPFgastos_2020$GASTO)

EPFhogar_2020$ANOENC <- as.factor(EPFhogar_2020$ANOENC )
EPFhogar_2020$NUMERO <- as.factor(EPFhogar_2020$NUMERO)
EPFhogar_2020$CCAA <- as.factor(EPFhogar_2020$CCAA)
EPFhogar_2020$DENSIDAD <- as.factor(EPFhogar_2020$DENSIDAD)
EPFhogar_2020$NUMOCU <- as.factor(EPFhogar_2020$NUMOCU)
EPFhogar_2020$ZONARES <- as.factor(EPFhogar_2020$ZONARES)
EPFhogar_2020$COMITOT <- as.integer(EPFhogar_2020$COMITOT)
EPFhogar_2020$IMPEXAC <- as.integer(EPFhogar_2020$IMPEXAC)
EPFhogar_2020$ECIVILLEGALSP <- as.factor(EPFhogar_2020$ECIVILLEGALSP)

EPFmhogar_2020$ANOENC <- as.factor(EPFmhogar_2020$ANOENC)
EPFmhogar_2020$NUMERO <- as.factor(EPFmhogar_2020$NUMERO)
EPFmhogar_2020$EDAD <- as.integer(EPFmhogar_2020$EDAD)
EPFmhogar_2020$SEXO <- as.factor(EPFmhogar_2020$SEXO)
EPFmhogar_2020$NACIONA <- as.factor(EPFmhogar_2020$NACIONA)
EPFmhogar_2020$ESTUDIOS <- as.factor(EPFmhogar_2020$ESTUDIOS)

# Exploració de les dades
dim(EPFgastos_2006);          str(EPFgastos_2006);          head(EPFgastos_2006);
summary(EPFgastos_2006)
dim(EPFgastos_2016);          str(EPFgastos_2016);          head(EPFgastos_2016);
summary(EPFgastos_2016)
dim(EPFgastos_2020);          str(EPFgastos_2020);          head(EPFgastos_2020);
summary(EPFgastos_2020)

dim(EPFhogar_2006);           str(EPFhogar_2006);           head(EPFhogar_2006);
summary(EPFhogar_2006)
dim(EPFhogar_2016);           str(EPFhogar_2016);           head(EPFhogar_2016);
summary(EPFhogar_2016)
dim(EPFhogar_2020);           str(EPFhogar_2020);           head(EPFhogar_2020);
summary(EPFhogar_2020)

dim(EPFmhogar_2006);          str(EPFmhogar_2006);          head(EPFmhogar_2006);
summary(EPFmhogar_2006)
dim(EPFmhogar_2016);          str(EPFmhogar_2016);          head(EPFmhogar_2016);
summary(EPFmhogar_2016)
dim(EPFmhogar_2020);          str(EPFmhogar_2020);          head(EPFmhogar_2020);
summary(EPFmhogar_2020)
```

Tractament dels valors *missings*

Una variable pot presentar NA's però també pot prendre valor -9 el qual indica que aquesta dada no consta. Es per això que primerament abans de crear la nova variable haurem de mirar si hi ha NA's i si tots els valors consten, es a dir, que no apareixen dades que prenguin valor -9. Si ens trobem que hi ha dades que prenen valor -9 els haurem de substituir per un NA.

```
# 2006
EPFhogar_2006$COMITOT[EPFhogar_2006$COMITOT == -9] <- NA
EPFhogar_2006$ECIVILSP[EPFhogar_2006$ECIVILSP == -9] <- NA

# 2016
EPFhogar_2016$NUMOCU[EPFhogar_2016$NUMOCU == -9] <- NA

# 2020
EPFhogar_2020$NUMOCU[EPFhogar_2020$NUMOCU == -9] <- NA
```

A continuació, es mostra un quadre resum que ens indica el nombre de missings per cada una de les variables:

```
# 2006
apply(apply(EPFgastos_2006,2,is.na),2,sum)
apply(apply(EPFhogar_2006,2,is.na),2,sum)
apply(apply(EPFmhogar_2006,2,is.na),2,sum)

# 2016
apply(apply(EPFgastos_2016,2,is.na),2,sum)
apply(apply(EPFhogar_2016,2,is.na),2,sum)
apply(apply(EPFmhogar_2016,2,is.na),2,sum)

# 2020
apply(apply(EPFgastos_2020,2,is.na),2,sum)
apply(apply(EPFhogar_2020,2,is.na),2,sum)
apply(apply(EPFmhogar_2020,2,is.na),2,sum)
```

i. Tractament *missings* BASE DE DADES 2006

- EPFgastos_2006: la base de dades NO presenta *missings*, no cal tractar-los.
- EPFhogar_2006: la base de dades presenta 2 *missings*, decidim eliminar-lo ja que 2 llars sobre el total de la mostra no és rellevant.
- EPFmhogar_2006: la base de dades presenta 9.277 *missings*, caldrà imputar-los amb la funció MICE.

```
# Eliminem els missings de la base de dades EPFhogar_2006
EPFhogar_2006 <- EPFhogar_2006[-which(is.na(EPFhogar_2006$COMITOT)),]
EPFhogar_2006 <- EPFhogar_2006[-which(is.na(EPFhogar_2006$ECIVILSP)),]

# Imputem els valors missings de la base de dades EPFmhogar_2006
library(mice)
tempData_2006 <- mice(EPFmhogar_2006,m=1,meth='pmm',seed=500)
completedData_2006 <- complete(tempData_2006,1)
```

ii. Tractament *missings* BASE DE DADES 2016

- EPFgastos_2016: la base de dades no presenta cap *missing*, no s'ha de tractar.
- EPFhogar_2016: la base de dades presenta 11 *missings*, decidim eliminar-lo ja que 11 llars sobre el total de la mostra no és rellevant.
- EPFmhogar_2016: la base de dades presenta 9.882 *missings*, caldrà imputarlos amb la funció MICE.

```
# Imputem els valors missings de la base de dades EPFmhogar_2006
library(mice)
tempData_2016 <- mice(EPFmhogar_2016,m=1,meth='pmm',seed=500)
completedData_2016 <- complete(tempData_2016,1)
```

iii. Tractament *missings* BASE DE DADES 2020

- EPFgastos_2020: la base de dades NO presenta *missings*, no cal tractar-los.
- EPFhogar_2020: la base de dades presenta 15 *missings*, decidim eliminar-lo ja que 15 llars sobre el total de la mostra no és rellevant.
- EPFmhogar_2020: la base de dades presenta 8.009 *missings*, caldrà imputar-los amb la funció MICE.

```
# Eliminem els missings de la base de dades EPFhogar_2020
EPFhogar_2020 <- EPFhogar_2020[-which(is.na(EPFhogar_2020$NUMOCU)),]
```

```
# Imputem els valors missings de la base de dades EPFmhogar_2006
library(mice)
tempData_2020 <- mice(EPFmhogar_2020,m=1,meth='pmm',seed=500)
completedData_2020 <- complete(tempData_2020,1)
```

Comprovem que cap base de dades tingui cap missing:

```
# 2006
apply(apply(EPFgastos_2006,2,is.na),2,sum)
apply(apply(EPFhogar_2006,2,is.na),2,sum)
apply(apply(completedData_2006,2,is.na),2,sum)

# 2016
apply(apply(EPFgastos_2016,2,is.na),2,sum)
apply(apply(EPFhogar_2016,2,is.na),2,sum)
apply(apply(completedData_2016,2,is.na),2,sum)

# 2020
apply(apply(EPFgastos_2020,2,is.na),2,sum)
apply(apply(EPFhogar_2020,2,is.na),2,sum)
apply(apply(completedData_2020,2,is.na),2,sum)
```

Exportació de les bases de dades Clean

```
# 2006
write.table(EPFgastos_2006, file = "EPFgastos_2006.csv", sep = ";", na = "NA",
dec = ".", row.names = FALSE, col.names = TRUE)
write.table(EPFhogar_2006, file = "EPFhogar_2006.csv", sep = ";", na = "NA", dec
= ".", row.names = FALSE, col.names = TRUE)
write.table(completedData_2006, file = "EPFmhogar_2006.csv", sep = ";", na =
"NA", dec = ".", row.names = FALSE, col.names = TRUE)

# 2016
write.table(EPFgastos_2016, file = "EPFgastos_2016.csv", sep = ";", na = "NA",
dec = ".", row.names = FALSE, col.names = TRUE)
write.table(EPFhogar_2016, file = "EPFhogar_2016.csv", sep = ";", na = "NA", dec
= ".", row.names = FALSE, col.names = TRUE)
write.table(completedData_2016, file = "EPFmhogar_2016.csv", sep = ";", na =
"NA", dec = ".", row.names = FALSE, col.names = TRUE)

# 2020
write.table(EPFgastos_2020, file = "EPFgastos_2020.csv", sep = ";", na = "NA",
dec = ".", row.names = FALSE, col.names = TRUE)
write.table(EPFhogar_2020, file = "EPFhogar_2020.csv", sep = ";", na = "NA", dec
= ".", row.names = FALSE, col.names = TRUE)
write.table(completedData_2020, file = "EPFmhogar_2020.csv", sep = ";", na =
"NA", dec = ".", row.names = FALSE, col.names = TRUE)
```

Creació de base de dades resultant

Importació base de dades

```
EPFgastos_2006 <- read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers Preprocessing
/EPFgastos_2006.csv", sep=";")
EPFhogar_2006 <- read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers Preprocessing
/EPFhogar_2006.csv", sep=";")
EPFmhogar_2006 <- read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers Preprocessing
/EPFmhogar_2006.csv", sep=";")

EPFgastos_2016 <- read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers Preprocessing
/EPFgastos_2016.csv", sep=";")
```

```
EPFhogar_2016  <-  read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers  Preprocessing
/EPFhogar_2016.csv", sep=";")
EPFmhogar_2016  <-  read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers  Preprocessing
/EPFmhogar_2016.csv", sep=";")

EPFgastos_2020  <-  read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers  Preprocessing
/EPFgastos_2020.csv", sep=";")
EPFhogar_2020  <-  read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers  Preprocessing
/EPFhogar_2020.csv", sep=";")
EPFmhogar_2020  <-  read.csv("~/Desktop/TFG/Dades/Fitxers/Fitxers  Preprocessing
/EPFmhogar_2020.csv", sep=";")
```

VARIABLE NUMOCU

La base de dades del 2006, la variable NUMOCUP és un factor i el seu domini es [0,4]. On 4 correspon a 4 o més ocupats. En les altres dues bases de dades la variable es numèrica i en el cas de la base de dades el domini va de [0,6] a la base de dades de 2016 i de [0,5] a la del 2020. Es per aixo que necessitem que aquestes ultimes dues siguin factors i hem de subtituir els valors 5 i 6 per un 4.

```
summary(EPFhogar_2006$NUMOCUP)
summary(EPFhogar_2016$NUMOCU)
summary(EPFhogar_2020$NUMOCU)

EPFhogar_2016$NUMOCU[EPFhogar_2016$NUMOCU>4] <- 4
EPFhogar_2020$NUMOCU[EPFhogar_2020$NUMOCU>4] <- 4
```

Definició variables auxiliars

En aquest apartat el que es farà es a partir, de les variables seleccionades anteriorment. Calcularem i seleccionarem les variables auxiliars necessàries per elaborar la nostra base de dades final. Cal especificar, que al final d'aquest document, obtindrem 3 bases de dades diferents, una per cada any (bd_2006, bd_2016, bd_2020) on l'estructura de la base de dades i les variables seran les mateixes per aquestes 3 bases de dades, i finalment les unirem en una sola base de dades.

VARIABLE GASTOT

```
# 2016
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01111"] <- "1111"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01112"] <- "1112"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01113"] <- "1113"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01114"] <- "1114"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01115"] <- "1115"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01116"] <- "1116"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01117"] <- "1117"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01118"] <- "1118"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01121"] <- "1121"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01122"] <- "1122"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01123"] <- "1123"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01124"] <- "1124"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01125"] <- "1125"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01126"] <- "1126"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01127"] <- "1127"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01128"] <- "1128"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01131"] <- "1131"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01132"] <- "1132"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01133"] <- "1133"
```

```
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01134"] <- "1134"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01135"] <- "1135"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01136"] <- "1136"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01141"] <- "1141"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01142"] <- "1142"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01143"] <- "1143"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01144"] <- "1144"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01145"] <- "1145"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01146"] <- "1146"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01147"] <- "1147"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01151"] <- "1151"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01152"] <- "1152"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01153"] <- "1153"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01154"] <- "1154"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01155"] <- "1155"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01161"] <- "1161"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01162"] <- "1162"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01163"] <- "1163"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01164"] <- "1164"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01171"] <- "1171"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01172"] <- "1172"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01173"] <- "1173"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01174"] <- "1174"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01175"] <- "1175"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01176"] <- "1176"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01181"] <- "1181"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01182"] <- "1182"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01183"] <- "1183"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01184"] <- "1184"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01185"] <- "1185"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01186"] <- "1186"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01191"] <- "1191"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01192"] <- "1192"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01193"] <- "1193"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01194"] <- "1194"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01199"] <- "1199"

EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01211"] <- "1211"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01212"] <- "1212"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01213"] <- "1213"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01221"] <- "1221"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01222"] <- "1222"
EPFgastos_2016$CODIGO[EPFgastos_2016$CODIGO == "01223"] <- "1223"

# 2016
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01111"] <- "1111"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01112"] <- "1112"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01113"] <- "1113"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01114"] <- "1114"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01115"] <- "1115"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01116"] <- "1116"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01117"] <- "1117"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01118"] <- "1118"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01121"] <- "1121"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01122"] <- "1122"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01123"] <- "1123"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01124"] <- "1124"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01125"] <- "1125"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01126"] <- "1126"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01127"] <- "1127"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01128"] <- "1128"
```

```
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01131"] <- "1131"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01132"] <- "1132"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01133"] <- "1133"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01134"] <- "1134"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01135"] <- "1135"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01136"] <- "1136"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01141"] <- "1141"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01142"] <- "1142"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01143"] <- "1143"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01144"] <- "1144"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01145"] <- "1145"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01146"] <- "1146"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01147"] <- "1147"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01151"] <- "1151"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01152"] <- "1152"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01153"] <- "1153"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01154"] <- "1154"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01155"] <- "1155"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01161"] <- "1161"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01162"] <- "1162"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01163"] <- "1163"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01164"] <- "1164"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01171"] <- "1171"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01172"] <- "1172"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01173"] <- "1173"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01174"] <- "1174"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01175"] <- "1175"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01176"] <- "1176"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01181"] <- "1181"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01182"] <- "1182"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01183"] <- "1183"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01184"] <- "1184"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01185"] <- "1185"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01186"] <- "1186"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01191"] <- "1191"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01192"] <- "1192"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01193"] <- "1193"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01194"] <- "1194"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01199"] <- "1199"

EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01211"] <- "1211"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01212"] <- "1212"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01213"] <- "1213"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01221"] <- "1221"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01222"] <- "1222"
EPFgastos_2020$CODIGO[EPFgastos_2020$CODIGO == "01223"] <- "1223"

# Només agafem els CODIGOS de gasto corresponents al GRUP 1: ALIMENTS I BEGUDES
NO ALCOHÒLIQUES
# 2006
EPFgastos_2006 = EPFgastos_2006 %>% filter(CODIGO == "1111"|CODIGO =="1112"
|CODIGO  =="1113"|CODIGO  =="1114"  |CODIGO  =="1115"|CODIGO  =="1116"|CODIGO
=="1117"|CODIGO =="1118"|CODIGO =="1121"|CODIGO =="1122"|CODIGO =="1123"|CODIGO
=="1124"|CODIGO =="1125"|CODIGO =="1126"|CODIGO =="1127"|CODIGO =="1128"|CODIGO
=="1131"  |CODIGO  =="1132"  |CODIGO  =="1133"  |CODIGO  =="1134"  |CODIGO
=="1135"|CODIGO =="1136"|CODIGO =="1141"|CODIGO =="1142"|CODIGO =="1143"|CODIGO
=="1144"|CODIGO    =="1145"|CODIGO    =="1146"    |CODIGO    =="1147"|CODIGO
=="1151"|CODIGO =="1152"|CODIGO =="1153"|CODIGO =="1154"|CODIGO =="1155"|CODIGO
=="1161"|CODIGO =="1162"|CODIGO =="1163"|CODIGO =="1164"|CODIGO =="1171"|CODIGO
=="1172"|CODIGO =="1173"|CODIGO =="1174"|CODIGO =="1175"|CODIGO =="1176"|CODIGO
=="1181"|CODIGO =="1182"|CODIGO =="1183"|CODIGO =="1184"|CODIGO =="1185"|CODIGO
```

```
=="1186"|CODIGO =="1191"|CODIGO =="1192"|CODIGO =="1193"|CODIGO =="1194"|CODIGO
=="1199"|CODIGO =="1211"|CODIGO =="1212"|CODIGO =="1213"|CODIGO =="1221"|CODIGO
=="1222"|CODIGO =="1223")

# 2016
EPFgastos_2016 = EPFgastos_2016 %>% filter(CODIGO == "1111"|CODIGO =="1112"
|CODIGO =="1113"|CODIGO =="1114" |CODIGO =="1115"|CODIGO =="1116"|CODIGO
=="1117"|CODIGO =="1118"|CODIGO =="1121"|CODIGO =="1122"|CODIGO =="1123"|CODIGO
=="1124"|CODIGO =="1125"|CODIGO =="1126"|CODIGO =="1127"|CODIGO =="1128"|CODIGO
=="1131" |CODIGO =="1132" |CODIGO =="1133" |CODIGO =="1134" |CODIGO
=="1135"|CODIGO =="1136"|CODIGO =="1141"|CODIGO =="1142"|CODIGO =="1143"|CODIGO
=="1144"|CODIGO =="1145"|CODIGO =="1146" |CODIGO =="1147"|CODIGO
=="1151"|CODIGO =="1152"|CODIGO =="1153"|CODIGO =="1154"|CODIGO =="1155"|CODIGO
=="1161"|CODIGO =="1162"|CODIGO =="1163"|CODIGO =="1164"|CODIGO =="1171"|CODIGO
=="1172"|CODIGO =="1173"|CODIGO =="1174"|CODIGO =="1175"|CODIGO =="1176"|CODIGO
=="1181"|CODIGO =="1182"|CODIGO =="1183"|CODIGO =="1184"|CODIGO =="1185"|CODIGO
=="1186"|CODIGO =="1191"|CODIGO =="1192"|CODIGO =="1193"|CODIGO =="1194"|CODIGO
=="1199"|CODIGO =="1211"|CODIGO =="1212"|CODIGO =="1213"|CODIGO =="1221"|CODIGO
=="1222"|CODIGO =="1223")

# 2020
EPFgastos_2020 = EPFgastos_2020 %>% filter(CODIGO == "1111"|CODIGO =="1112"
|CODIGO =="1113"|CODIGO =="1114" |CODIGO =="1115"|CODIGO =="1116"|CODIGO
=="1117"|CODIGO =="1118"|CODIGO =="1121"|CODIGO =="1122"|CODIGO =="1123"|CODIGO
=="1124"|CODIGO =="1125"|CODIGO =="1126"|CODIGO =="1127"|CODIGO =="1128"|CODIGO
=="1131" |CODIGO =="1132" |CODIGO =="1133" |CODIGO =="1134" |CODIGO
=="1135"|CODIGO =="1136"|CODIGO =="1141"|CODIGO =="1142"|CODIGO =="1143"|CODIGO
=="1144"|CODIGO =="1145"|CODIGO =="1146" |CODIGO =="1147"|CODIGO
=="1151"|CODIGO =="1152"|CODIGO =="1153"|CODIGO =="1154"|CODIGO =="1155"|CODIGO
=="1161"|CODIGO =="1162"|CODIGO =="1163"|CODIGO =="1164"|CODIGO =="1171"|CODIGO
=="1172"|CODIGO =="1173"|CODIGO =="1174"|CODIGO =="1175"|CODIGO =="1176"|CODIGO
=="1181"|CODIGO =="1182"|CODIGO =="1183"|CODIGO =="1184"|CODIGO =="1185"|CODIGO
=="1186"|CODIGO =="1191"|CODIGO =="1192"|CODIGO =="1193"|CODIGO =="1194"|CODIGO
=="1199"|CODIGO =="1211"|CODIGO =="1212"|CODIGO =="1213"|CODIGO =="1221"|CODIGO
=="1222"|CODIGO =="1223")

# VARIABLE GASTOT
GASTOT_2006 = EPFgastos_2006 %>% group_by(NUMERO) %>%
summarise(GASTOT=sum(GASTO))
GASTOT_2016 = EPFgastos_2016 %>% group_by(NUMERO) %>%
summarise(GASTOT=sum(GASTO))
GASTOT_2020 = EPFgastos_2020 %>% group_by(NUMERO) %>%
summarise(GASTOT=sum(GASTO))
```

VARIABLE EDAD

```
EDAD_2006 = EPFmhogar_2006 %>% group_by(NUMERO) %>% summarise(EDAD = mean(EDAD))
EDAD_2016 = EPFmhogar_2016 %>% group_by(NUMERO) %>% summarise(EDAD = mean(EDAD))
EDAD_2020 = EPFmhogar_2020 %>% group_by(NUMERO) %>% summarise(EDAD = mean(EDAD))
```

VARIABLE SEXO

```
## 2006
DONES_2006 = EPFmhogar_2006 %>%
  filter(SEXO == 1) %>%
  group_by(NUMERO) %>%
  summarise(nDONES=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

SEX_2006 = MEMBRES_2006 %>% left_join(DONES_2006)
SEX_2006$nDONES[is.na(SEX_2006$nDONES)] <- 0 # Tota la llar esta formada per
homes
SEX_2006 = SEX_2006 %>% mutate(SEXO = nDONES/nTOTAL*100)
```

```
## 2016
DONES_2016 = EPFmhogar_2016 %>%
  filter(SEXO == 1) %>%
  group_by(NUMERO) %>%
  summarise(nDONES=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

SEX_2016 = MEMBRES_2016 %>% left_join(DONES_2016)
SEX_2016$nDONES[is.na(SEX_2016$nDONES)] <- 0 # Tota la llar esta formada per
homes
SEX_2016 = SEX_2016 %>% mutate(SEXO = nDONES/nTOTAL*100)

## 2020
DONES_2020 = EPFmhogar_2020 %>%
  filter(SEXO == 1) %>%
  group_by(NUMERO) %>%
  summarise(nDONES=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

SEX_2020 = MEMBRES_2020 %>% left_join(DONES_2020)
SEX_2020$nDONES[is.na(SEX_2020$nDONES)] <- 0 # Tota la llar esta formada per
homes
SEX_2020 = SEX_2020 %>% mutate(SEXO = nDONES/nTOTAL*100)

# La variable SEXO no presenta missings
sum(is.na(SEX_2006$SEXO))
sum(is.na(SEX_2016$SEXO))
sum(is.na(SEX_2020$SEXO))
```

VARIABLE NACION

```
## 2006
ESP_2006 = EPFmhogar_2006 %>%
  filter(NACIONA == 1|NACIONA == 3) %>%
  group_by(NUMERO) %>%
  summarise(nESP=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NACI_2006 = MEMBRES_2006 %>% left_join(ESP_2006)
NACI_2006$nESP[is.na(NACI_2006$nESP)] <- 0 # Tota la llar esta formada per
individus amb nacionalitat estranjera.
NACI_2006 = NACI_2006 %>% mutate(NACION = nESP/nTOTAL*100)

## 2016
ESP_2016 = EPFmhogar_2016 %>%
  filter(NACIONA == 1|NACIONA == 3) %>%
  group_by(NUMERO) %>%
  summarise(nESP=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NACI_2016 = MEMBRES_2016 %>% left_join(ESP_2016)
NACI_2016$nESP[is.na(NACI_2016$nESP)] <- 0 # Tota la llar esta formada per
individus amb nacionalitat estranjera.
NACI_2016 = NACI_2016 %>% mutate(NACION = nESP/nTOTAL*100)
```

```
## 2020
ESP_2020 = EPFmhogar_2020 %>%
  filter(NACIONA == 1|NACIONA == 3) %>%
  group_by(NUMERO) %>%
  summarise(nESP=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NACI_2020 = MEMBRES_2020 %>% left_join(ESP_2020)
NACI_2020$nESP[is.na(NACI_2020$nESP)] <- 0 # Tota la llar esta formada per
individus amb nacionalitat estranjera.
NACI_2020 = NACI_2020 %>% mutate(NACION = nESP/nTOTAL*100)

# La variable NACION no presenta missings
sum(is.na(NACI_2006$NACION))
sum(is.na(NACI_2016$NACION))
sum(is.na(NACI_2020$NACION))
```

VARIABLE ZONARESI

```
EPFhogar_2006$ZONARES[EPFhogar_2006$ZONARES>=1 & EPFhogar_2006$ZONARES<=4] <- 1
EPFhogar_2016$ZONARES[EPFhogar_2016$ZONARES>=1 & EPFhogar_2016$ZONARES<=4] <- 1
EPFhogar_2020$ZONARES[EPFhogar_2020$ZONARES>=1 & EPFhogar_2020$ZONARES<=4] <- 1

EPFhogar_2006$ZONARES[EPFhogar_2006$ZONARES>4] <- 2
EPFhogar_2016$ZONARES[EPFhogar_2016$ZONARES>4] <- 2
EPFhogar_2020$ZONARES[EPFhogar_2020$ZONARES>4] <- 2

VARIABLE DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARES

EPFhogar_2016$CCAA[EPFhogar_2016$CCAA==19] <- 18
EPFhogar_2020$CCAA[EPFhogar_2020$CCAA==19] <- 18

EPFhogar_2006$ECIVILSP[EPFhogar_2006$ECIVILSP==5] <- 4
EPFhogar_2016$ECIVILLEGALSP[EPFhogar_2016$ECIVILLEGALSP==5] <- 4
EPFhogar_2020$ECIVILLEGALSP[EPFhogar_2020$ECIVILLEGALSP==5] <- 4


EPFhogar_2006$CCAA[EPFhogar_2006$CCAA  ==  2  |  EPFhogar_2006$CCAA  ==  9  |
EPFhogar_2006$CCAA == 15 | EPFhogar_2006$CCAA == 16 | EPFhogar_2006$CCAA == 13
| EPFhogar_2006$CCAA == 17] <-  99

EPFhogar_2006$CCAA[EPFhogar_2006$CCAA  ==  10  |  EPFhogar_2006$CCAA  ==  3  |
EPFhogar_2006$CCAA == 12 | EPFhogar_2006$CCAA == 4 | EPFhogar_2006$CCAA == 6 |
EPFhogar_2006$CCAA == 7] <-  999

EPFhogar_2006$CCAA[EPFhogar_2006$CCAA  ==  14  |  EPFhogar_2006$CCAA  ==  18  |
EPFhogar_2006$CCAA == 8 | EPFhogar_2006$CCAA == 11 | EPFhogar_2006$CCAA ==1  |
EPFhogar_2006$CCAA == 5] <- 9999

EPFhogar_2016$CCAA[EPFhogar_2016$CCAA==2     |     EPFhogar_2016$CCAA==9     |
EPFhogar_2016$CCAA==15  |  EPFhogar_2016$CCAA==16  |  EPFhogar_2016$CCAA==13  |
EPFhogar_2016$CCAA==17] <- 99

EPFhogar_2016$CCAA[EPFhogar_2016$CCAA==10     |     EPFhogar_2016$CCAA==3     |
EPFhogar_2016$CCAA==12  |  EPFhogar_2016$CCAA==4  |  EPFhogar_2016$CCAA==6  |
EPFhogar_2016$CCAA==7] <- 999

EPFhogar_2016$CCAA[EPFhogar_2016$CCAA==14     |     EPFhogar_2016$CCAA==18     |
EPFhogar_2016$CCAA==8  |  EPFhogar_2016$CCAA==11  |  EPFhogar_2016$CCAA==1  |
EPFhogar_2016$CCAA==5] <- 9999

EPFhogar_2020$CCAA[EPFhogar_2020$CCAA==2     |     EPFhogar_2020$CCAA==9     |
EPFhogar_2020$CCAA==15  |  EPFhogar_2020$CCAA==16  |  EPFhogar_2020$CCAA==13  |
EPFhogar_2020$CCAA==17] <- 99
```

```
EPFhogar_2020$CCAA[EPFhogar_2020$CCAA==10     |     EPFhogar_2020$CCAA==3     |
EPFhogar_2020$CCAA==12   |   EPFhogar_2020$CCAA==4   |   EPFhogar_2020$CCAA==6   |
EPFhogar_2020$CCAA==7] <- 999

EPFhogar_2020$CCAA[EPFhogar_2020$CCAA==14     |     EPFhogar_2020$CCAA==18     |
EPFhogar_2020$CCAA==8   |   EPFhogar_2020$CCAA==11   |   EPFhogar_2020$CCAA==1   |
EPFhogar_2020$CCAA==5] <- 9999

EPFhogar_2006$CCAA[EPFhogar_2006$CCAA==99] <- 1
EPFhogar_2006$CCAA[EPFhogar_2006$CCAA==999] <- 2
EPFhogar_2006$CCAA[EPFhogar_2006$CCAA==9999] <- 3

EPFhogar_2016$CCAA[EPFhogar_2016$CCAA==99] <- 1
EPFhogar_2016$CCAA[EPFhogar_2016$CCAA==999] <- 2
EPFhogar_2016$CCAA[EPFhogar_2016$CCAA==9999] <- 3

EPFhogar_2020$CCAA[EPFhogar_2020$CCAA==99] <- 1
EPFhogar_2020$CCAA[EPFhogar_2020$CCAA==999] <- 2
EPFhogar_2020$CCAA[EPFhogar_2020$CCAA==9999] <- 3

DENSI_CCAA_NUMOCU_ECIVIL_2006 = EPFhogar_2006 %>% select(NUMERO, DENSI, CCAA,
COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARES)
DENSI_CCAA_NUMOCU_ECIVIL_2016 = EPFhogar_2016 %>% select(NUMERO, DENSIDAD, CCAA,
COMITOT, NUMOCU, IMPEXAC, ECIVILLEGALSP, ZONARES)
DENSI_CCAA_NUMOCU_ECIVIL_2020 = EPFhogar_2020 %>% select(NUMERO, DENSIDAD, CCAA,
COMITOT, NUMOCU, IMPEXAC, ECIVILLEGALSP, ZONARES)
```

VARIABLE NMIEMBR1-NMIEMBR3

```
## 2006
NMIEMB1_2006 = EPFmhogar_2006 %>%
  filter(EDAD>=0 & EDAD<=15) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR1=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR1_2006 = MEMBRES_2006 %>% left_join(NMIEMB1_2006)
NMIEMBR1_2006$NMIEMBR1[is.na(NMIEMBR1_2006$NMIEMBR1)] <- 0
NMIEMBR1_2006 = NMIEMBR1_2006 %>% mutate(NMIEMBR1 = NMIEMBR1/nTOTAL*100)

NMIEMB2_2006 = EPFmhogar_2006 %>%
  filter(EDAD>=16 & EDAD<=64) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR2=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR2_2006 = MEMBRES_2006 %>% left_join(NMIEMB2_2006)
NMIEMBR2_2006$NMIEMBR2[is.na(NMIEMBR2_2006$NMIEMBR2)] <- 0
NMIEMBR2_2006 = NMIEMBR2_2006 %>% mutate(NMIEMBR2 = NMIEMBR2/nTOTAL*100)

NMIEMB3_2006 = EPFmhogar_2006 %>%
  filter(EDAD>=65) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR3=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR3_2006 = MEMBRES_2006 %>% left_join(NMIEMB3_2006)
```

```
NMIEMBR3_2006$NMIEMBR3[is.na(NMIEMBR3_2006$NMIEMBR3)] <- 0
NMIEMBR3_2006 = NMIEMBR3_2006 %>% mutate(NMIEMBR3 = NMIEMBR3/nTOTAL*100)

## 2016
NMIEMB1_2016 = EPFmhogar_2016 %>%
  filter(EDAD>=0 & EDAD<=15) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR1=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR1_2016 = MEMBRES_2016 %>% left_join(NMIEMB1_2016)
NMIEMBR1_2016$NMIEMBR1[is.na(NMIEMBR1_2016$NMIEMBR1)] <- 0
NMIEMBR1_2016 = NMIEMBR1_2016 %>% mutate(NMIEMBR1 = NMIEMBR1/nTOTAL*100)

NMIEMB2_2016 = EPFmhogar_2016 %>%
  filter(EDAD>=16 & EDAD<=64) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR2=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR2_2016 = MEMBRES_2016 %>% left_join(NMIEMB2_2016)
NMIEMBR2_2016$NMIEMBR2[is.na(NMIEMBR2_2016$NMIEMBR2)] <- 0
NMIEMBR2_2016 = NMIEMBR2_2016 %>% mutate(NMIEMBR2 = NMIEMBR2/nTOTAL*100)

NMIEMB3_2016 = EPFmhogar_2016 %>%
  filter(EDAD>=65) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR3=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR3_2016 = MEMBRES_2016 %>% left_join(NMIEMB3_2016)
NMIEMBR3_2016$NMIEMBR3[is.na(NMIEMBR3_2016$NMIEMBR3)] <- 0
NMIEMBR3_2016 = NMIEMBR3_2016 %>% mutate(NMIEMBR3 = NMIEMBR3/nTOTAL*100)

## 2020
NMIEMB1_2020 = EPFmhogar_2020 %>%
  filter(EDAD>=0 & EDAD<=15) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR1=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR1_2020 = MEMBRES_2020 %>% left_join(NMIEMB1_2020)
NMIEMBR1_2020$NMIEMBR1[is.na(NMIEMBR1_2020$NMIEMBR1)] <- 0
NMIEMBR1_2020 = NMIEMBR1_2020 %>% mutate(NMIEMBR1 = NMIEMBR1/nTOTAL*100)

NMIEMB2_2020 = EPFmhogar_2020 %>%
  filter(EDAD>=16 & EDAD<=64) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR2=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR2_2020 = MEMBRES_2020 %>% left_join(NMIEMB2_2020)
NMIEMBR2_2020$NMIEMBR2[is.na(NMIEMBR2_2020$NMIEMBR2)] <- 0
```

```
NMIEMBR2_2020 = NMIEMBR2_2020 %>% mutate(NMIEMBR2 = NMIEMBR2/nTOTAL*100)

NMIEMB3_2020 = EPFmhogar_2020 %>%
  filter(EDAD>=65) %>%
  group_by(NUMERO) %>%
  summarise(NMIEMBR3=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

NMIEMBR3_2020 = MEMBRES_2020 %>% left_join(NMIEMB3_2020)
NMIEMBR3_2020$NMIEMBR3[is.na(NMIEMBR3_2020$NMIEMBR3)] <- 0
NMIEMBR3_2020 = NMIEMBR3_2020 %>% mutate(NMIEMBR3 = NMIEMBR3/nTOTAL*100)
```

VARIABLE ESTUDIOS1-ESTUDIOS3

```
## 2006
ESTUDIO1_2006 = EPFmhogar_2006 %>%
  filter(ESTUDIOS == 1) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO1=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

ESTUDIOS1_2006 = MEMBRES_2006 %>% left_join(ESTUDIO1_2006)
ESTUDIOS1_2006$ESTUDIO1[is.na(ESTUDIOS1_2006$ESTUDIO1)] <- 0
ESTUDIOS1_2006 = ESTUDIOS1_2006 %>% mutate(ESTUDIOS1 = ESTUDIO1/nTOTAL*100)

ESTUDIO2_2006 = EPFmhogar_2006 %>%
  filter(ESTUDIOS == 2 | ESTUDIOS == 3) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO2=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

ESTUDIOS2_2006 = MEMBRES_2006 %>% left_join(ESTUDIO2_2006)
ESTUDIOS2_2006$ESTUDIO2[is.na(ESTUDIOS2_2006$ESTUDIO2)] <- 0
ESTUDIOS2_2006 = ESTUDIOS2_2006 %>% mutate(ESTUDIOS2 = ESTUDIO2/nTOTAL*100)

ESTUDIO3_2006 = EPFmhogar_2006 %>%
  filter(ESTUDIOS > 3) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO3=n())

MEMBRES_2006 = EPFmhogar_2006 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

ESTUDIOS3_2006 = MEMBRES_2006 %>% left_join(ESTUDIO3_2006)
ESTUDIOS3_2006$ESTUDIO3[is.na(ESTUDIOS3_2006$ESTUDIO3)] <- 0
ESTUDIOS3_2006 = ESTUDIOS3_2006 %>% mutate(ESTUDIOS3 = ESTUDIO3/nTOTAL*100)

## 2016
ESTUDIO1_2016 = EPFmhogar_2016 %>%
  filter(ESTUDIOS == 1) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO1=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())
```

```
ESTUDIOS1_2016 = MEMBRES_2016 %>% left_join(ESTUDIO1_2016)
ESTUDIOS1_2016$ESTUDIO1[is.na(ESTUDIOS1_2016$ESTUDIO1)] <- 0
ESTUDIOS1_2016 = ESTUDIOS1_2016 %>% mutate(ESTUDIOS1 = ESTUDIO1/nTOTAL*100)


ESTUDIO2_2016 = EPFmhogar_2016 %>%
  filter(ESTUDIOS == 2 | ESTUDIOS == 3) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO2=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

ESTUDIOS2_2016 = MEMBRES_2016 %>% left_join(ESTUDIO2_2016)
ESTUDIOS2_2016$ESTUDIO2[is.na(ESTUDIOS2_2016$ESTUDIO2)] <- 0
ESTUDIOS2_2016 = ESTUDIOS2_2016 %>% mutate(ESTUDIOS2 = ESTUDIO2/nTOTAL*100)

ESTUDIO3_2016 = EPFmhogar_2016 %>%
  filter(ESTUDIOS > 3) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO3=n())

MEMBRES_2016 = EPFmhogar_2016 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

ESTUDIOS3_2016 = MEMBRES_2016 %>% left_join(ESTUDIO3_2016)
ESTUDIOS3_2016$ESTUDIO3[is.na(ESTUDIOS3_2016$ESTUDIO3)] <- 0
ESTUDIOS3_2016 = ESTUDIOS3_2016 %>% mutate(ESTUDIOS3 = ESTUDIO3/nTOTAL*100)

## 2020
ESTUDIO1_2020 = EPFmhogar_2020 %>%
  filter(ESTUDIOS == 1) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO1=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

ESTUDIOS1_2020 = MEMBRES_2020 %>% left_join(ESTUDIO1_2020)
ESTUDIOS1_2020$ESTUDIO1[is.na(ESTUDIOS1_2020$ESTUDIO1)] <- 0
ESTUDIOS1_2020 = ESTUDIOS1_2020 %>% mutate(ESTUDIOS1 = ESTUDIO1/nTOTAL*100)


ESTUDIO2_2020 = EPFmhogar_2020 %>%
  filter(ESTUDIOS == 2 | ESTUDIOS == 3) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO2=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())

ESTUDIOS2_2020 = MEMBRES_2020 %>% left_join(ESTUDIO2_2020)
ESTUDIOS2_2020$ESTUDIO2[is.na(ESTUDIOS2_2020$ESTUDIO2)] <- 0
ESTUDIOS2_2020 = ESTUDIOS2_2020 %>% mutate(ESTUDIOS2 = ESTUDIO2/nTOTAL*100)


ESTUDIO3_2020 = EPFmhogar_2020 %>%
  filter(ESTUDIOS > 3) %>%
  group_by(NUMERO) %>%
  summarise(ESTUDIO3=n())

MEMBRES_2020 = EPFmhogar_2020 %>%
  group_by(NUMERO) %>%
  summarise(nTOTAL=n())
```

```
ESTUDIOS3_2020 = MEMBRES_2020 %>% left_join(ESTUDIO3_2020)
ESTUDIOS3_2020$ESTUDIO3[is.na(ESTUDIOS3_2020$ESTUDIO3)] <- 0
ESTUDIOS3_2020 = ESTUDIOS3_2020 %>% mutate(ESTUDIOS3 = ESTUDIO3/nTOTAL*100)
```

Base de dades conjunta de variables auxiliars

Un cop creades totes les variables auxiliars procedirem a realitzar la unio de la base de dades dels respectius anys:

```
# 2006
bd_2006 = EPFgastos_2006 %>% select(ANOENC, NUMERO, GASTO, CODIGO)
bd_2006 = bd_2006 %>% left_join(GASTOT_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(EDAD_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(SEX_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(NACI_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(DENSI_CCAA_NUMOCU_ECIVIL_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(NMIEMBR1_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(NMIEMBR2_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(NMIEMBR3_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(ESTUDIOS1_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(ESTUDIOS2_2006, by="NUMERO")
bd_2006 = bd_2006 %>% left_join(ESTUDIOS3_2006, by="NUMERO")

bd_2006 = bd_2006 %>% select(ANOENC, NUMERO, CODIGO, GASTO, GASTOT, EDAD, SEXO,
NACION, DENSI, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARES, NMIEMBR1,
NMIEMBR2, NMIEMBR3, ESTUDIOS1, ESTUDIOS2, ESTUDIOS3)

# ----------------------------------------
# NETEJA BASE DE DADES FINAL 2006
# ----------------------------------------
apply(apply(bd_2006,2,is.na),2,sum)

# VARIABLE DENSI_CCAA_NUMOCU_ECIVIL_2006
# Hi hauran missings perquè recordem que en el document anterior hi havien molts
pocs missings a la base de dades EPFhogar_2006 i vam decidir eliminar-los, per
tant aquestes llars també s'han d'eliminar de la base de dades EPFhogar_2006
bd_2006 <- bd_2006[-which(is.na(bd_2006$DENSI)),]
# Comprovem que tota la base de dades no conté missings
apply(apply(bd_2006,2,is.na),2,sum)

# 2016
bd_2016 = EPFgastos_2016 %>% select(ANOENC, NUMERO, GASTO, CODIGO)
bd_2016 = bd_2016 %>% left_join(GASTOT_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(EDAD_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(SEX_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(NACI_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(DENSI_CCAA_NUMOCU_ECIVIL_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(NMIEMBR1_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(NMIEMBR2_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(NMIEMBR3_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(ESTUDIOS1_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(ESTUDIOS2_2016, by="NUMERO")
bd_2016 = bd_2016 %>% left_join(ESTUDIOS3_2016, by="NUMERO")

bd_2016 = bd_2016 %>% select(ANOENC, NUMERO, CODIGO, GASTO, GASTOT, EDAD, SEXO,
NACION, DENSIDAD, CCAA, COMITOT, NUMOCU, IMPEXAC, ECIVILLEGALSP, ZONARES,
NMIEMBR1, NMIEMBR2, NMIEMBR3, ESTUDIOS1, ESTUDIOS2, ESTUDIOS3)

# ----------------------------------------
# NETEJA BASE DE DADES FINAL 2016
# ----------------------------------------
apply(apply(bd_2016,2,is.na),2,sum)

# VARIABLE TAMAMU_CCAA_NUMOCU_ECIVIL_2006
```

```
# Hi hauran missings perquè recordem que en el document anterior hi havien
missings a la base de dades EPFhogar_2016 i vam decidir eliminar-los, per tant
aquestes llars també s'han d'eliminar de la base de dades EPFhogar_2016
bd_2016 <- bd_2016[-which(is.na(bd_2016$DENSIDAD)),]
# Comprovem que tota la base de dades no conté missings
apply(apply(bd_2016,2,is.na),2,sum)

# 2020
bd_2020 = EPFgastos_2020 %>% select(ANOENC, NUMERO, GASTO, CODIGO)
bd_2020 = bd_2020 %>% left_join(GASTOT_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(EDAD_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(SEX_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(NACI_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(DENSI_CCAA_NUMOCU_ECIVIL_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(NMIEMBR1_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(NMIEMBR2_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(NMIEMBR3_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(ESTUDIOS1_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(ESTUDIOS2_2020, by="NUMERO")
bd_2020 = bd_2020 %>% left_join(ESTUDIOS3_2020, by="NUMERO")

bd_2020 = bd_2020 %>% select(ANOENC, NUMERO, CODIGO, GASTO, GASTOT, EDAD, SEXO,
NACION, DENSIDAD, CCAA, COMITOT, NUMOCU, IMPEXAC, ECIVILLEGALSP, ZONARES,
NMIEMBR1, NMIEMBR2, NMIEMBR3, ESTUDIOS1, ESTUDIOS2, ESTUDIOS3)

# ------------------------------------------
# NETEJA BASE DE DADES FINAL 2020
# ------------------------------------------
apply(apply(bd_2020,2,is.na),2,sum)
# VARIABLE TAMAMU_CCAA_NUMOCU_ECIVIL_2020
# Hi hauran missings perque recordem que en el document anterior hi havien
missings a la base de dades EPFhogar_2016 i vam decidir elimnar-los, per tant
aquestes llars també s'han d'eliminar de la base de dades EPFhogar_2016
bd_2020 <- bd_2020[-which(is.na(bd_2020$DENSIDAD)),]
# Comprovem que tota la base de dades no conté missings
apply(apply(bd_2020,2,is.na),2,sum)
```

Finalment, un cop tenim les tres bases de dades netejades. Procedirem a unir aquestes tres bases de dades en una sola:

```
# Classifiquem cada una de les variables
# 2006
bd_2006$ANOENC <- as.factor(bd_2006$ANOENC)
bd_2006$NUMERO <- as.factor(bd_2006$NUMERO)
bd_2006$GASTO <-  as.integer(bd_2006$GASTO)
bd_2006$CODIGO <- as.factor(bd_2006$CODIGO)
bd_2006$EDAD <- as.integer(bd_2006$EDAD)
bd_2006$SEXO <- as.integer(bd_2006$SEXO)
bd_2006$NACION <- as.integer(bd_2006$NACION)
bd_2006$DENSI <- as.factor(bd_2006$DENSI)
bd_2006$ZONARES <- as.factor(bd_2006$ZONARES)
bd_2006$CCAA <- as.factor(bd_2006$CCAA)
bd_2006$COMITOT <- as.integer(bd_2006$COMITOT)
bd_2006$NUMOCUP <-  as.integer(bd_2006$NUMOCUP)
bd_2006$IMPEXAC <- as.integer(bd_2006$IMPEXAC)
bd_2006$ECIVILSP <- as.factor(bd_2006$ECIVILSP)
bd_2006$NMIEMBR1 <- as.integer(bd_2006$NMIEMBR1)
bd_2006$NMIEMBR2 <- as.integer(bd_2006$NMIEMBR2)
bd_2006$NMIEMBR3 <- as.integer(bd_2006$NMIEMBR3)
bd_2006$ESTUDIOS1 <- as.integer(bd_2006$ESTUDIOS1)
bd_2006$ESTUDIOS2 <- as.integer(bd_2006$ESTUDIOS2)
bd_2006$ESTUDIOS3 <- as.integer(bd_2006$ESTUDIOS3)

# 2016
bd_2016$ANOENC <- as.factor(bd_2016$ANOENC)
bd_2016$NUMERO <- as.factor(bd_2016$NUMERO)
```

```
bd_2016$GASTO <-  as.integer(bd_2016$GASTO)
bd_2016$CODIGO <- as.factor(bd_2016$CODIGO)
bd_2016$EDAD <- as.integer(bd_2016$EDAD)
bd_2016$SEXO <- as.integer(bd_2016$SEXO)
bd_2016$NACION <- as.integer(bd_2016$NACION)
bd_2016$DENSIDAD <- as.factor(bd_2016$DENSIDAD)
bd_2016$ZONARES <- as.factor(bd_2016$ZONARES)
bd_2016$CCAA <- as.factor(bd_2016$CCAA)
bd_2016$COMITOT <- as.integer(bd_2016$COMITOT)
bd_2016$NUMOCU <-  as.integer(bd_2016$NUMOCU)
bd_2016$IMPEXAC <- as.integer(bd_2016$IMPEXAC)
bd_2016$ECIVILLEGALSP <- as.factor(bd_2016$ECIVILLEGALSP)
bd_2016$NMIEMBR1 <- as.integer(bd_2016$NMIEMBR1)
bd_2016$NMIEMBR2 <- as.integer(bd_2016$NMIEMBR2)
bd_2016$NMIEMBR3 <- as.integer(bd_2016$NMIEMBR3)
bd_2016$ESTUDIOS1 <- as.integer(bd_2016$ESTUDIOS1)
bd_2016$ESTUDIOS2 <- as.integer(bd_2016$ESTUDIOS2)
bd_2016$ESTUDIOS3 <- as.integer(bd_2016$ESTUDIOS3)

# 2020
bd_2020$ANOENC <- as.factor(bd_2020$ANOENC)
bd_2020$NUMERO <- as.factor(bd_2020$NUMERO)
bd_2020$GASTO <-  as.integer(bd_2020$GASTO)
bd_2020$CODIGO <- as.factor(bd_2020$CODIGO)
bd_2020$EDAD <- as.integer(bd_2020$EDAD)
bd_2020$SEXO <- as.integer(bd_2020$SEXO)
bd_2020$NACION <- as.integer(bd_2020$NACION)
bd_2020$DENSIDAD <- as.factor(bd_2020$DENSIDAD)
bd_2020$ZONARES <- as.factor(bd_2020$ZONARES)
bd_2020$CCAA <- as.factor(bd_2020$CCAA)
bd_2020$COMITOT <- as.integer(bd_2020$COMITOT)
bd_2020$NUMOCU <-  as.integer(bd_2020$NUMOCU)
bd_2020$IMPEXAC <- as.integer(bd_2020$IMPEXAC)
bd_2020$ECIVILLEGALSP <- as.factor(bd_2020$ECIVILLEGALSP)
bd_2020$NMIEMBR1 <- as.integer(bd_2020$NMIEMBR1)
bd_2020$NMIEMBR2 <- as.integer(bd_2020$NMIEMBR2)
bd_2020$NMIEMBR3 <- as.integer(bd_2020$NMIEMBR3)
bd_2020$ESTUDIOS1 <- as.integer(bd_2020$ESTUDIOS1)
bd_2020$ESTUDIOS2 <- as.integer(bd_2020$ESTUDIOS2)
bd_2020$ESTUDIOS3 <- as.integer(bd_2020$ESTUDIOS3)

# Canviem nom de les variables
colnames(bd_2006)[9] <- "DENSIDAD"
colnames(bd_2016)[14] <- "ECIVILSP"
colnames(bd_2020)[14] <- "ECIVILSP"

colnames(bd_2016)[12] <- "NUMOCUP"
colnames(bd_2020)[12] <- "NUMOCUP"

bd <- bind_rows(bd_2006, bd_2016, bd_2020)
summary(bd)
```

Exportació de la base de dades netejada

```
write.table(bd, file = "EPFClean.csv", sep = ";", na = "NA", dec = ".", row.names
= FALSE, col.names = TRUE)
```

Normalització de la base de dades resultant

Instal·lació de llibreries

```
# install.packages("modeest")
# install.packages("descr")
```

```
#install.packages("RColorBrewer")
library(lubridate)
library(ggplot2)
library(modeest)
library(descr)
library(RColorBrewer)
```

Importació de la base de dades

A continuació ens construirem una nova base de dades que només ens mostra els codis dels diferents aliments i a més creem una nova variable que ens indiqui el subgrup al que pertany.

Filtratge i base de dades

```
# ECOICOP -> GRUPO 1.  ALIMENTOS Y BEBIDAS NO ALCOHÓLICAS
### GRUP
nom_grup <- data.frame(CODIGO = c("1111", "1112", "1113", "1114", "1115",
"1116", "1117", "1118", "1121", "1122", "1123", "1124", "1125", "1126", "1127",
"1128","1131", "1132", "1133", "1134", "1135", "1136", "1141","1142", "1143",
"1144", "1145", "1146", "1147", "1151", "1152", "1153", "1154", "1155", "1161",
"1162", "1163", "1164", "1171", "1172", "1173", "1174", "1175", "1176", "1181",
"1182", "1183", "1184", "1185", "1186", "1191", "1192", "1193", "1194", "1199",
"1211", "1212", "1213", "1221", "1222", "1223"), GRUP = c(rep("Pa i cereals",
8), rep("Carn", 8), rep("Peix i Marisc", 6), rep("Llet, Formatges i Ous", 7),
rep("Olis i greixos", 5), rep("Fruits", 4), rep("Llegums i verdures", 6),
rep("Sucre, Confitura, Mel, Xocolata i Confiteria", 6), rep("Productes n.c.o.p",
5), rep("Begudes no alcoholíques", 6)))

### GRUP ESPECIFIC
nom_espe <- data.frame(CODIGO = c("1111", "1112", "1113", "1114", "1115",
"1116", "1117", "1118", "1121", "1122", "1123", "1124", "1125", "1126", "1127",
"1128","1131", "1132", "1133", "1134", "1135", "1136", "1141","1142", "1143",
"1144", "1145", "1146", "1147", "1151", "1152", "1153", "1154", "1155", "1161",
"1162", "1163", "1164", "1171", "1172", "1173", "1174", "1175", "1176", "1181",
"1182", "1183", "1184", "1185", "1186", "1191", "1192", "1193", "1194", "1199",
"1211", "1212", "1213", "1221", "1222", "1223"), SUBGRUP = c("Arroç", "Farina i
altres cereals", "Pà", "Altres productes fleca", "Pizza i quiche", "Pastes
alimentàries i cuscús", "Cereals d'esmorzar", "Altres productes basats en
cereals", "Carn de boví","Carn de porcí", "Carn d'Oví i Caprí", "Carn d'au",
"Altres Carns", "Despulles comestibles", "Carn seca, salada i fumada", "Altres
preparats de carn", "Peix fresc o refrigerat", "Peix congelat", "Marisc fresc o
refrigerat" , "Marisc congelat", "Peix i marisc sec, fumat i salat", "Altres
preparats de peix i marisc conservats o processats", "Llet fresca sencera",
"Llet fresca desnatada", "Llet en conserva", "Iogurt", "Formatge i quallada",
"Altres productes lactis",  "Ous", "Mantega", "Margarina i altres greixos
vegetals", "Oli d'oliva" , "Altres olis comestibles", "Altres greixos animals
comestibles", "Fruites fresques o refrigerades", "Fruites congelades", "Fruits
secs i fruits de closca", "Fruites en conserva i productes a base de fruites",
"Llegums i hortalisses fresques o refrigerades, excepte patates i altres
tubercles", "Llegums i hortalisses congelades, excepte patates i altres
tubercles", "Llegums i hortalisses seques o conservades d'una altra manera o
processades", "Patates", "Patates xips", "Altres tubercles i els seus
productes", "Sucre", "Confitures, melmelades i mel", "Xocolata", "Productes de
confiteria", "Gelats", "Succedanis artificials del sucre", "Salses i
condiments", "Sal, espècies i herbes culinàries", "Aliments per a nadons", "Plats
preparats", "Altres productes alimentaris n. c. o. p.", "Café", "Te", "Cacau i
xocolata en pols", "Aigua mineral o de font", "Refrescs", "Sucs de fruites i
vegetals"))

# install.packages("dplyr")
library(dplyr)
# ----------------------------------------------------------
## BASE DE DADES GRUP 1. ALMENTS I BEGUDES NO ALCOHOLIQUES
```

```
# ------------------------------------------------------------
bd_aliments <- left_join(bd, nom_grup, by = "CODIGO")
bd_aliments <- left_join(bd_aliments, nom_espe, by = "CODIGO")
```

A continuació crearem una nova base de dades la qual només tindrà dues variables, la variable NUMERO i la variable ALIMENTACIO. Aquesta ultima ens indicarà el tipus d'alimentació de la llar.

```
# ------------------------------------------------------------
## BASE DE DADES OMNÍVORS I VEGETARIANS
# ------------------------------------------------------------

###### 1. OMNÍVORS

Omnivors = bd_aliments %>% filter(GRUP=="Carn")

Llars_omnivores = Omnivors %>% select(ANOENC,NUMERO) %>% group_by(ANOENC,NUMERO)
%>% summarise(N=n())
Llars_omnivores = Llars_omnivores %>% select(ANOENC, NUMERO)

bd_Omnivors = left_join(Llars_omnivores, bd_aliments)

###### 2. VEGETARIANS

Vegetarians = anti_join(bd_aliments, Llars_omnivores, by = c("ANOENC",
"NUMERO"))
Llars_vegetarianes = Vegetarians %>% select(ANOENC,NUMERO) %>%
group_by(ANOENC,NUMERO) %>% summarise(N=n())
Llars_vegetarianes = Llars_vegetarianes %>% select(ANOENC, NUMERO)

bd_Vegetarians = left_join(Llars_vegetarianes, bd_aliments)

# Exportem les dues bases de dades
# write.table(bd_Omnivors, file = "bd_Omnivors.csv", sep = ";", na = "NA", dec
= ".", row.names = FALSE, col.names = TRUE)
# write.table(bd_Vegetarians, file = "bd_Vegetarians.csv", sep = ";", na = "NA",
dec = ".", row.names = FALSE, col.names = TRUE)
```

Un cop tenim la base de dades tant de vegetarians com omnívors el que farem serà crear la variable resposta:

```
bd_Omnivors$Y = 1
bd_Vegetarians$Y = 0
```

Conèixer la base de dades

```
# View(bd_aliments) ### Veure la base de dades
class(bd) ### Classe
dim(bd) ### Dimensio
nrow(bd) ### Nombre de files
ncol(bd) ### Nombre de columnes
colnames(bd) ### Nom de les columnes
# str(bd_aliments) ### Coneixer l'estructura de la base de dades
```

NORMALITZACIÓ BASE DE DADES

La normalització és el procés per el qual s'organitza la informació per evitar la redundància de dades i així obtenir una base de dades optimitzada. Alguns dels objectius que volem aconseguir amb la normalització de la base de dades són:

- Evitar dades repetides.
- Simplificar dependències entre columnes.
- Administrar la mida de la base de dades.
- Proporcionar flexibilitat d'accés a la informació.
- Mantenir la integritat de les dades.

- Assegurar el bon desenvolupament futur.

Els passos a seguir en el procés de normalització venen donats per les denominades cinc Formes Normals:

i. Primera forma normal (1FN)

A 1FN es necessari que tots els atributs (variables) de la relació (taula) siguin atòmics, es a dir, que no hi hagin atributs (variables) compostos ni multivalorats. Per tant hem d'identificar si existeixen grups de repetició que es donin sobre el mateix registre.

La nostra clau primària és c(ANOENC, NUMERO) i com es pot observar, aquest es repeteix.

- **No 1FN**: R(**ANOENC, NUMERO**, GASTO, GASTOT, SEXO, EDAD, NACION, DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARESI, NMIEMBR1, NMIEMBR2, NMIEMBR3,ESTUDIOS1, ESTUDIOS2, ESTUDIOS3,GRUP, SUBGRUP, Y)
- **Sí 1FN:** R1(**ANOENC, NUMERO**, GASTOT, SEXO, EDAD, NACION, DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARESI, NMIEMBR1, NMIEMBR2, NMIEMBR3,ESTUDIOS1, ESTUDIOS2, ESTUDIOS3,Y) , R2(**ANOENC, NUMERO**, GASTO, CODIGO, GRUP, SUBGRUP)

```
R = bd
R_1 = R %>% select(ANOENC, NUMERO, GASTOT, SEXO, EDAD, NACION, DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARES, NMIEMBR1, NMIEMBR2, NMIEMBR3,ESTUDIOS1, ESTUDIOS2, ESTUDIOS3, Y) %>% group_by(ANOENC, NUMERO) %>% filter(row_number(NUMERO) == 1)
R_2 = R %>% select(ANOENC, NUMERO, GASTO, CODIGO, GRUP, SUBGRUP) %>% group_by(ANOENC, NUMERO)

head(R_1)
head(R_2)
```

ii. Segona forma normal (2FN)

La 2FN requereix que la relació R estigui en 1FN i a més tots els atributs tipus no principal tinguin una dependència funcional completa amb algun atribut tipus clau de R. Es a dir perquè la nostra base de dades es trobi en la segona forma normal, ha d'estar en 1FN i identificar les dependències transitives i funcionals. La **dependència funcional** per exemple si tenim A,B,C,D com atributs, on A es la clau candidata/primària, els atributs B,C,D,A dependent funcional i totalment de la clau candidata A, es adir que si A no existeix, la B no tindrà raó de ser i així amb C i D. La dependència transitiva, si agafem el mateix exemple, però B depèn funcionalment de A i C depèn funcionalment de B i D depèn funcionalment de C. Llavors de D a A hi ha una dependència transitiva.

En el nostre cas, existeix una dependència transitiva que es CODIGO -> SUBGRUP -> GRUP i la clau primària c(ANOENC, NUMERO) no té una dependència funcional amb les variables (CODIGO, SUBGRUP, GRUP). Per tant, en aquest cas no es compleix la 2FN i s'hauran de dividir la base de dades.

- **No 2FN**: R1(**ANOENC, NUMERO**, GASTOT, SEXO, NACION, DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARESI, NMIEMBR1, NMIEMBR2, NMIEMBR3,ESTUDIOS1, ESTUDIOS2, ESTUDIOS3,Y) , R2(**ANOENC, NUMERO**, CODIGO, GRUP, SUBGRUP)
- **Sí 2FN**: R1(**ANOENC, NUMERO**, GASTOT, SEXO, NACION, DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARESI, NMIEMBR1, NMIEMBR2, NMIEMBR3,ESTUDIOS1, ESTUDIOS2, ESTUDIOS3,Y), R21(**ANOENC, NUMERO**, CODIGO), R22(**CODIGO**, GRUP, SUBGRUP)

```
R_21 = R %>% select(ANOENC, NUMERO, CODIGO)
R_22 = R %>% select(CODIGO, SUBGRUP, GRUP) %>% group_by(CODIGO) %>%
filter(row_number(CODIGO) == 1)
R_22 = R_22 %>% select(CODIGO, SUBGRUP, GRUP)

head(R_21)
head(R_22)
```

iii. Tercera forma normal (3FN)

La 3FN ens diu que ha d'estar en 2FN i tot atribut no principal i ha de dependre total i funcionalment de la clau principal i eliminar les dependències transitives.

En el nostre cas no tenim dependències funcionals i hem d'eliminar la dependència transitiva de CODIGO -> GRUP -> SUBGRUP. Per tant, hem de seguir dividint les taules.

- **No 3FN**: R1(**ANOENC, NUMERO**, GASTOT, SEXO, EDAD,NACION, TAMAMU, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARESI, NMIEMBR1, NMIEMBR2, NMIEMBR3,ESTUDIOS1, ESTUDIOS2, ESTUDIOS3,Y), R21(**ANOENC, NUMERO**, CODIGO), R22(**CODIGO**, GRUP, SUBGRUP)

- **Sí 3FN:** R1(**ANOENC, NUMERO**, GASTOT, SEXO, EDAD, NACION, TAMAMU, CCAA, COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARESI, NMIEMBR1, NMIEMBR2, NMIEMBR3,ESTUDIOS1, ESTUDIOS2, ESTUDIOS3,Y), R21(**ANOENC, NUMERO**, CODIGO), R221(**CODIGO**, SUBGRUP), R222(**GRUP**, SUBGRUP)

```
R_221 = R %>% select(CODIGO, SUBGRUP) %>% group_by(CODIGO) %>%
filter(row_number(CODIGO) == 1)
R_221 = R_221 %>% select(CODIGO, SUBGRUP)
R_222 = R %>% select(GRUP, SUBGRUP) %>% group_by(SUBGRUP) %>%
filter(row_number(GRUP) == 1)
R_222 = R_222 %>% select(GRUP, SUBGRUP)

head(R_221)
head(R_222)
```

iv. Forma normal de Boyce-Codd (BCNF o 3.5FN)

La BCNF requereix que la relació estigui en 3 FN i que totes les dependències funcionals X -> Y no trivials siguin superclau, es a dir, X es una clau candidata o un subconjunt de la mateixa. En el cas que no estigui en BCNF haurem d'eliminar de cada relació els atributs que no són clau. En el nostre cas ja esta en BCNF. Per tant, una relació esta en forma FNBC si i només si les úniques dependències funcionals elementals son aquelles on la clau primària determinen un atribut. En el nostre cas, les nostres taules ja es troben en la Forma normal de Boyce-Codd.

v. Quarta forma normal (4FN)

La 4FN ens diu que la relació R ha d'estar en BCNF i que cada una de les dependències multievaludes X ->-> Y, X és superClau. Per tant, X es una clau candidata o un subconjunt de la mateixa. En nostre cas particular, ja es troba en 4FN. Es a dir, que la taula ha d'estar en forma FNBC i s'han d'eliminar totes les dependències multievaluades. Una dependència multievaluades es un atribut que per una mateixa entitat pot prendre varis valors diferents.

En el nostre cas no es donen casos de dependències multievaluades. Per tant, les nostres taules ja es troben en la quarta forma normal.

Finalment, exportem les relacions a arxius externs en format csv:

```
write.table(R_1, file = "R_1.csv", sep = ";", na = "NA", dec = ".", row.names =
FALSE, col.names = TRUE)
write.table(R_21, file = "R_2.csv", sep = ";", na = "NA", dec = ".", row.names
= FALSE, col.names = TRUE)
write.table(R_221, file = "R_3.csv", sep = ";", na = "NA", dec = ".", row.names
= FALSE, col.names = TRUE)
write.table(R_222, file = "R_4.csv", sep = ";", na = "NA", dec = ".", row.names
= FALSE, col.names = TRUE)
```

A.2) Anàlisi exploratori de les dades

```
# llibreries
# install.packages("caret", dependencies = c("Depends", "Suggests"))
library(caret)
library(tidyverse)
# Importem la base de dades: dades_11
datos <- read.csv("R_1.csv", sep = ";")
datos$NUMERO <- NULL
datos$LlarID <- 1:nrow(datos)

# Assignem la classe corresponent a cada variable (numèrica o categòrica)

   # Covariables
     datos$GASTOT <- as.numeric(datos$GASTOT)
     datos$SEXO <- as.numeric(datos$SEXO)
     datos$EDAD <- as.numeric(datos$EDAD)
     datos$NACION <- as.numeric(datos$NACION)
     datos$COMITOT  <- as.numeric(datos$COMITOT)
     datos$IMPEXAC <- as.numeric(datos$IMPEXAC)
     datos$NMIEMBR1 <- as.numeric(datos$NMIEMBR1)
     datos$NMIEMBR2 <- as.numeric(datos$NMIEMBR2)
     datos$NMIEMBR3 <- as.numeric(datos$NMIEMBR3)
     datos$ESTUDIOS1 <- as.numeric(datos$ESTUDIOS1)
     datos$ESTUDIOS2 <- as.numeric(datos$ESTUDIOS2)
     datos$ESTUDIOS3 <- as.numeric(datos$ESTUDIOS3)

   # Factors
     datos$DENSIDAD <- as.factor(datos$DENSIDAD)
```

```
datos$CCAA <- as.factor(datos$CCAA)
datos$NUMOCUP <- as.factor(datos$NUMOCUP)
datos$ECIVILSP <- as.factor(datos$ECIVILSP)
datos$ZONARES <- as.factor(datos$ZONARES)
```

Distribució de la Variable Resposta (Y)

```
ggplot(data = datos, aes(x = Y, y = ..count.., fill = Y)) +
  geom_bar(alpha = 0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "Variable Resposta", subtitle = "Consum de Carn") +
  theme_bw() +
  theme(legend.position = "none") + ylab("Freqüència")

# Taula de Frequencies
table(datos$Y)
# Taula de Frequencies (proporcions)
prop.table(table(datos$Y)) %>% round(digits = 2)

# Percentatge d'encerts si es prediu per totes  les llars no mengen carn.
n_observaciones <- nrow(datos)
predicciones <- rep(x = "No",  n_observaciones)
mean(predicciones == datos$Y) * 100
```

Distribució de la Variables Contínues (X)

Variable GASTOT

```
library(ggpubr) # install.packages("ggpubr")
p1 <- ggplot(data = datos, aes(x = GASTOT, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p2 <- ggplot(data = datos, aes(x = Y, y = GASTOT, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot1 <- ggarrange(p1, p2, legend = "top")
final_plot1 <- annotate_figure(final_plot1, top = text_grob("GASTOT", size =
15))
final_plot1

library(ggplot2)
library(gridExtra) # install.packages("gridExtra")
p1 <- ggplot(data = datos, aes(x = log(GASTOT), fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p2 <- ggplot(data = datos, aes(x = Y, y = log(GASTOT), color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot1 <- ggarrange(p1, p2, legend = "top")
final_plot1 <- annotate_figure(final_plot1, top = text_grob("Log(GASTOT)", size
=15))
final_plot1
```

Variable SEXO

```
p3 <- ggplot(data = datos, aes(x = SEXO, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p4 <- ggplot(data = datos, aes(x = Y, y = SEXO, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot2 <- ggarrange(p3, p4, legend = "top")
final_plot2 <- annotate_figure(final_plot2, top = text_grob("SEXO", size = 15))
final_plot2
```

Variable NACION

```
p5 <- ggplot(data = datos, aes(x = NACION, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p6 <- ggplot(data = datos, aes(x = Y, y = NACION, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot3 <- ggarrange(p5, p6, legend = "top")
final_plot3 <- annotate_figure(final_plot3, top = text_grob("NACION", size =
15))
final_plot3
```

Variable EDAD

```
p7 <- ggplot(data = datos, aes(x = EDAD, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p8 <- ggplot(data = datos, aes(x = Y, y = EDAD, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot4 <- ggarrange(p7, p8, legend = "top")
final_plot4 <- annotate_figure(final_plot4, top = text_grob("EDAD", size = 15))
final_plot4
datos <- datos %>%
         mutate(EDAD_grupo = case_when(EDAD <= 30  ~ "[0,30]",
                                       EDAD > 30 & EDAD <= 64  ~ "(30,64]",
                                       EDAD > 64 ~ "(64, 100)"))
datos$EDAD_grupo <- as.factor(datos$EDAD_grupo)
```

Variable COMITOT

```
p9 <- ggplot(data = datos, aes(x = COMITOT, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
```

```
p10 <- ggplot(data = datos, aes(x = Y, y = COMITOT, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot5 <- ggarrange(p9, p10, legend = "top")
final_plot5 <- annotate_figure(final_plot5, top = text_grob("COMITOT", size =
15))
final_plot5

p9 <- ggplot(data = datos, aes(x = log(COMITOT), fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p10 <- ggplot(data = datos, aes(x = Y, y = log(COMITOT), color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot5 <- ggarrange(p9, p10, legend = "top")
final_plot5 <- annotate_figure(final_plot5, top = text_grob("log(COMITOT)", size
= 15))
final_plot5
```

Variable IMPEXAC

```
p10 <- ggplot(data = datos, aes(x = IMPEXAC, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p11 <- ggplot(data = datos, aes(x = Y, y = IMPEXAC, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot6 <- ggarrange(p10, p11, legend = "top")
final_plot6 <- annotate_figure(final_plot6, top = text_grob("IMPEXAC", size =
15))
final_plot6
```

```
datos <- datos %>%
      mutate(IMPEXAC_grupo = case_when(IMPEXAC <= 1100  ~ "[0, 1100]",
                                       IMPEXAC > 1100 & IMPEXAC <= 1800  ~ "(1100, 1800]",
                                       IMPEXAC > 1800 ~ "(1800, Inf)"))
datos$IMPEXAC_grupo <- as.factor(datos$IMPEXAC_grupo)

datos <- datos %>%
         mutate(IMPEXAC_grupo = case_when(IMPEXAC <= 1100  ~ "[0, 1100]",
                                          IMPEXAC > 1100 & IMPEXAC <= 1800  ~
"(1100, 1800]",
                                          IMPEXAC > 1800 ~ "(1800, Inf)"))
datos$IMPEXAC_grupo <- as.factor(datos$IMPEXAC_grupo)
```

Variable NMIEMBR1

```
p12 <- ggplot(data = datos, aes(x = NMIEMBR1, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
```

```
      theme_bw()
p13 <- ggplot(data = datos, aes(x = Y, y = NMIEMBR1, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot7 <- ggarrange(p12, p13, legend = "top")
final_plot7 <- annotate_figure(final_plot7, top = text_grob("NMIEMBR1", size =
15))
final_plot7
```

Variable NMIEMBR2

```
p14 <- ggplot(data = datos, aes(x = NMIEMBR2, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p15 <- ggplot(data = datos, aes(x = Y, y = NMIEMBR2, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot8 <- ggarrange(p14, p15, legend = "top")
final_plot8 <- annotate_figure(final_plot8, top = text_grob("NMIEMBR2", size =
15))
final_plot8
datos <- datos %>%
         mutate(NMIEMBR2_grupo = case_when(NMIEMBR2 <= 25  ~ " [0%,25%]",
                                           NMIEMBR2 > 25 & NMIEMBR2 <= 75  ~ "(25%,
75%]",
                                            NMIEMBR2 > 75   ~ "(75%, 100%]"))
datos$NMIEMBR2_grupo <- as.factor(datos$NMIEMBR2_grupo)
```

Variable NMIEMBR3

```
p16 <- ggplot(data = datos, aes(x = NMIEMBR3, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p17 <- ggplot(data = datos, aes(x = Y, y = NMIEMBR3, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot9 <- ggarrange(p16, p17, legend = "top")
final_plot9 <- annotate_figure(final_plot9, top = text_grob("NMIEMBR3", size =
15))
final_plot9
datos <- datos %>%
         mutate(NMIEMBR3_grupo = case_when(NMIEMBR3 <= 25  ~ " [0%,25%]",
                                           NMIEMBR3 > 25 & NMIEMBR3 <= 75  ~ "(25%,
75%]",
                                            NMIEMBR3 > 75  ~ "(75%, 100%]"))
datos$NMIEMBR3_grupo <- as.factor(datos$NMIEMBR3_grupo)
```

Variable ESTUDIOS1

```
p18 <- ggplot(data = datos, aes(x = ESTUDIOS1, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
```

```
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p19 <- ggplot(data = datos, aes(x = Y, y = ESTUDIOS1, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot10 <- ggarrange(p18, p19, legend = "top")
final_plot10 <- annotate_figure(final_plot10, top = text_grob("ESTUDIOS1", size
= 15))
final_plot10
p18 <- ggplot(data = datos, aes(x = log(ESTUDIOS1), fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p19 <- ggplot(data = datos, aes(x = Y, y = log(ESTUDIOS1), color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot10 <- ggarrange(p18, p19, legend = "top")
final_plot10 <- annotate_figure(final_plot10, top = text_grob("Log(ESTUDIOS1)",
size =15))
final_plot10
```

Variables ESTUDIOS2

```
p20 <- ggplot(data = datos, aes(x = ESTUDIOS2, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p21 <- ggplot(data = datos, aes(x = Y, y = ESTUDIOS2, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot11 <- ggarrange(p20, p21, legend = "top")
final_plot11 <- annotate_figure(final_plot11, top = text_grob("ESTUDIOS2", size
= 15))
final_plot11
datos <- datos %>%
         mutate(ESTUDIOS2_grupo = case_when(ESTUDIOS2 <= 20  ~ " [0%,20%]",
                                      ESTUDIOS2 > 20 & ESTUDIOS2 <= 75  ~ "(20%,
75%]",
                                       ESTUDIOS2 > 75 ~ "(75%, 100%]"))
datos$ESTUDIOS2_grupo <- as.factor(datos$ESTUDIOS2_grupo)
```

Variable ESTUDIOS3

```
p22 <- ggplot(data = datos, aes(x = ESTUDIOS3, fill = Y)) +
      geom_density(alpha = 0.5) +
      scale_fill_manual(values = c("gray50", "#7EC3E5")) +
      geom_rug(aes(color = Y), alpha = 0.5) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
p23 <- ggplot(data = datos, aes(x = Y, y = ESTUDIOS3, color = Y)) +
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15) +
      scale_color_manual(values = c("gray50", "#7EC3E5")) +
      theme_bw()
final_plot11 <- ggarrange(p22, p23, legend = "top")
```

```
final_plot11 <- annotate_figure(final_plot11, top = text_grob("ESTUDIOS3", size
= 15))
final_plot11
datos <- datos %>%
         mutate(ESTUDIOS3_grupo = case_when(ESTUDIOS3 <= 20  ~ " [0%,20%]",
                                            ESTUDIOS3 > 20 & ESTUDIOS3 <= 75  ~ "(20%,
75%]",
                                             ESTUDIOS3 > 75  ~ "(75%, 10%]"))
datos$ESTUDIOS3_grupo <- as.factor(datos$ESTUDIOS3_grupo)
```

Distribució de la Variables Qualitatives (X)

Variable DENSIDAD

```
dd <- datos
dd$DENSIDAD <- factor(dd$DENSIDAD, labels = c("Zona densament poblada", "Zona
intermedia",
                                              "Zona diseminada"))
ggplot(data = dd, aes(x = DENSIDAD, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "DENSIDAD") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(NULL)
```

Variable CCAA

```
dd$CCAA <- factor(dd$CCAA, labels = c("PIBpc > 25.000 EUR", "20.000 <= PIB <=
25.000 EUR", "PIBpc < 20.000 EUR" ))
ggplot(data = dd, aes(x = CCAA, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "CCAA") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(NULL)
```

Variable NUMOCUP

```
dd$NUMOCUP <- factor(dd$NUMOCUP, labels = c("Cap ocupat", "1 Ocupat",
                                            "2 Ocupats", "3 Ocupats", "4 o més
Ocuptas"))
ggplot(data = dd, aes(x = NUMOCUP, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "NUMOCUP") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(NULL)
```

Variable ECIVILSP

```
dd$ECIVILSP <- factor(dd$ECIVILSP, labels = c("Solter", "Casat",
                                              "Vidu", "Separat o divorciat"))
ggplot(data = dd, aes(x = ECIVILSP, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "ECIVILSP") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(NULL)
```

Variable ZONARES

```
dd$ZONARES <- factor(dd$ZONARES, labels = c("Urbana", "Rural"))
ggplot(data = dd, aes(x = ZONARES, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
```

```
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "ZONARES") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(NULL)
```

Variable EDAD_grupo

```
dd$EDAD_grupo <- factor(dd$EDAD_grupo, labels = c( "(30,64]", "(64,100)",
"[0,30]"))
ggplot(data = dd, aes(x = EDAD_grupo, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "EDAD_grupo") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(limits = c("[0,30]",
"(30,64]", "(64,100)"), NULL)
```

Variable IMPEXAC_grupo

```
dd$IMPEXAC_grupo <- factor(dd$IMPEXAC_grupo, labels = c( "(1100, 1800]", "(1800,
Inf)", "[0, 1100]"))
ggplot(data = dd, aes(x = IMPEXAC_grupo, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "IMPEXAC_grupo") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(limits = c("[0, 1100]",
"(1100, 1800]", "(1800, Inf)"), NULL)
```

Variable NMIEMBR2_grupo

```
dd$NMIEMBR2_grupo <- factor(dd$NMIEMBR2_grupo, labels = c( "[0%,25%]", "(25%,
75%]", "(75%, 100%]"))
ggplot(data = dd, aes(x = NMIEMBR2_grupo, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "NMIEMBR2_grupo") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(limits = c("[0%,25%]",
"(25%, 75%]", "(75%, 100%]"), NULL)
```

Variable NMIEMBR3_grupo

```
dd$NMIEMBR3_grupo <- factor(dd$NMIEMBR3_grupo, labels = c( "[0%,25%]", "(25%,
75%]", "(75%, 100%]"))
ggplot(data = dd, aes(x = NMIEMBR3_grupo, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "NMIEMBR3_grupo") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(limits = c("[0%,25%]",
"(25%, 75%]", "(75%, 100%]"), NULL)
```

Variable ESTUDIOS2_grupo

```
dd$ESTUDIOS2_grupo <- factor(dd$ESTUDIOS2_grupo, labels = c( "[0%,20%]", "(20%,
75%]", "(75%, 100%]"))
ggplot(data = dd, aes(x = ESTUDIOS2_grupo, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "ESTUDIOS2_grupo") +
  theme_bw() +
```

```
  theme(legend.position = "bottom") + scale_x_discrete(limits = c("[0%,20%]",
"(20%, 75%]", "(75%, 100%]"), NULL)
```

Variable ESTUDIOS3_grupo

```
dd$ESTUDIOS3_grupo <- factor(dd$ESTUDIOS3_grupo, labels = c( "[0%,20%]", "(20%,
75%]", "(75%, 100%]"))
ggplot(data = dd, aes(x = ESTUDIOS3_grupo, y = ..count.., fill = Y)) +
  geom_bar(alpha=0.5) +
  scale_fill_manual(values = c("gray50", "#7EC3E5")) +
  labs(title = "ESTUDIOS3_grupo") +
  theme_bw() +
  theme(legend.position = "bottom") + scale_x_discrete(limits = c("[0%,20%]",
"(20%, 75%]", "(75%, 100%]"), NULL)
```

Importància de les variables

Correlació entre variables contínues

```
library(ggcorrplot)
vcontinuas = datos %>% select(GASTOT, SEXO, EDAD, NACION, COMITOT, IMPEXAC,
NMIEMBR1, NMIEMBR2, NMIEMBR3, ESTUDIOS1, ESTUDIOS2, ESTUDIOS3)
corr <- round(cor(vcontinuas), 1) # coef correlacio
p.mat <- cor_pmat(vcontinuas) # pvalors
ggcorrplot(corr, hc.order = TRUE, type = "lower",
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"), lab = TRUE)

library(ggcorrplot)
vcontinuas = datos %>% select(GASTOT, SEXO, NACION, COMITOT, IMPEXAC, ESTUDIOS1)
corr <- round(cor(vcontinuas), 1) # coef correlacio
p.mat <- cor_pmat(vcontinuas) # pvalors
ggcorrplot(corr, hc.order = TRUE, type = "lower",
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"), lab = TRUE)
```

A.3) Divisió de les dades en entrenament i test

```
set.seed(123)
# Es creen els índex de les observacions d'entrenament
train <- createDataPartition(y = datos$Y, p = 0.8, list = FALSE, times = 1)
datos_train <- datos[train, ]
datos_test  <- datos[-train, ]
prop.table(table(datos_train$Y))
prop.table(table(datos_test$Y))
```

A.4) Preprocessament de les dades

```
Exclusió de variables amb variància propera a zero

objeto_recipe <- recipe(formula = Y ~ GASTOT + SEXO + EDAD_grupo + NACION +
DENSIDAD + CCAA + COMITOT + NUMOCUP + IMPEXAC_grupo + ECIVILSP + ZONARES +
ESTUDIOS1 + ESTUDIOS2_grupo + ESTUDIOS3_grupo,
                        data =  datos_train)
objeto_recipe

datos %>% select(GASTOT, SEXO, EDAD_grupo, NACION, DENSIDAD, CCAA, COMITOT,
NUMOCUP,  IMPEXAC_grupo,  ECIVILSP,  ZONARES,  ESTUDIOS1,  ESTUDIOS2_grupo,
ESTUDIOS3_grupo) %>%
  nearZeroVar(saveMetrics = TRUE)
```

A.5) Models. Ajust, optimització i validació del model

```
# DEFINICIÓ ENTRENAMENT
control_train_original = trainControl(method = "repeatedcv",
                                      repeats = 5,
                                      classProbs = TRUE,
                                      summaryFunction = twoClassSummary)
control_train_down = trainControl(method = "repeatedcv",
                                      repeats = 5,
                                      classProbs = TRUE,
                                      summaryFunction = twoClassSummary,
                                      sampling = "down")
control_train_up = trainControl(method = "repeatedcv",
                                      repeats = 5,
                                      classProbs = TRUE,
                                      summaryFunction = twoClassSummary,
                                      sampling = "up")
```

Regressió logística

```
# AJUST DEL MODEL
set.seed(342)
modelo_logistic_original <- train(Y ~ ., data = datos_train_prep,
                          method = "glm",
                          preProcess = c("center", "scale"),
                          metric = "ROC",
                          trControl = control_train_original,
                          family = "binomial")

set.seed(342)
modelo_logistic_down <- train(Y ~ ., data = datos_train_prep,
                          method = "glm",
                          preProcess = c("center", "scale"),
                          metric = "ROC",
                          trControl = control_train_down,
                          family = "binomial")

set.seed(342)
modelo_logistic_up <- train(Y ~ ., data = datos_train_prep,
                          method = "glm",
                          preProcess = c("center", "scale"),
                          metric = "ROC",
                          trControl = control_train_up,
                          family = "binomial")

# REPRESENTACIÓ GRÀFICA
library(pROC) # install.packages("pROC")
predicciones_logistic_original <- predict(object = modelo_logistic_original,
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_logistic_down <- predict(object = modelo_logistic_down,
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_logistic_up <- predict(object = modelo_logistic_up,
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_logistic_original_resp         <-         predict(object         =
modelo_logistic_original,
                        newdata = datos_test_prep)
predicciones_logistic_down_resp <- predict(object = modelo_logistic_down,
                        newdata = datos_test_prep)
predicciones_logistic_up_resp <- predict(object = modelo_logistic_up,
                        newdata = datos_test_prep)
curva_roc_logistic_original <- roc(response = datos_test_prep$Y,
                 predictor = predicciones_logistic_original$Si)
curva_roc_logistic_down <- roc(response = datos_test_prep$Y,
                 predictor = predicciones_logistic_down$Si)
```

```
curva_roc_logistic_up <- roc(response = datos_test_prep$Y,
                 predictor = predicciones_logistic_up$Si)
roc_logistic_original          <-          data.frame(x          =          (1-
curva_roc_logistic_original$specificities),
y=curva_roc_logistic_original$sensitivities,  Model  =  "Regresió  Logística
Original")
roc_logistic_down <- data.frame(x = (1-curva_roc_logistic_down$specificities),
y=curva_roc_logistic_down$sensitivities, Model = "Regresió Logística Down")
roc_logistic_up  <-  data.frame(x  =  (1-curva_roc_logistic_up$specificities),
y=curva_roc_logistic_up$sensitivities, Model = "Regresió Logística Up")

droc <- rbind(roc_logistic_original,roc_logistic_down,roc_logistic_up)
library(paletteer)
ggplot(droc, aes(x = x, y = y, color = Model)) +
  geom_line() + labs(x= "1-Specificity", y="Sensitivity", title = "Corbes ROC.
Regressió Logística")+
  scale_fill_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  scale_colour_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  theme_bw()

# MATRIU DE CONFUSIÓ
library("vcd") # install.packages("grid")
cm_logistic_original  <-  confusionMatrix(predicciones_logistic_original_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_logistic_down      <-      confusionMatrix(predicciones_logistic_down_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_logistic_up        <-        confusionMatrix(predicciones_logistic_up_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))

a  <-  fourfoldplot(round(cm_logistic_original$table,2),  color  =  c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Original")
b   <-   fourfoldplot(round(cm_logistic_down$table,2),   color   =   c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Submostreig (Down)")
c   <-   fourfoldplot(round(cm_logistic_up$table,2),   color   =   c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Sobremostreig (Up)")
```

Arbre de decisió

```
# AJUST DEL MODEL
set.seed(342)
modelo_C50Tree_original <- train(Y ~ ., data = datos_train_prep,
                    method = "C5.0Tree",
                    preProcess = c("center", "scale"),
                    metric = "ROC",
                    trControl = control_train_original)

set.seed(342)
modelo_C50Tree_down <- train(Y ~ ., data = datos_train_prep,
                    method = "C5.0Tree",
                    preProcess = c("center", "scale"),
                    metric = "ROC",
                    trControl = control_train_down)

set.seed(342)
modelo_C50Tree_up <- train(Y ~ ., data = datos_train_prep,
                    method = "C5.0Tree",
                    preProcess = c("center", "scale"),
                    metric = "ROC",
                    trControl = control_train_up)

# REPRESENTACIÓ GRÀFICA
library(pROC)
# Se obtienen las probabilidades predichas para cada clase
predicciones_C50Tree_original <- predict(object = modelo_C50Tree_original,
```

```
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_C50Tree_down <- predict(object = modelo_C50Tree_down,
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_C50Tree_up <- predict(object = modelo_C50Tree_up,
                        newdata = datos_test_prep,
                        type = "prob")

predicciones_C50Tree_original_resp <- predict(object = modelo_C50Tree_original,
                        newdata = datos_test_prep)
predicciones_C50Tree_down_resp <- predict(object = modelo_C50Tree_down,
                        newdata = datos_test_prep)
predicciones_C50Tree_up_resp <- predict(object = modelo_C50Tree_up,
                        newdata = datos_test_prep)

curva_roc_C50Tree_original <- roc(response = datos_test_prep$Y,
                 predictor = predicciones_C50Tree_original$Si)
curva_roc_C50Tree_down <- roc(response = datos_test_prep$Y,
                 predictor = predicciones_C50Tree_down$Si)
curva_roc_C50Tree_up <- roc(response = datos_test_prep$Y,
                 predictor = predicciones_C50Tree_up$Si)
roc_C50Tree_original          <-          data.frame(x          =          (1-
curva_roc_C50Tree_original$specificities),
y=curva_roc_C50Tree_original$sensitivities, Model = "C50Tree Original")
roc_C50Tree_down  <-  data.frame(x  =  (1-curva_roc_C50Tree_down$specificities),
y=curva_roc_C50Tree_down$sensitivities, Model = "C50Tree Down")
roc_C50Tree_up   <-   data.frame(x   =   (1-curva_roc_C50Tree_up$specificities),
y=curva_roc_C50Tree_up$sensitivities, Model = "C50Tree Up")

droc <- rbind(roc_C50Tree_original,roc_C50Tree_down,roc_C50Tree_up)
library(paletteer)
ggplot(droc, aes(x = x, y = y, color = Model)) +
  geom_line() + labs(x= "1-Specificity", y="Sensitivity", title = "Corbes ROC.
Àrbre de Decisió Simple (C.5.0)")+
  scale_fill_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  scale_colour_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  theme_bw()

# MATRIU DE CONFUSIÓ
library("vcd") # install.packages("grid")
cm_C50Tree_original    <-    confusionMatrix(predicciones_C50Tree_original_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_C50Tree_down        <-        confusionMatrix(predicciones_C50Tree_down_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_C50Tree_up          <-          confusionMatrix(predicciones_C50Tree_up_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))

a   <-   fourfoldplot(round(cm_C50Tree_original$table,2),   color   =   c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Original")
b   <-   fourfoldplot(round(cm_C50Tree_down$table,2),   color   =   c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Submostreig (Down)")
c   <-   fourfoldplot(round(cm_C50Tree_up$table,2),   color   =   c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Sobremostreig (Up)")
```

Bagging

```
# AJUST DEL MODEL
set.seed(342)
modelo_bag_original <- train(Y ~ ., data = datos_train_prep,
                   method = "treebag",
                   preProcess = c("center", "scale"),
                   nbagg = 25,
                   metric = "ROC",
```

```
                    trControl = control_train_original)

set.seed(342)
modelo_bag_down <- train(Y ~ ., data = datos_train_prep,
                    method = "treebag",
                    preProcess = c("center", "scale"),
                    nbagg = 25,
                    metric = "ROC",
                    trControl = control_train_down)

set.seed(342)
modelo_bag_up <- train(Y ~ ., data = datos_train_prep,
                    method = "treebag",
                    preProcess = c("center", "scale"),
                    nbagg = 25,
                    metric = "ROC",
                    trControl = control_train_up)

# REPRESENTACIÓ GRÀFICA
library(pROC)
predicciones_bag_original <- predict(object = modelo_bag_original,
                         newdata = datos_test_prep,
                         type = "prob")
predicciones_bag_down <- predict(object = modelo_bag_down,
                         newdata = datos_test_prep,
                         type = "prob")
predicciones_bag_up <- predict(object = modelo_bag_up,
                         newdata = datos_test_prep,
                         type = "prob")

predicciones_bag_original_resp <- predict(object = modelo_bag_original,
                         newdata = datos_test_prep)
predicciones_bag_down_resp <- predict(object = modelo_bag_down,
                         newdata = datos_test_prep)
predicciones_bag_up_resp <- predict(object = modelo_bag_up,
                         newdata = datos_test_prep)

curva_roc_bag_original <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_bag_original$Si)
curva_roc_bag_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_bag_down$Si)
curva_roc_bag_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_bag_up$Si)

roc_bag_original  <-  data.frame(x  =  (1-curva_roc_bag_original$specificities),
y=curva_roc_bag_original$sensitivities, Model = "Bagging Original")
roc_bag_down    <-    data.frame(x    =    (1-curva_roc_bag_down$specificities),
y=curva_roc_bag_down$sensitivities, Model = "Bagging Down")
roc_bag_up     <-     data.frame(x     =     (1-curva_roc_bag_up$specificities),
y=curva_roc_bag_up$sensitivities, Model = "Bagging Up")

droc <- rbind(roc_bag_original,roc_bag_down,roc_bag_up)

library(paletteer)
ggplot(droc, aes(x = x, y = y, color = Model)) +
  geom_line() + labs(x= "1-Specificity", y="Sensitivity", title = "Corbes ROC.
Bagging")+
  scale_fill_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  scale_colour_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  theme_bw()

# MATRIU DE CONFUSIÓ
library("vcd") # install.packages("grid")
cm_bag_original       <-       confusionMatrix(predicciones_bag_original_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_bag_down  <-  confusionMatrix(predicciones_bag_down_resp,  datos_test_prep$Y,
dnn = c("Prediccions", "Observacions"))
```

```
cm_bag_up <- confusionMatrix(predicciones_bag_up_resp, datos_test_prep$Y, dnn =
c("Prediccions", "Observacions"))

a   <-   fourfoldplot(round(cm_bag_original$table,2),   color   =   c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Original")
b    <-    fourfoldplot(round(cm_bag_down$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Submostreig (Down)")
c    <-    fourfoldplot(round(cm_bag_up$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Sobremostreig (Up)")
```

Random Forest

```
# AJUST DEL MODEL
set.seed(342)
modelo_rf_original <- train(Y ~ ., data = datos_train_prep,
                    method = "rf",
                    preProcess = c("center", "scale"),
                    ntree = 50,
                    metric = "ROC",
                    trControl = control_train_original)

set.seed(342)
modelo_rf_down <- train(Y ~ ., data = datos_train_prep,
                    method = "rf",
                    preProcess = c("center", "scale"),
                    ntree = 50,
                    metric = "ROC",
                    trControl = control_train_down)

set.seed(342)
modelo_rf_up <- train(Y ~ ., data = datos_train_prep,
                    method = "rf",
                    preProcess = c("center", "scale"),
                    ntree = 50,
                    metric = "ROC",
                    trControl = control_train_up)

# REPRESENTACIÓ GRÀFICA
library(pROC)
predicciones_rf_original <- predict(object = modelo_rf_original,
                         newdata = datos_test_prep,
                         type = "prob")
predicciones_rf_down <- predict(object = modelo_rf_down,
                         newdata = datos_test_prep,
                         type = "prob")
predicciones_rf_up <- predict(object = modelo_rf_up,
                         newdata = datos_test_prep,
                         type = "prob")
predicciones_rf_original_resp <- predict(object = modelo_rf_original,
                         newdata = datos_test_prep)
predicciones_rf_down_resp <- predict(object = modelo_rf_down,
                         newdata = datos_test_prep)
predicciones_rf_up_resp <- predict(object = modelo_rf_up,
                         newdata = datos_test_prep)
curva_roc_rf_original <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_original$Si)
curva_roc_rf_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_down$Si)
curva_roc_rf_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_up$Si)
roc_rf_original  <-  data.frame(x  =  (1-curva_roc_rf_original$specificities),
y=curva_roc_rf_original$sensitivities, Model = "Random Forest Original")
roc_rf_down    <-    data.frame(x    =    (1-curva_roc_rf_down$specificities),
y=curva_roc_rf_down$sensitivities, Model = "Random Forest Down")
```

```
roc_rf_up     <-     data.frame(x     =     (1-curva_roc_rf_up$specificities),
y=curva_roc_rf_up$sensitivities, Model = "Random Forest Up")

droc <- rbind(roc_rf_original,roc_rf_down,roc_rf_up)

library(paletteer)
ggplot(droc, aes(x = x, y = y, color = Model)) +
  geom_line() + labs(x= "1-Specificity", y="Sensitivity", title = "Corbes ROC.
Random Forest")+
  scale_fill_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  scale_colour_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  theme_bw()

# REPRESENTACIÓ GRÀFICA
ggplot(modelo_rf_original, highlight = TRUE) +
  scale_x_continuous(breaks = 1:30) +
  labs(title = "Evolució ROC del model Random Forest Original")  +
  theme_bw() + geom_line(color = "#656A59")

ggplot(modelo_rf_down, highlight = TRUE) +
  scale_x_continuous(breaks = 1:30) +
  labs(title = "Evolució ROC del model Random Forest Submostreig (Down)")  +
  theme_bw() + geom_line(color = "#F8B323")

ggplot(modelo_rf_up, highlight = TRUE) +
  scale_x_continuous(breaks = 1:30) +
  labs(title = "Evolució ROC del model Random Forest Sobremostreig (Up)")  +
  theme_bw() + geom_line(color = "#46B2B5")

# MATRIU DE CONFUSIÓ
library("vcd") # install.packages("grid")
cm_rf_original        <-        confusionMatrix(predicciones_rf_original_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_rf_down <- confusionMatrix(predicciones_rf_down_resp, datos_test_prep$Y, dnn
= c("Prediccions", "Observacions"))
cm_rf_up <- confusionMatrix(predicciones_rf_up_resp, datos_test_prep$Y, dnn =
c("Prediccions", "Observacions"))

a    <-    fourfoldplot(round(cm_rf_original$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Original")
b    <-    fourfoldplot(round(cm_rf_down$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Submostreig (Down)")
c    <-    fourfoldplot(round(cm_rf_up$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Sobremostreig (Up)")
```

Gradient Boosting

```
# AJUST DEL MODEL
set.seed(342)
modelo_boost_original <- train(Y ~ ., data = datos_train_prep,
                   method = "gbm",
                   preProcess = c("center", "scale"),
                   metric = "ROC",
                   tuneLength = 5,
                   trControl = control_train_original)
set.seed(342)
modelo_boost_down <- train(Y ~ ., data = datos_train_prep,
                   method = "gbm",
                   preProcess = c("center", "scale"),
                   metric = "ROC",
                   tuneLength = 5,
                   trControl = control_train_down)
set.seed(342)
modelo_boost_up <- train(Y ~ ., data = datos_train_prep,
```

```
                    method = "gbm",
                    preProcess = c("center", "scale"),
                    metric = "ROC",
                    tuneLength = 5,
                    trControl = control_train_up)

# REPRESENTACIÓ GRÀFICA
library(pROC)
predicciones_boost_original <- predict(object = modelo_boost_original,
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_boost_down <- predict(object = modelo_boost_down,
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_boost_up <- predict(object = modelo_boost_up,
                        newdata = datos_test_prep,
                        type = "prob")

predicciones_boost_original_resp <- predict(object = modelo_boost_original,
                        newdata = datos_test_prep)
predicciones_boost_down_resp <- predict(object = modelo_boost_down,
                        newdata = datos_test_prep)
predicciones_boost_up_resp <- predict(object = modelo_boost_up,
                        newdata = datos_test_prep)

curva_roc_boost_original <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_original$Si)
curva_roc_boost_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_down$Si)
curva_roc_boost_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_up$Si)
roc_boost_original           <-           data.frame(x           =           (1-
curva_roc_boost_original$specificities),
y=curva_roc_boost_original$sensitivities, Model = "Gradient Boosting Original")
roc_boost_down   <-   data.frame(x   =   (1-curva_roc_boost_down$specificities),
y=curva_roc_boost_down$sensitivities, Model = "Gradient Boosting Down")
roc_boost_up    <-    data.frame(x    =    (1-curva_roc_boost_up$specificities),
y=curva_roc_boost_up$sensitivities, Model = "Gradient Boosting Up")

droc <- rbind(roc_boost_original,roc_boost_down,roc_boost_up)

library(paletteer)
ggplot(droc, aes(x = x, y = y, color = Model)) +
  geom_line() + labs(x= "1-Specificity", y="Sensitivity", title = "Corbes ROC.
Gradient Boosting") +
  scale_fill_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  scale_colour_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  theme_bw()

# MATRIU DE CONFUSIÓ
library("vcd") # install.packages("grid")
cm_boost_original      <-      confusionMatrix(predicciones_boost_original_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_boost_down          <-          confusionMatrix(predicciones_boost_down_resp,
datos_test_prep$Y, dnn = c("Prediccions", "Observacions"))
cm_boost_up  <-  confusionMatrix(predicciones_boost_up_resp,  datos_test_prep$Y,
dnn = c("Prediccions", "Observacions"))

a   <-   fourfoldplot(round(cm_boost_original$table,2),   color   =   c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Original")
b    <-    fourfoldplot(round(cm_boost_down$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Submostreig (Down)")
c    <-    fourfoldplot(round(cm_boost_up$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió
Sobremostreig (Up)")
```

A.6) Comparació dels resultats

Test Friedman i Wicoxson

```
# Sampling = DOWN
resultados_resamples     <-     resamples(list("Logistic     Regression"     =
modelo_logistic_down
                                  ,"Single Tree" = modelo_C50Tree_down,
                                "Bagged" = modelo_bag_down,
                                "Random Forest" = modelo_rf_down,
                                "GBM" = modelo_boost_down))
resultados_resamples$values %>% head(10)

metricas_resamples <- resultados_resamples$values %>%
                          gather(key = "modelo", value = "valor", -Resample) %>%
                          separate(col = "modelo", into = c("modelo", "metrica"),
                                   sep = "~", remove = TRUE)
metricas_resamples %>% head()

matriz_metricas <- metricas_resamples %>% filter(metrica == "ROC") %>%
                   spread(key = modelo, value = valor) %>%
                   select(-Resample, -metrica) %>% as.matrix()
friedman.test(y = matriz_metricas)

# Comparacions múltiples amb un test suma de rangs de Wilcoxon
metricas_ROC<- metricas_resamples %>% filter(metrica == "ROC")
comparaciones  <- pairwise.wilcox.test(x = metricas_ROC$valor,
                                       g = metricas_ROC$modelo,
                                       paired = TRUE,
                                       p.adjust.method = "holm")

comparaciones <- comparaciones$p.value %>%
  as.data.frame() %>%
  rownames_to_column(var = "modeloA") %>%
  gather(key = "modeloB", value = "p_value", -modeloA) %>%
  na.omit() %>%
  arrange(modeloA)
metricas_resamples <- resultados_resamples$values %>%
                          gather(key = "modelo", value = "valor", -Resample) %>%
                          separate(col = "modelo", into = c("modelo", "metrica"),
                                   sep = "~", remove = TRUE)
metricas_resamples %>% head()

matriz_metricas <- metricas_resamples %>% filter(metrica == "Spec") %>%
                   spread(key = modelo, value = valor) %>%
                   select(-Resample, -metrica) %>% as.matrix()
friedman.test(y = matriz_metricas)

metricas_Sens <- metricas_resamples %>% filter(metrica == "Sens")
comparaciones  <- pairwise.wilcox.test(x = metricas_Spec$valor,
                                       g = metricas_Spec$modelo,
                                       paired = TRUE,
                                       p.adjust.method = "holm")
comparaciones <- comparaciones$p.value %>%
  as.data.frame() %>%
  rownames_to_column(var = "modeloA") %>%
  gather(key = "modeloB", value = "p_value", -modeloA) %>%
  na.omit() %>%
  arrange(modeloA)
metricas_resamples <- resultados_resamples$values %>%
                          gather(key = "modelo", value = "valor", -Resample) %>%
                          separate(col = "modelo", into = c("modelo", "metrica"),
                                   sep = "~", remove = TRUE)
metricas_resamples %>% head()

matriz_metricas <- metricas_resamples %>% filter(metrica == "Sens") %>%
                   spread(key = modelo, value = valor) %>%
```

```
                    select(-Resample, -metrica) %>% as.matrix()
friedman.test(y = matriz_metricas)

metricas_Spec <- metricas_resamples %>% filter(metrica == "Spec")
comparaciones  <- pairwise.wilcox.test(x = metricas_Spec$valor,
                                       g = metricas_Spec$modelo,
                                       paired = TRUE,
                                       p.adjust.method = "holm")
comparaciones <- comparaciones$p.value %>%
  as.data.frame() %>%
  rownames_to_column(var = "modeloA") %>%
  gather(key = "modeloB", value = "p_value", -modeloA) %>%
  na.omit() %>%
  arrange(modeloA)

# Sampling = DOWN
model_resamples_down     <-     resamples(list("Logistic     Regression"     =
modelo_logistic_down
                                   ,"Single Tree" = modelo_C50Tree_down,
                                "Bagged" = modelo_bag_down,
                                "Random Forest" = modelo_rf_down,
                                "GBM" = modelo_boost_down))
summary(model_resamples_down)

bwplot(model_resamples_down)
dotplot(model_resamples_down, metric = "ROC")
dotplot(model_resamples_down, metric = "Sens")
dotplot(model_resamples_down, metric = "Spec")

curva_roc_logistic_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_logistic_down$Si)
curva_roc_C50Tree_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_C50Tree_down$Si)
curva_roc_boost_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_boost_down$Si)
curva_roc_rf_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_down$Si)
curva_roc_bag_down <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_bag_down$Si)


roc_logistic   <-   data.frame(x   =   (1-curva_roc_logistic_down$specificities),
y=curva_roc_logistic_down$sensitivities, Model = "Regresió Logística")
roc_C50Tree    <-    data.frame(x    =    (1-curva_roc_C50Tree_down$specificities),
y=curva_roc_C50Tree_down$sensitivities, Model = "C50Tree")
roc_boost     <-     data.frame(x     =     (1-curva_roc_boost_down$specificities),
y=curva_roc_boost_down$sensitivities, Model = "Gradient Boosting")
roc_rf      <-      data.frame(x      =      (1-curva_roc_rf_down$specificities),
y=curva_roc_rf_down$sensitivities, Model = "Random Forest")
roc_bag     <-     data.frame(x     =     (1-curva_roc_bag_down$specificities),
y=curva_roc_bag_down$sensitivities, Model = "Bagging")

droc <- rbind(roc_logistic,roc_C50Tree,roc_boost, roc_rf, roc_bag)

ggplot(droc, aes(x = x, y = y, color = Model)) +
  geom_line() + labs(x= "1-Specificity", y="Sensitivity", title = "Corbes ROC
(Sampling = down)")+
  scale_fill_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  scale_colour_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  theme_bw()

# Sampling = UP
model_resamples_up <- resamples(list("Logistic Regression" = modelo_logistic_up
                                   ,"Single Tree" = modelo_C50Tree_up,
                                "Bagged" = modelo_bag_up,
                                "Random Forest" = modelo_rf_up,
                                "GBM" = modelo_boost_up))
summary(model_resamples_up)
```

```
bwplot(model_resamples_up)
dotplot(model_resamples_up, metric = "ROC")
dotplot(model_resamples_up, metric = "Sens")
dotplot(model_resamples_up, metric = "Spec")

curva_roc_logistic_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_logistic_up$Si)
curva_roc_C50Tree_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_C50Tree_up$Si)
curva_roc_boost_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_boost_up$Si)
curva_roc_rf_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_rf_up$Si)
curva_roc_bag_up <- roc(response = datos_test_prep$Y,
                  predictor = predicciones_bag_up$Si)

roc_logistic   <-   data.frame(x   =   (1-curva_roc_logistic_up$specificities),
y=curva_roc_logistic_up$sensitivities, Model = "Regresió Logística")
roc_C50Tree    <-    data.frame(x    =    (1-curva_roc_C50Tree_up$specificities),
y=curva_roc_C50Tree_up$sensitivities, Model = "C50Tree")
roc_boost     <-     data.frame(x     =     (1-curva_roc_boost_up$specificities),
y=curva_roc_boost_up$sensitivities, Model = "Gradient Boosting")
roc_rf      <-      data.frame(x      =      (1-curva_roc_rf_up$specificities),
y=curva_roc_rf_up$sensitivities, Model = "Random Forest")
roc_bag     <-     data.frame(x     =     (1-curva_roc_bag_up$specificities),
y=curva_roc_bag_up$sensitivities, Model = "Bagging")

droc <- rbind(roc_logistic,roc_C50Tree,roc_boost, roc_rf, roc_bag)

ggplot(droc, aes(x = x, y = y, color = Model)) +
  geom_line() + labs(x= "1-Specificity", y="Sensitivity", title = "Corbes ROC
(Sampling = up)")+
  scale_fill_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  scale_colour_manual(values = paletteer_d("ggthemes::excel_Badge")) +
  theme_bw()
```

Errors de test

```
# Sampling = DOWN
library(tidyr)
modelos_down <- list(Logistic = modelo_logistic_down,
                Single_Tree = modelo_C50Tree_down,
                Random_Forest = modelo_rf_down,
                GBM = modelo_boost_down,
                Bagging = modelo_bag_down)

predicciones_down <- extractPrediction(
                  models = modelos_down,
                  testX = datos_test_prep %>% select(-Y),
                  testY = datos_test_prep$Y
                )
predicciones_down %>% head()

metricas_predicciones_down <- predicciones_down %>%
                        mutate(acierto = ifelse(obs == pred, TRUE, FALSE)) %>%
                        group_by(object, dataType) %>%
                        summarise(ROC = mean(acierto))

metricas_predicciones_down %>%
  spread(key = dataType, value = ROC) %>%
  arrange(desc(Test))

ggplot(data = metricas_predicciones_down,
       aes(x = reorder(object, ROC), y = ROC,
           color = dataType, label = round(ROC, 2))) +
  geom_point( size = 10) +
```

```
  scale_color_manual(values = c("#7EC3E5", "gray50")) +
  geom_text(color = "white", size = 3) +
  scale_y_continuous(limits = c(0, 1)) +
  # ROC
  geom_hline(yintercept = 0.62, linetype = "dashed") +
  annotate(geom = "text", y = 0.66, x = 5, label = "ROC") +
  coord_flip() +
  labs(title = "ROC d'entrenament i test Submostreig (Down)",
       x = "Model") +
  theme_bw() +
  theme(legend.position = "bottom")

# Sampling = UP
library(tidyr)
modelos_up <- list(Logistic = modelo_logistic_up,
                   Single_Tree = modelo_C50Tree_up,
                   Random_Forest = modelo_rf_up,
                   GBM = modelo_boost_up,
                   Bagging = modelo_bag_up)

predicciones_up <- extractPrediction(
                     models = modelos_up,
                     testX = datos_test_prep %>% select(-Y),
                     testY = datos_test_prep$Y
                   )
predicciones_up %>% head()

metricas_predicciones_up <- predicciones_up %>%
                            mutate(acierto = ifelse(obs == pred, TRUE, FALSE)) %>%
                            group_by(object, dataType) %>%
                            summarise(ROC = mean(acierto))

metricas_predicciones_up %>%
  spread(key = dataType, value = ROC) %>%
  arrange(desc(Test))
ggplot(data = metricas_predicciones_up,
       aes(x = reorder(object, ROC), y = ROC,
           color = dataType, label = round(ROC, 2))) +
  geom_point(size = 8) +
  scale_color_manual(values = c("#7EC3E5", "gray50")) +
  geom_text(color = "white", size = 3) +
  scale_y_continuous(limits = c(0, 1)) +
  # ROC
  geom_hline(yintercept = 0.62, linetype = "dashed") +
  annotate(geom = "text", y = 0.66, x = 5, label = "ROC") +
  coord_flip() +
  labs(title = "ROC d'entrenament i test Sobremostreig (Up)",
       x = "Model") +
  theme_bw() +
  theme(legend.position = "bottom")
```

A.7) Anàlisi dels resultats

Influència de les variables sobre la variable resposta

summary(modelo_logistic_up$finalModel)

Predicció de la variable resposta

```
cm_boost_down          <-          confusionMatrix(predicciones_boost_down_resp,
datos_test_prep$Y, dnn = c("Predicciones", "Observaciones"))

c    <-    fourfoldplot(round(cm_boost_down$table,2),    color    =    c("tomato",
"darkolivegreen2"), conf.level = 0, margin = 2, main = "Matriu de confusió Down")
library(ggplot2)
```

```
library(gridExtra) # install.packages("gridExtra")
library(paletteer) # install.packages("paletteer")

predicciones_boost_down <- predict(object = modelo_boost_down,
                        newdata = datos_test_prep,
                        type = "prob")
predicciones_bag_up$Obs <- datos_test_prep$Y
er1      <-      predicciones_boost_down[predicciones_boost_down$No>0.5      &
predicciones_boost_down$Obs=="Si",]
er2      <-      predicciones_boost_down[predicciones_boost_down$No<0.5      &
predicciones_boost_down$Obs=="No",]

errors <- rbind(er1, er2)
errors$tipo <- "Error"

ac1      <-      predicciones_boost_down[predicciones_boost_down$No>0.5      &
predicciones_boost_down$Obs=="No",]
ac2      <-      predicciones_boost_down[predicciones_boost_down$No<0.5      &
predicciones_boost_down$Obs=="Si",]

aciertos <- rbind(ac1,ac2)
aciertos$tipo <- "Correcte"

ee <- rbind(errors, aciertos)

siii <- ee[,c("Si", "Obs", "tipo")]

siii1 <- siii[siii$Obs=="No", ]
siii11 <- siii1[siii1$tipo=="Correcte",]
siii12 <- siii1[siii1$tipo=="Error",]

siii2 <- siii[siii$Obs=="Si", ]
siii21 <- siii2[siii2$tipo=="Correcte",]
siii22 <- siii2[siii2$tipo=="Error",]

p1 <- ggplot() +
      # Correcte
      geom_boxplot(data=siii11, aes(x=Obs, y=Si, colour = "Correcte")) +
      # Error
      geom_jitter(data=siii12, aes(x=Obs, y=Si, colour = "Errors")) +
  scale_color_manual(values = c("olivedrab3", "red2")) + theme_bw()  +
  labs(title = "Prob. Predites vs Valor Observat", subtitle = "Prob = Si, Obs =
No")    +  theme(legend.title  =  element_blank(),  legend.position  =  "top")  +
geom_hline(yintercept=0.5,  linetype='dashed')

p2 <- ggplot() +
      # Correcte
      geom_boxplot(data=siii21, aes(x=Obs, y=Si, colour = "Correcte")) +
      # Error
      geom_jitter(data=siii22, aes(x=Obs, y=Si, colour = "Errors")) +
  scale_color_manual(values = c("olivedrab3", "red2")) + theme_bw()  +
  labs(title = "Prob. Predites vs Valor Observat", subtitle = "Prob = Si, Obs =
Si")  +   theme(legend.title  =  element_blank(),  legend.position  =  "top")  +
geom_hline(yintercept=0.5,  linetype='dashed')

grid.arrange(p1, p2, ncol=2, nrow = 1)

nooo <- ee[,c("No", "Obs", "tipo")]

nooo1 <- nooo[nooo$Obs=="No", ]
nooo11 <- nooo1[nooo1$tipo=="Correcte",]
nooo12 <- nooo1[nooo1$tipo=="Error",]

nooo2 <- nooo[nooo$Obs=="Si", ]
nooo21 <- nooo2[nooo2$tipo=="Correcte",]
nooo22 <- nooo2[nooo2$tipo=="Error",]

p3 <- ggplot() +
```

```
      # Correcte
      geom_boxplot(data=nooo11, aes(x=Obs, y=No, colour = "Correcte")) +
      # Error
      geom_jitter(data=nooo12, aes(x=Obs, y=No, colour = "Errors")) +
  scale_color_manual(values = c("olivedrab3", "red2")) + theme_bw()  +
  labs(title = "Prob. Predites vs Valor Observat", subtitle = "Prob = No, Obs =
No")   + theme(legend.title = element_blank(), legend.position = "top")  +
geom_hline(yintercept=0.5,  linetype='dashed')

p4 <- ggplot() +
      # Correcte
      geom_boxplot(data=nooo21, aes(x=Obs, y=No, colour = "Correcte")) +
      # Error
      geom_jitter(data=nooo22, aes(x=Obs, y=No, colour = "Errors")) +
  scale_color_manual(values = c("olivedrab3", "red2")) + theme_bw()  +
  labs(title = "Prob. Predites vs Valor Observat", subtitle = "Prob = No, Obs =
Si") +   theme(legend.title = element_blank(), legend.position = "top")  +
geom_hline(yintercept=0.5,  linetype='dashed')

grid.arrange(p3, p4, ncol=2, nrow = 1)
```

Importància de Variables (VarImp())

```
library(gbm)
bagging <- data.frame(varImp(modelo_bag_up)$importance)
ggplot2::ggplot(bagging, aes(x=reorder(rownames(bagging),Overall), y=Overall))
+
geom_point( color="#7EC3E5", size=4, alpha=0.6)+
geom_segment(     aes(x=rownames(bagging),     xend=rownames(bagging),     y=0,
yend=Overall),
color='gray50') +
ggtitle("Importancia de las variables", subtitle = "Bagging")+
xlab('Variable')+
ylab('Overall Importance')+
theme_light() +
coord_flip()

predicciones <- predict(object = modelo_boost_down,
                        newdata = datos_test_prep)
datos_test_prep$Yhat <- predicciones

# CORRECTES: PERFIL FAMILIES Y = 1 i PERFIL FAMILIES Y = 0
correcte1          <-          datos_test_prep[datos_test_prep$Y=="Si"          &
datos_test_prep$Yhat=="Si",]
correcte2          <-          datos_test_prep[datos_test_prep$Y=="No"          &
datos_test_prep$Yhat=="No",]

perfil_1 = correcte1 %>% select(GASTOT, SEXO, COMITOT, EDAD_grupo_X.64..100.,
ECIVILSP_X4, IMPEXAC_grupo_X.0..1100.)
summary(perfil_1)
perfil_0 = correcte2 %>% select(GASTOT, SEXO, COMITOT, EDAD_grupo_X.64..100.,
ECIVILSP_X4, IMPEXAC_grupo_X.0..1100.)
summary(perfil_0)

# INCORRECTES: PERFIL FAMILIES Y = 1 i PERFIL FAMILIES Y = 0
incorrecte1         <-         datos_test_prep[datos_test_prep$Y=="Si"         &
datos_test_prep$Yhat=="No",]
incorrecte2         <-         datos_test_prep[datos_test_prep$Y=="No"         &
datos_test_prep$Yhat=="Si",]
perfil_1 = incorrecte1 %>% select(GASTOT, SEXO, COMITOT, EDAD_grupo_X.64..100.,
ECIVILSP_X4, IMPEXAC_grupo_X.0..1100.)
summary(perfil_1)
perfil_0 = incorrecte2 %>% select(GASTOT, SEXO, COMITOT, EDAD_grupo_X.64..100.,
ECIVILSP_X4, IMPEXAC_grupo_X.0..1100.)
summary(perfil_0)
```

Bloc B. Problema de regressió

B.1) Base de dades

Importació de la base de dades

```
library(readr)
library(readxl)
library(dplyr)
setwd("~/Desktop/TFG/Dades/OBJECTIU2")
Omnivors <- read_delim("EPFClean.csv", delim = ";", escape_double = FALSE,
trim_ws = TRUE)
R <- read_delim("R_3.csv", delim = ";", escape_double = FALSE, trim_ws = TRUE)
R1 <- read_delim("R_1.csv", delim = ";", escape_double = FALSE, trim_ws = TRUE)
R1$Y <- NULL
Omnivors = Omnivors %>% left_join(R, by="CODIGO")

preuXkg <- read_excel("PreuKg.xlsx")
preuXkg

# Seleccionem files amb SUBGRUP = PROCÍ, BOVI, OVI I CAPRI I POLLASTRE
Omnivors = Omnivors %>% filter(SUBGRUP=="Carn de boví" | SUBGRUP=="Carn de
caprí"| SUBGRUP=="Carn d'Oví i Caprí" | SUBGRUP=="Au")

# Creem la variable COMPRA ; HC ; HH
bd = Omnivors %>% left_join(preuXkg, by=c("ANOENC", "SUBGRUP"))

# creació variable Y
bd = bd %>% mutate(PES = GASTO/Preu_kg_Carn) # Pes (kg) de carn consumit
bd = bd %>% mutate(Cost_MA = Preu_kg_MA*PES) # Cost (eur) Medi Ambiental x tipus
de carn
Resp = bd %>%
  group_by(NUMERO) %>%
  summarise(Y = sum(Cost_MA))

bd = R1 %>% inner_join(Resp, by ="NUMERO")
bd

write.table(bd, file = "bd.csv", sep = ";", na = "NA", dec = ".", row.names =
FALSE, col.names = TRUE)
```

B.2) Anàlisi exploratori de les dades

```
# Directori
setwd("~/Desktop/TFG/Dades/OBJECTIU2")
#setwd("C:/Users/lluis/OneDrive - Universitat de
Barcelona/Public/docencia/tfg/carn")

library(caret)
library(tidyverse)
library(dplyr)
library(skimr) # install.packages("devtools")
# Importem la base de dades: dades_11
datos <- read.csv("bd.csv", sep = ";")
datos$NUMERO <- NULL
datos$LlarID <- 1:nrow(datos)

# Assignem la classe corresponent a cada variable (numèrica o categòrica)

   # Covariables
     datos$GASTOT <- as.numeric(datos$GASTOT)
     datos$SEXO <- as.numeric(datos$SEXO)
     datos$EDAD <- as.numeric(datos$EDAD)
     datos$NACION <- as.numeric(datos$NACION)
     datos$COMITOT  <- as.numeric(datos$COMITOT)
```

```
    datos$IMPEXAC <- as.numeric(datos$IMPEXAC)
    datos$NMIEMBR1 <- as.numeric(datos$NMIEMBR1)
    datos$NMIEMBR2 <- as.numeric(datos$NMIEMBR2)
    datos$NMIEMBR3 <- as.numeric(datos$NMIEMBR3)
    datos$ESTUDIOS1 <- as.numeric(datos$ESTUDIOS1)
    datos$ESTUDIOS2 <- as.numeric(datos$ESTUDIOS2)
    datos$ESTUDIOS3 <- as.numeric(datos$ESTUDIOS3)

  # Factors
    datos$DENSIDAD <- as.factor(datos$DENSIDAD)
    datos$CCAA <- as.factor(datos$CCAA)
    datos$NUMOCUP <- as.factor(datos$NUMOCUP)
    datos$ECIVILSP <- as.factor(datos$ECIVILSP)
    datos$ZONARES <- as.factor(datos$ZONARES)
```

Distribució de la Variable Resposta (Y)

```
library(ggpubr) # install.packages("ggpubr")
p1 <- ggplot(data = datos, aes(x = Y)) +
  geom_density(lwd = 1.2, linetype = 1, colour = "#7EC3E5") +
  theme_bw()
p2 <- ggplot(data = datos, aes(x="",y = Y)) +
      geom_point(colour = "#7EC3E5")+
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15, colour = "#7EC3E5" ) +
      scale_color_manual(values = c("#7EC3E5")) +
      scale_fill_manual(values = c("#7EC3E5")) +
      theme_bw()
final_plot1 <- ggarrange(p1, p2, legend = "top")
final_plot1 <- annotate_figure(final_plot1, top = text_grob("Y", size = 15))
final_plot1

p1 <- ggplot(data = datos, aes(x = log(Y))) +
  geom_density(lwd = 1.2, linetype = 1, colour = "#7EC3E5") +
  theme_bw()
p2 <- ggplot(data = datos, aes(x="",y = log(Y))) +
      geom_point(colour = "#7EC3E5")+
      geom_boxplot(outlier.shape = NA) +
      geom_jitter(alpha = 0.3, width = 0.15, colour = "#7EC3E5" ) +
      scale_color_manual(values = c("#7EC3E5")) +
      scale_fill_manual(values = c("#7EC3E5")) +
      theme_bw()
final_plot1 <- ggarrange(p1, p2, legend = "top")
final_plot1 <- annotate_figure(final_plot1, top = text_grob("log(Y)", size =
15))
final_plot1
```

Anàlisi de Correlacions

Correlació variables contínues

```
library(ggcorrplot)
library(dplyr)
library(car)
vcontinuas = datos %>% select (SEXO, EDAD, NACION, COMITOT, IMPEXAC, NMIEMBR1,
NMIEMBR2, NMIEMBR3, ESTUDIOS1, ESTUDIOS2, ESTUDIOS3, Y)
corr <- round(cor(vcontinuas), 1) # coef correlacio
p.mat <- cor_pmat(vcontinuas) # pvalors
ggcorrplot(corr, hc.order = TRUE, type = "lower",
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"), lab = TRUE)

# simple_lm <- lm (Y ~., data = datos)
# vif (simple_lm)
```

```
# datos_cor = cor(datos[,c("SEXO", "EDAD", "NACION", "COMITOT", "IMPEXAC",
"NMIEMBR1", "NMIEMBR2", "NMIEMBR3", "ESTUDIOS1", "ESTUDIOS2", "ESTUDIOS3")])
# findCorrelation(datos_cor, verbose = T, names = T)
library(dplyr)
# creem la nova variable ESTUDIOS
estud = datos %>% select(LlarID, ESTUDIOS2, ESTUDIOS3) %>% group_by(LlarID) %>%
  summarise(max=max(ESTUDIOS2, ESTUDIOS3))

dd = datos %>% left_join(estud, by="LlarID")

CLASE1 = dd %>% filter(max==ESTUDIOS2 & max!=ESTUDIOS3) %>% mutate(ESTUDIOS =
1)
CLASE2 = dd %>% filter(max==ESTUDIOS3 & max!=ESTUDIOS2) %>% mutate(ESTUDIOS =
2)
CLASE3 = dd %>% filter(max==ESTUDIOS2 & max==ESTUDIOS3) %>% mutate(ESTUDIOS =
3)

df = rbind(CLASE1, CLASE2, CLASE3)

datos = df %>% select(ANOENC, GASTOT, SEXO, EDAD, NACION, DENSIDAD, CCAA,
COMITOT, NUMOCUP, IMPEXAC, ECIVILSP, ZONARES, Y, ESTUDIOS1, ESTUDIOS)
datos$ESTUDIOS <- as.factor(datos$ESTUDIOS)
library(ggcorrplot)
library(dplyr)
library(car)
vcontinuas = datos %>% select (SEXO, EDAD, NACION, COMITOT, IMPEXAC, ESTUDIOS1,
Y)
corr <- round(cor(vcontinuas), 1) # coef correlacio
p.mat <- cor_pmat(vcontinuas) # pvalors
ggcorrplot(corr, hc.order = TRUE, type = "lower",
   outline.col = "white",
   ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"), lab = TRUE)

datos_cor = cor(datos[,c("SEXO", "EDAD", "NACION", "COMITOT", "IMPEXAC", "Y")])
findCorrelation(datos_cor, verbose = T, names = T)

library(tidyverse)
datos %>%
  select(SEXO, EDAD, NACION, COMITOT, IMPEXAC, ESTUDIOS1, Y) %>%
  pivot_longer(!Y) %>%
  mutate(name = factor(name, levels = c("SEXO", "EDAD", "NACION", "COMITOT",
"IMPEXAC", "ESTUDIOS1"))) %>%
  ggplot(aes(x=value, y=Y)) +
  geom_point() +
  facet_wrap(~name, scales="free") +
  theme_bw()
library(car)
simple_lm <- lm (Y ~., data = datos)
vif (simple_lm)
```

Valors atípics o *outliers*

```
ggplot(datos) +
  aes(x = "", y = SEXO) +
  geom_boxplot() +
  theme_bw() + labs(title = "Variable SEXO")

ggplot(datos) +
  aes(x = "", y = EDAD) +
  geom_boxplot() +
  theme_bw() + labs(title = "Variable EDAD")
ggplot(datos) +
  aes(x = "", y = COMITOT) +
  geom_boxplot() +
  theme_bw() + labs(title = "Variable COMITOT")
```

```
out <- boxplot.stats(datos$COMITOT)$out
out_ind <- which(datos$COMITOT %in% c(out))

datos[out_ind, ]
datos = datos[-out_ind, ]
library(ggpubr) # install.packages("ggpubr")
p1 <- ggplot(data = datos, aes(x = COMITOT)) +
  geom_density(lwd = 1.2, linetype = 1, colour = "gray50") +
  theme_bw()
p2 <- ggplot(data = datos, aes(x="",y = COMITOT)) +
  geom_point(colour = "gray50")+
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15, colour = "gray50" ) +
  scale_color_manual(values = c("gray50")) +
  scale_fill_manual(values = c("gray50")) +
  theme_bw()
final_plot1 <- ggarrange(p1, p2, legend = "top")
final_plot1 <- annotate_figure(final_plot1, top = text_grob("COMITOT", size =
15))
final_plot1
ggplot(datos) +
  aes(x = "", y = IMPEXAC) +
  geom_boxplot() +
  theme_bw() + labs(title = "Variable IMPEXAC")

out <- boxplot.stats(datos$IMPEXAC)$out
out_ind <- which(datos$IMPEXAC %in% c(out))

datos[out_ind, ]
datos = datos[-out_ind, ]
p1 <- ggplot(data = datos, aes(x = IMPEXAC)) +
  geom_density(lwd = 1.2, linetype = 1, colour = "gray50") +
  theme_bw()
p2 <- ggplot(data = datos, aes(x="",y = IMPEXAC)) +
  geom_point(colour = "gray50")+
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15, colour = "gray50" ) +
  scale_color_manual(values = c("gray50")) +
  scale_fill_manual(values = c("gray50")) +
  theme_bw()
final_plot1 <- ggarrange(p1, p2, legend = "top")
final_plot1 <- annotate_figure(final_plot1, top = text_grob("IMPEXAC", size =
15))
final_plot1
ggplot(datos) +
  aes(x = "", y = Y) +
  geom_boxplot() +
  theme_bw() + labs(title = "Variable Y")

out <- boxplot.stats(datos$Y)$out
out_ind <- which(datos$Y %in% c(out))

datos[out_ind, ]
datos = datos[-out_ind, ]

p1 <- ggplot(data = datos, aes(x = Y)) +
  geom_density(lwd = 1.2, linetype = 1, colour = "gray50") +
  theme_bw()
p2 <- ggplot(data = datos, aes(x="",y = Y)) +
  geom_point(colour = "gray50")+
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(alpha = 0.3, width = 0.15, colour = "gray50" ) +
  scale_color_manual(values = c("gray50")) +
  scale_fill_manual(values = c("gray50")) +
  theme_bw()
final_plot1 <- ggarrange(p1, p2, legend = "top")
final_plot1 <- annotate_figure(final_plot1, top = text_grob("Y", size = 15))
final_plot1
```

B.3) Divisió de les dades en entrenament i test

```
set.seed(123)
train <- createDataPartition(y = datos$Y, p = 0.8, list = FALSE, times = 1)
datos_train <- datos[train, ]
datos_test  <- datos[-train, ]
summary(datos_train$Y)
summary(datos_test$Y)
```

B.4) Preprocessament de les dades

Exclusió de variables amb variància propera a zero

```
library(recipes) # install.packages("recipes", dependencies = TRUE)

objeto_recipe <- recipe(formula = Y ~ SEXO + EDAD + NACION + DENSIDAD + CCAA +
COMITOT + NUMOCUP + IMPEXAC + ECIVILSP + ZONARES + ESTUDIOS1 + ESTUDIOS ,data =
datos_train)
objeto_recipe

datos %>% select(SEXO, EDAD, NACION, DENSIDAD, CCAA, COMITOT, NUMOCUP, IMPEXAC,
ECIVILSP, ZONARES, ESTUDIOS1, ESTUDIOS) %>%
  nearZeroVar(saveMetrics = TRUE)
```

B.5) Models. Ajust, optimització i validació del model

```
library(tidyverse)
p1 <- ggplot(data = datos_train, aes(x = Y)) +
     geom_histogram(fill = "cornflowerblue") +
     labs(title = "Distribució del Cost Ocult",
          x = "Y") +
     theme_bw() +
     theme(plot.title = element_text(face = "bold"))

p2 <- ggplot(data =  datos_train, aes(x = Y)) +
      stat_ecdf(geom = "step", color = "cornflowerblue", size = 1) +
      labs(title = "Funció de distribució empírica",
           x = "Y",
           y = "CDF") +
      theme_bw() +
      theme(plot.title = element_text(face = "bold"))

ggpubr::ggarrange(plotlist = list(p1, p2), ncol = 2)

library(univariateML)
comparacion_aic <- AIC(
                    mlbetapr(datos_train$Y),
                    mlexp(datos_train$Y),
                    mlinvgamma(datos_train$Y),
                    mlgamma(datos_train$Y),
                    mllnorm(datos_train$Y),
                    mlrayleigh(datos_train$Y),
                    mlinvgauss(datos_train$Y),
                    mlweibull(datos_train$Y),
                    mlinvweibull(datos_train$Y),
                    mllgamma(datos_train$Y)
                   )
comparacion_aic %>% rownames_to_column(var = "Distribució") %>% arrange(AIC)
```

```
comparacion_bic <- BIC(
                    mlbetapr(datos_train$Y),
                    mlexp(datos_train$Y),
                    mlinvgamma(datos_train$Y),
                    mlgamma(datos_train$Y),
                    mllnorm(datos_train$Y),
                    mlrayleigh(datos_train$Y),
                    mlinvgauss(datos_train$Y),
                    mlweibull(datos_train$Y),
                    mlinvweibull(datos_train$Y),
                    mllgamma(datos_train$Y)
                   )
comparacion_bic %>% rownames_to_column(var = "Distribució") %>% arrange(BIC)


p1 <- ggplot(data =  datos_train_prep) +
  geom_histogram(aes(x = Y, y =  after_stat(density)),
                 bins = 40,
                 alpha = 0.3, color = "black") +
  geom_rug(aes(x = Y)) +
  stat_function(fun    =    function(.x){dml(x    =    .x,    obj    =    mlweibull(
datos_train_prep$Y))},
                aes(color = "Weibull"),
                size = 1) +
  stat_function(fun    =    function(.x){dml(x    =    .x,    obj    =    mlgamma(
datos_train_prep$Y))},
                aes(color = "Gamma"),
                size = 1) +
  scale_color_manual(breaks = c("Weibull", "Gamma"),
                     values = c("Weibull" = "red", "Gamma" = "blue")) +
  labs(title = "Funció de densitat",
       color = "Distribució") +
  theme_bw() +
  theme(legend.position = "bottom")

p2 <- ggplot(data =  datos_train_prep) +
  stat_ecdf(aes(x = Y), geom = "step", color = "black", size = 1) +
  geom_rug(aes(x = Y)) +
  stat_function(fun    =    function(.x){pml(q    =    .x,    obj    =    mlweibull(
datos_train_prep$Y))},
                aes(color = "Weibull"),
                size = 1) +
  stat_function(fun    =    function(.x){pml(q    =    .x,    obj    =    mlgamma(
datos_train_prep$Y))},
                aes(color = "Gamma"),
                size = 1) +
  scale_color_manual(breaks = c("Weibull", "Gamma"),
                     values = c("Weibull" = "red", "Gamma" = "blue")) +
 labs(title = "Funció de Distribució Acumulada",
       color = "Distribució",
       y = "CDF") +
  theme_bw() +
  theme(legend.position = "bottom")

ggpubr::ggarrange(plotlist = list(p1, p2), ncol = 2)
```

Validació distribució Weibull i Gamma

```
library(fitdistrplus)
# Descripció de la distribució
descdist(data = datos_train_prep$Y, graph = FALSE)

fit.weibull <- fitdist(datos_train_prep$Y, distr = "weibull", method = "mle",
lower=c(0,0))
fit.gamma  <-  fitdist(datos_train_prep$Y,  distr  =  "gamma",  lower=c(0,0),
start=list(scale=1,shape=1))
```

```
# Representació gràfica de las distribucions
plot(fit.weibull)
plot(fit.gamma)

validacio_WEIBULL <- gofstat(fit.weibull)
validacio_WEIBULL

# Test Kolmogorov – Smirnoff (K-S)
validacio_WEIBULL$kstest

validacio_GAMMA <- gofstat(fit.gamma)
validacio_GAMMA

# Test Kolmogorov – Smirnoff (K-S)
validacio_GAMMA$kstest
comparacio <- gofstat(f = list(fit.weibull, fit.gamma))
comparacio
```

Model Weibull i Gamma

```
# AJUST DEL MODEL GAMLSS: WEIBULL
library(gamlss)
library(splines) # install.packages("splines")
library(gamlss.data) # install.packages("gamlss.data")
library(gamlss.dist) # install.packages("gamlss.dist")
library(MASS) # install.packages("MASS")
library(nlme) # install.packages("nlme")
m0 <- gamlss(
              formula = Y ~ SEXO + EDAD + COMITOT + IMPEXAC + DENSIDAD + CCAA +
NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS,
              sigma.formula =  Y ~ SEXO + EDAD + COMITOT + IMPEXAC + DENSIDAD +
CCAA + NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS,
              family  = WEI,
              data    =  datos_train_prep,
              trace   = FALSE
             )
summary(m0)
plot(m0)
# Worm plot de los residus
wp(m0,ylim.all = 1)
term.plot(m0, pages = 1, ask = FALSE, rug = TRUE)
# AJUST DEL MODELO GAMLSS: WEIBULL (p splines)
library(gamlss)
library(splines) # install.packages("splines")
library(gamlss.data) # install.packages("gamlss.data")
library(gamlss.dist) # install.packages("gamlss.dist")
library(MASS) # install.packages("MASS")
library(nlme) # install.packages("nlme")
m1 <- gamlss(formula = Y ~ pb(SEXO) + pb(EDAD) + pb(COMITOT) + pb(IMPEXAC) +
DENSIDAD + CCAA + NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS,
              sigma.formula =   Y  ~   pb(SEXO)  +  pb(EDAD)  +  pb(COMITOT)  +
pb(IMPEXAC) + DENSIDAD + CCAA + NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS ,
              family  = WEI,
              data    =  datos_train_prep)
summary(m1)
plot(m1)
# Worm plot de los residus
wp(m1,ylim.all = 1)
term.plot(m1, pages = 1, ask = FALSE, rug = TRUE)

# AJUST DEL MODEL GAMLSS: WEIBULL (cubic splines)
library(gamlss)
library(splines) # install.packages("splines")
library(gamlss.data) # install.packages("gamlss.data")
library(gamlss.dist) # install.packages("gamlss.dist")
library(MASS) # install.packages("MASS")
```

```
library(nlme) # install.packages("nlme")
m2 <- gamlss(formula = Y ~ cs(SEXO,df=10) + cs(EDAD,df=10) + cs(COMITOT,df=10)
+ cs(IMPEXAC,df=10) + DENSIDAD + CCAA + NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS,
             sigma.formula  =    Y  ~  cs(SEXO,df=10)  +  cs(EDAD,df=10)  +
cs(COMITOT,df=10) + cs(IMPEXAC,df=10) + DENSIDAD + CCAA + NUMOCUP + ECIVILSP +
ZONARES + ESTUDIOS ,
             family  = WEI,
             data    =  datos_train_prep)
summary(m2)
plot(m2)
# Worm plot de los residus
wp(m2,ylim.all = 1)
term.plot(m2, pages = 1, ask = FALSE, rug = TRUE)
GAIC(m0,m1,m2)

# AJUST DEL MODEL GAMLSS: GAMMA
library(gamlss)
library(splines) # install.packages("splines")
library(gamlss.data) # install.packages("gamlss.data")
library(gamlss.dist) # install.packages("gamlss.dist")
library(MASS) # install.packages("MASS")
library(nlme) # install.packages("nlme")
m3 <- gamlss(
             formula = Y ~ SEXO + EDAD + COMITOT + IMPEXAC + DENSIDAD + CCAA +
NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS,
             sigma.formula =  Y ~ SEXO + EDAD + COMITOT + IMPEXAC + DENSIDAD +
CCAA + NUMOCUP + ECIVILSP + ZONARES + ESTUDIOS,
             family  = GA,
             data    =  datos_train_prep,
             trace   = FALSE
            )
summary(m0)
plot(m0)
# Worm plot de los residus
wp(m3,ylim.all = 1)
term.plot(m3, pages = 1, ask = FALSE, rug = TRUE)
AIC(m0, m3)
```

Validació model Weibull

1. Bondat d'ajust

```
set.seed(123)
predictions_train <-  predict(
  object  = m0,
  what    = "mu",
  type    = "response",
  newdata = datos_train_prep)

data.frame( R2 = R2(predictions_train, datos_train_prep$Y),
            RMSE = RMSE(predictions_train, datos_train_prep$Y),
            MAE = MAE(predictions_train, datos_train_prep$Y))

n <- 35398
k <- 10
R2ajustat <- 1 - (((n-1)/(n-k-1)))*(1-0.01067289)
R2ajustat
```

2. Anàlisi dels residus

2.1. Anàlisi gràfic dels residus

```
res <- resid(m0)
data <- data.frame(Residuals = res, Fitted =fitted(m0) )
ggplot(data, aes(x = Fitted, y = Residuals)) +
```

```
  geom_point() + theme_bw() + labs(title = "Residuals vs Fitted Values. Weibull
distribution") +  geom_hline(yintercept = 0,
             linetype = 3,
             color = "red",
             lwd = 1)
```

2.2. Anàlisi quantitatiu dels residus

```
# Test Durbin-Watson
library(lmtest)
dwtest(formula = m0,  alternative = "two.sided")
```

3. Avaluació fora de la mostra (dades training)

```
# predicting the target variable
set.seed(123)
predictions_test <-  predict(
  object  = m0,
  what    = "mu",
  type    = "response",
  newdata = datos_test_prep)

set.seed(123)
predictions_train <-  predict(
  object  = m0,
  what    = "mu",
  type    = "response",
  newdata = datos_train_prep)

# computing model performance metrics
# Reference
data.frame( R2 = R2(predictions_train, datos_train_prep$Y),
            RMSE = RMSE(predictions_train, datos_train_prep$Y),
            MAE = MAE(predictions_train, datos_train_prep$Y))
# Predictions
data.frame( R2 = R2(predictions_test, datos_test_prep$Y),
            RMSE = RMSE(predictions_test, datos_test_prep$Y),
            MAE = MAE(predictions_test, datos_test_prep$Y))
```

Validació model Gamma

1. Bondat d'ajust

```
set.seed(123)
predictions_train <-  predict(
  object  = m3,
  what    = "mu",
  type    = "response",
  newdata = datos_train_prep)

data.frame( R2 = R2(predictions_train, datos_train_prep$Y),
            RMSE = RMSE(predictions_train, datos_train_prep$Y),
            MAE = MAE(predictions_train, datos_train_prep$Y))

n <- 35398
k <- 10
R2ajustat <- 1 - (((n-1)/(n-k-1)))*(1-0.01067516)
R2ajustat
```

2. Anàlisi dels residus

2.1. Anàlisi gràfic dels residus

```
res <- resid(m3)
data <- data.frame(Residuals = res, Fitted =fitted(m3) )
```

```
ggplot(data, aes(x = Fitted, y = Residuals)) +
  geom_point() + theme_bw() + labs(title = "Residuals vs Fitted Values. Gamma
distribution") +  geom_hline(yintercept = 0,
             linetype = 3,
             color = "red",
             lwd = 1)
```

2.2. Anàlisi quantitatiu dels residus

```
# Test Durbin-Watson
library(lmtest)
dwtest(formula = m3,  alternative = "two.sided")
```

3. Avaluació fora de la mostra (dades training)

```
# predicting the target variable
set.seed(123)
predictions_test <-  predict(
  object  = m3,
  what    = "mu",
  type    = "response",
  newdata = datos_test_prep)

set.seed(123)
predictions_train <-  predict(
  object  = m3,
  what    = "mu",
  type    = "response",
  newdata = datos_train_prep)

# computing model performance metrics
# Reference
data.frame( R2 = R2(predictions_train, datos_train_prep$Y),
            RMSE = RMSE(predictions_train, datos_train_prep$Y),
            MAE = MAE(predictions_train, datos_train_prep$Y))
# Predictions
data.frame( R2 = R2(predictions_test, datos_test_prep$Y),
            RMSE = RMSE(predictions_test, datos_test_prep$Y),
            MAE = MAE(predictions_test, datos_test_prep$Y))
```

B.6) Anàlisi dels resultats

Influència de les variables sobre la variable resposta

Model Weibull

```
summary(m0)
```

Model Gamma

```
summary(m3)
```