TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi de la severitat dels accidents de trànsit a la ciutat de Barcelona

Autor: Gemma Ruiz Mascaró

Director: Lluís Bermúdez Morata

Departament: Matemàtica Econòmica, Financera i Actuarial

Convocatòria: 2n Semestre del curs 2023/24 (Juny 2024)

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

AGRAÏMENTS

Primerament, agrair a la meva mare i la meva germana per tenir tanta paciència i impulsar-me a estudiar el grau i donar-me suport des del principi.

A l'Albert, per estar en els meus moments més durs i animar-me i evadir-me quan ha calgut.

I, al tutor del treball, en Lluís Bermúdez, que m'ha ajudat des del principi i ha dedicat el seu temps en corregir-me els errors i aportar els seus coneixements en el tema per dur a terme un bon treball.

RESUM

Aquesta anàlisi té com a finalitat determinar els factors que incideixen en la gravetat dels accidents de trànsit a Barcelona durant el període comprés entre el 2020 i el 2023. Per aconseguir-ho, utilitzarem el conjunt de dades disponible a la plataforma Open Data Barcelona i emprarem models interpretables i d'altres de caixa negra.

A través d'aquests models, serem capaços d'identificar les diferents causes i patrons subjacents als accidents, així com les característiques que defineixen els usuaris amb major risc d'estar implicats en ells. Amb aquest coneixement, l'objectiu és dissenyar estratègies efectives per reduir la severitat d'accidents i, per tant, millorar la seguretat viària a la ciutat.

Paraules clau: Model de regressió logística, Arbres de decisió, Random Forest, XGBoost, anàlisi de dades, anàlisi descriptiu de les dades, SQL.

Title: ANALYSIS OF THE SEVERITY OF TRAFFIC ACCIDENTS IN THE CITY OF BARCELONA

ABSTRACT

The objective of the analysis is to determine the factor that impacts the severity of traffic accidents in Barcelona between 2020 and 2023.We are making use of data available at the Open Data Barcelona platform, We will use interpretable models as well as black box models

Making use of these models, we will be able to identify the different traffic accidents causes and subjacent patterns, as well as the characteristics which will determine the subjects who are at higher risk to be involved in a traffic accident. The aim is to use the knowledge to be able to design and implement strategies to reduce the severity of traffic accidents and therefore improve road safety in the city.

Keywords: Logistic regression model, Decision trees, Random Forest, XGBoost, data analysis, descriptive data analysis, SQL,

CLASSIFICACIÓ AMS

62M20 Prediction

62-07 Data analysis

62J12 – Generalized Linear Models

TAULA DE CONTINGUTS

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

I. INTRODUCCIÓ

El *Global status report on road safety 2023* organitzat per la *World Health Organization* amb les dades del 2021, comptabilitzen 50 milions de persones ferides i 1,19 milions de morts a causa d'accidents de trànsit arreu del món i, estimen de cara al 2030 reduir les víctimes mortals a la meitat.

Així doncs, els accidents de trànsit, actualment, són un greu problema, sobretot a les grans ciutats, degut a que més del 50% de les morts són vianants, ciclistes o motociclistes i, més del 90% dels accidents són causa del factor humà.

Per tant, per tal de millorar la seguretat viària i reduir el nombre d'accidents, és necessari aprofundir en els factors que influeixen en la gravetat dels accidents. D'aquesta manera, mitjançant les dades proporcionades per les entitats responsables i una modelització adequada, permetrà definir perfils de risc, patrons de conducta i diferents indicadors de risc.

I. Objectius del treball

L'objectiu principal del treball consisteix en comprendre els factors que afecten i en quin grau ho fan en la severitat dels accidents de trànsit.

En primer lloc, haurem d'estudiar i tractar la base de dades per aconseguir la òptima i així poder fer un anàlisi exhaustiu i de qualitat per aconseguir unes conclusions basades en evidències. Per fer això, s'han hagut d'agafar un conjunt de taules i estudiar les variables que realment interessaven per l'estudi i solucionar els possibles problemes que podrien tenir per tal, de fer un bon anàlisi posterior.

En segon lloc, es modelitzaran les dades a través de l'ús de models interpretables i d'altres de caixa negra. Amb l'objectiu d'interpretar quins són de major risc i, veure quin dels models ens dona millors prediccions i poder comparar-les totes.

Finalment, l'objectiu específic consisteix en crear un escenari que ens indiqui quin seria el pitjor cas mitjançant el model escollit. Aquest escenari, es podran desenvolupar estratègies per a reduir el nombre d'accidents fatals i millorar la seguretat viària a la ciutat. De la mateixa manera, es podrà contribuir al disseny de mesures preventives per tal de garantir la màxima seguretat a la ciutadania.

Aquest estudi, pot ser de gran utilitat pel Servei de Trànsit, l'Ajuntament de Barcelona i la Guardia Urbana, qui s'encarrega de la seguretat viària del territori, amb la finalitat de preveure on cal tenir major control i reforços en cas d'accident.

II. METODOLOGIA

Al llarg d'aquest apartat, es descriurà el procediment emprat per dur a terme el treball per aplicar les diferents tècniques de modelització en el problema dels accidents de trànsit.

Per dur a terme el projecte, hem utilitzat un conjunt bases de dades públiques que es poden trobar al Open Data de Barcelona. Aquestes dades, fan referència als accidents gestionats per la Guàrdia Urbana durant el període comprés entre 2020 i 2023.

Aquestes dades han estat tractades per tal d'aconseguir una única base de dades amb tota la informació necessària, triant únicament aquelles variables i creant d'altres en funció de les que ja havien amb la finalitat de transformar-les en altres més eficients per obtenir resultats. I, posteriorment s'han creat dos conjunts de dades, un de prova i l'altre d'entrenament.

Tot seguit, s'han aplicat 4 algorismes basats en l'aprenentatge automàtic. S'han estudiat, el model de classificació logística, que s'utilitza per predir la probabilitat de que una observació pertanyi a una classe particular; arbres de decisió, on seguint una estructura jeràrquica de regles, es prenen decisions; *Random Forest* és un conjunt d'arbres de decisió que entrenament i prediuen com millorar la precisió de classificació; XGBoost, crea models molt eficients utilitzant tècniques de regularització i optimització avançades.

Obtinguts els resultats, s'han analitzat individualment i comparat amb la resta de models utilitzant diferents mètriques d'avaluació i rendiment com ara la sensibilitat, especificitat, matriu de confusió i l'àrea sota la corba de ROC.

Finalment, s'ha triat el millor model mitjançant metodologia xAI[1] i s'han creat escenaris segons el grau de severitat de l'accident en si és mortal o no.

Tot aquest procediment, s'ha dut a terme emprant el software R, mitjançant la interfície R-Studio. Principalment, s'han fet servir els parquets "*sqldf"* , "*caret", "dalex"* i *"iml".*

[1]La Intel·ligència Artificial eXplicable, és una metodologia que permet comprendre i explicar els resultats creats amb algoritmes d'aprenentatge automàtic.

III. MARC CONCEPTUAL

I. Model de regressió logístic

El model de regressió logística és un tipus d'anàlisi de classificació on hi ha una variable resposta qualitativa i una o més variables explicatives que, a diferencia de la variable depenent, aquestes poden ser qualitatives o quantitatives.

L'objectiu d'aquest model consisteix en estudiar l'efecte que les variables independents tenen sobre la variable dependent. Així doncs, estimarem la probabilitat de que es presenti l'esdeveniment d'interès.

Anàlisi de regressió logística, està dins dels Models Lineals Generalitzats (GLM, en anglès) i, per tant, farà servir la funció lògit com a funció d'enllaç. A més, aquest model, generalment analitza dades que segueixen una distribució Binomial.

D'aquesta manera, al utilitzar la funció enllaç lògit , per representar els resultats del model de manera lineal, s'haurà de fer una transformació logarítmica, en conseqüència quedarà una expressió pel model donada per:

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$$

Aquesta expressió, és equivalent a:

$$p_i = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_m X_m}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_m X_m}}$$

On p_i és la probabilitat condicionada per tal de que la variable resposta sigui igual a 1, en el cas d'aquest estudi, que l'accident sigui mortal.

Per poder calcular la probabilitat, primer, hem de calcular els valors dels paràmetres. L'estimació d'aquests en el model de regressió logística es fa basant-se en el criteri de màxima versemblança.

Sigui (x_i, y_i) amb i=1...n una mostra d'observacions independents, podem expressar la funció de màxima versemblança com:

$$L(\beta|Y,X) = \prod_{i=1}^{n} \frac{(e^{\beta_0+\beta_1 X_1+\cdots+\beta_m X_m)^{\delta_i}}}{1 + (e^{\beta_0+\beta_1 X_1+\cdots+\beta_m X_m)'}}$$

On $\beta' = (\beta_0, \beta_1 \ldots \beta_m)'$ i $\delta_i = 1$ si Y=1 i $\delta_i = 0$ si Y=0.

Per la interpretació dels paràmetres, es fan servir els *odds ratio*, ja que ens permet quantificar l'associació entre la variable resposta i la variable explicativa i, el calcularem de la següent manera:

$$OR_{x_i} = \frac{odds(Y = 1|x_1, \dots x_i = 1, \dots x_m)}{odds(Y = 1|x_1, \dots x_i = 0, \dots x_m)} = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_i + \cdots + \beta_m}}{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_m}}$$

I, el resultat l'interpretarem com un increment o disminució de la probabilitat de que l'accident sigui mortal o no. Un coeficient positiu, al transformar-lo, serà superior a 1, i per tant, el valor de la *Odds Ratio* augmentarà. En cas contrari, si un coeficient és negatiu, la transformació serà inferior a 1 i la *Odds Ratio* disminuirà.

II. Arbres de decisió

Els arbres de decisió consisteixen, en un algoritme d'aprenentatge supervisat no paramètric que permet dur a terme tasques de classificació i regressió.

És un model predictiu on donat un conjunt de dades es fabriquen diagrames de construccions lògiques que consisteixen en dividir els espais dels predictors agrupant observacions amb valors similars de la variable resposta.

Per construir un arbre de decisió, necessitem un node arrel que és el punt inicial d'on comencen totes les branques i, representa la decisió o la condició principal que divideix el conjunt de dades en diferents subgrups. Les branques que surten del node arrel coneguts com nodes de decisió, van dividint el subconjunt de dades en grups més petits fins arribar als nodes fulla, que se'ls hi assiga una classe, és a dir, un valor de la variable resposta. L'estructura de l'arbre seria la següent:

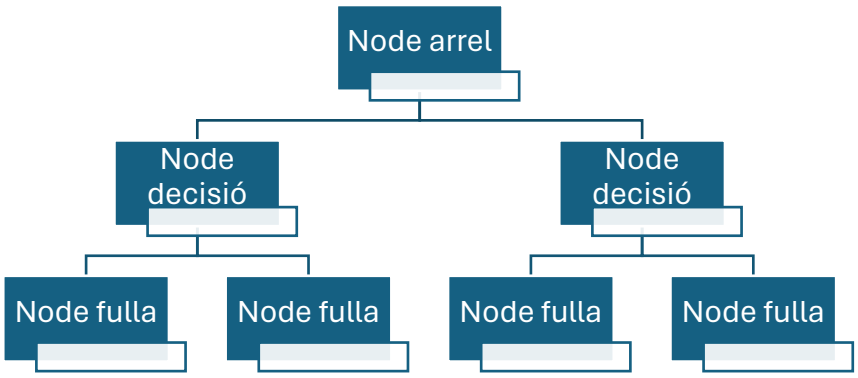

Figura III.1: Estructura arbres de decisió

Aquest tipus d'algoritme, és simple i clara, pel que permet una fàcil presa de decisions degut a que ens permet veure el camí seguit fins arribar a la decisió presa. Encara que, a vegades els arbres de decisió poden estar sobreajustats i força inestables ja que petites variacions a les dades poden resultar en arbres molt diferents és, per aquest motiu, que s'apliquen mètodes com els que es veuran a continuació, per solucionar el sobreajut i la inestabilitat i, per tant, prendre una interpretació directa.

III. Random Forest

El *Random Forest* consisteix en una tècnica d'aprenentatge automàtic que segueix un mètode ensemblista on amb la combinació de múltiples arbres de decisió genera prediccions precises i robustes.

Cada arbre és entrenat amb una mostra aleatòria de les dades i es fa servir una combinació de bootstrap i mostreig aleatori per garantir una diversitat de mostres. S'escolliran les característiques aleatòriament, per poder crear observacions i així, crear un bosc d'arbres de decisió on es realitzarà la moda dels resultats.

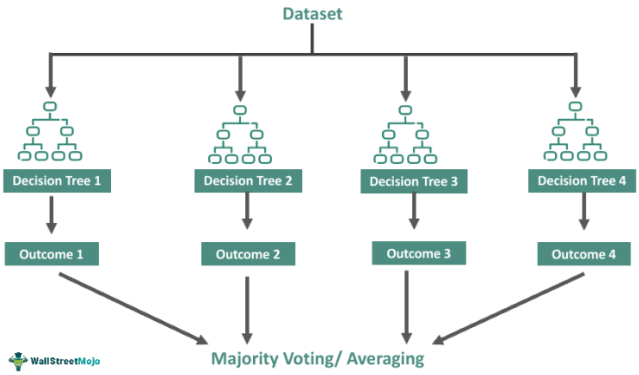

Figura III.2: Estructura del Random Forest

Font: *WallStreetMojo*

L'algoritme del *Random Forest* és el següent:

1. *Bagging:* es prenen mostres aleatòries amb reemplaçament de les dades d'entrenament.
2. Construcció d'arbres de decisió (*bootstrap sampling)*: es construeix i s'entrena un arbre de decisió per a cada mostra de dades mitjançant unes característiques diferents en cada divisió de l'arbre
3. Predicció i resultat final: una vegada s'han construït tots els arbres, cada arbre escull la classe o valor de regressió corresponen a una nova instancia de dades. En el cas d'una predicció per classificació, l'algoritme realitzarà una votació la

moda decidirà el arbre final. En canvi, si tenim un model de regressió, s'escollirà la mitjana aritmètica.

IV. XGBoost

XGBoost o eXtreme Gradient Boosting és un algoritme d'aprenentatge supervisat basat en la tècnica del *boosting*. Aquest proporciona estimacions de la importància de les característiques.

Per poder definir aquest algoritme, primerament hem de saber que és el *boosting.* El *boosting* és una técnica que consisteix en combinar múltiples models dèbils per crear-ne un de fort, construint models seqüencialment on cada model nou intenta corregir els errors anteriors. En el cas del XGBoost, aquests models dèbils son els arbres de decisió.

L'algoritme del XGBoost segueix un seguit de pasos per crear el model final fent servir el *boosting* basat en gradients. Els passos són els següents:

1. Agafem un model base feble per iniciar el procés, és a dir, un arbre de decisió poc profund.
2. S'entrena el model base amb les dades d'entrenament i es genera una predicció lineal.
3. Es calculen els residus, que són la diferència entre els valors observats i els esperats. Aquest càlcul, representa l'error que hem de corregir pel model.
4. Creem un nou model, corregint els errors de l'anterior.
5. S'actualitza el model afegint-hi el nou model ajustat, sumant les prediccions del nou model a les anteriors prediccions.
6. S'iteren els pasos 3 i 4 millorant els models a cada iteracció.
7. Per acabar, la predicció s'obté sumant les prediccions de tots els models ajustats al pas anterior.

V. Mètodes d'avaluació

Hi ha diverses maneres per calcular la bondat d'ajust d'un model, en aquest cas, ens centrarem en la matriu de confusió i l'àrea sota la corba de ROC (AUC).

La matriu de confusió és una eina que ens permet conèixer i avaluar quant de bo és el nostre model de classificació. Aquesta matriu consisteix en una taula on es mostra la distribució dels objectes en funció del valor que pren la variable resposta, així doncs, aconseguirem una matriu que tindrà la següent forma:

	Valors predits	
Valors reals	**Positius**	**Negatius**
Positius	Veritable positiu	Fals negatiu
Negatius	Fals positiu	Veritable negatiu

Taula III.1: Matriu de confusió

D'altra banda, tenim l'àrea sota la corba de ROC, que s'utilitza per mesurar l'encert en la predicció d'esdeveniments binaria i, representa la relació entre la taxa de veritables positius (sensibilitat) i la de falsos positius (especificitat) per diferents nivells possibles per tant, el que s'estudia amb la corba de ROC és la sensibilitat en funció de l'especificitat i pren valors entre 0 i 1, on un valor de 1 o pròxim indicarà un model perfecte, i un valor inferior indicarà que no és el millor model.

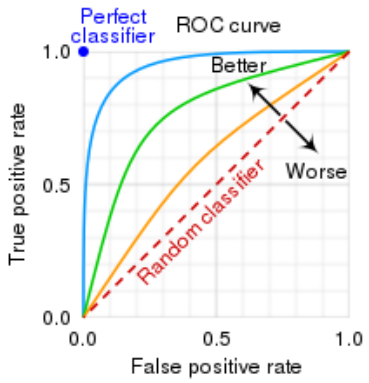

Figura III.3: Corba de ROC

Font: Wikipedia

IV. ESTUDI DE LES DADES

Per dur a terme l'estudi s'ha treballat amb una selecció de variables que provenen de 6 bases de dades extretes de *Open Bank Barcelona* , una font de dades obertes. I s'han analitzat mitjançant diversos recursos d'anàlisis de dades com la visualització gràfica, models predictius i mètodes d'avaluació dels models.

Per crear la base de dades s'ha fet servir el llenguatge SQL de la llibreria "*sqldf"* de RStudio.

I. Estructura de les dades

A continuació, trobem les bases de dades originals sobre les que hem treballat:

Nom de la taula	Descripció de la taula
Accidents	Ofereix un llistat d'accidents gestionats per la Guàrdia Urbana, el nombre de lesionats, la gravetat de l'accident, el nombre de vehicles implicats i el lloc de l'impacte
Accidents segons causa del conductor	Indica el lloc on ha passat, els motius i causes de l'accident i el moment en que ha ocorregut
Accidents segons causa mediata	Indica la causa mediata en temps, lloc i grau
Persones involucrades als accidents	Ofereix la informació sobre les persones que han estat involucrades, el tipus de lesió i grau, perfil de persona
Accidents segons la tipologia	Ens indica com ha estat l'impacte (xoc contra element, col·lisió lateral...)
Vehicles implicats en els accidents	Vehicles implicats i qui ha estat el causant de l'accident

Taula IV.1: Descripció de les bases de dades originals

Com hi havia un gran nombre de variables, es va triar una selecció i es van transformar algunes per tal de fer un anàlisi més precís, basant-nos en un estudi realitzat anteriorment pel tutor del treball[2] aconseguint la següent base de dades:

[2] L. Bermúdez and I. Morillo, Assessing the effectiveness of road safety measures in Barcelona (2013-2018), Heliyon

Variable	Descripció i nivells
Gravetat de l'accident	
Severity	Grau de gravetat de l'accident. (1: Accidents fatals o greus; 0: Accidents lleus)
Moment de l'accident	
Any	Any en que ha ocorregut l'accident (2020,2021,2022,2023)
Dia	Tipus de dia en que ha ocorregut l'accident (Laborable, Weekend)
Time	Part del dia en que ha ocorregut l'accident (Matí, Tarda o Nit)
Lloc de l'accident	
Tipus_via	Tipus de via on ha ocorregut l'accident (Carrer, Avinguda o Fast Lane)
Característiques dels vehicles implicats	
Vehicles	1: Si hi ha 1 o més vehicles implicats; 0: no hi ha vehicles implicats
vehicle_bici	1: presència de bicicletes; 0: no presència
vehicle_2rodes	1: presència de vehicles de 2 rodes ; 0: no presència
vehicle_light	1: presència de vehicle *light* ; 0 no presència
vehicle_heavy	1: presència de vehicle *heavy*; 0: no presència
Causes de l'accident	
causa_beguda	1: presència d'alcohol ; 0: no presència
causa_carretera	1: presència de la via en mal estat ; 0: no presència
causa_velocitat	1: presència de velocitat inadequada; no presència
causa_vianant	1: culpa del vianant; 0: no és culpa d'un vianant
Tipus d'accident	
accident_atropellament	1: atropellament; 0: una altre tipus d'accident
accident_caiguda	1: caiguda ; 0: una altre tipus d'accident
accident_colisio	1: col·lisió ; 0: una altre tipus d'accident
accident_xoc	1:xoc ; 0: una altre tipus d'accident

Taula IV.2: Descripció de les variables

II. Anàlisis descriptiva de les dades

Com s'ha vist a l'apartat anterior, totes les variables de la base de dades, són variables categories. Per tenir una comprensió completa i detallada de la base de dades a estudiar, es durà a terme un anàlisi descriptiu de les dades.

A continuació, es pot observar un resum de les variables de la base de dades:

La variable resposta, Severity, ens mostra una gran quantitat d'accidents lleus respecte els fatals/greus.

Severity	Freqüència
0	28.900
1	724

Taula IV.3: Summary de la variable Severity

Pel que fa referència al moment de l'accident individual i respecte la variable resposta, tenim les següents dades:

		Freqüència	Accident fatal/greu	Accidents lleus
Dia				
	Laborable	23.455	568	22.887
	Cap de setmana	6.169	156	6.013
Time				
	Matí	8.705	8.538	167
	Tarda	18.003	17.562	441
	Nit	2.916	2.800	116
Tipus de via				
	Carrer	12.979	13.651	307
	Avinguda	13.958	12.637	342
	Fast Lane	2.687	2.612	75

Taula IV.4: Summary de les variables de moment i localització de l'accident

Com podem observar, en la taula anterior, la majoria dels accidents es produeixen en dies laborables, generalment a la tarda, això ens pot indicar, que en el moment en que podria haver-hi més accidents, és quan la gent surt de la feina. Però en proporció, quan més greus són les lesions són a la nit seguit de la tarda.

Pel que fa als vehicles implicats, es mostra a continuació el anàlisis descriptiu:

		Freqüència	Accident fatal/greu	Accidents lleus
Vehicles				
	Presència	29.604	724	28.880
	No presència	20	0	20
Bicicleta				
	Presència	2.733	62	2.671
	No presència	26.891	662	2.622
2 rodes				
	Presència	13.342	375	12.965
	No presència	16.282	349	15.933
Light				
	Presència	9.586	164	9.422
	No presència	20.038	560	19.478
Heavy				
	Presència	2.804	91	2.713
	No presència	26.187	633	26.187

Taula IV.5: Summary de les variables sobre els vehicles implicats en els accidents

El tipus de vehicle implicat que més accidents presenta, són els vehicles de 2 rodes seguit dels vehicles *light* (turismes, taxis o furgonetes) i, també un nombre major de morts i lesions greus.

En tercer lloc, es mostra la taula de freqüències segons la causa de l'accident:

		Freqüència	Accident fatal/greu	Accidents lleus
Vianant				
	Presència	7.712	235	7.477
	No presència	21.912	489	21.423
Beguda				
	Presència	1.665	42	1.623
	No presència	27.959	682	27.277
Carretera				
	Presència	334	4	330
	No presència	29.290	720	28.570
Velocitat				
	Presència	148	28	120
	No presència	29.476	696	28.780

Taula IV.6: Summary de les variables sobre les causes de l'accident

Les morts i lesions més greus són més freqüents quan l'accident és causat pels vianants. Aquest factor és el de major risc ja que és el que major nombre de morts i lesions greus presenta, seguit de un consum d'alcohol excessiu i una velocitat inadequada.

Per acabar, es mostra la informació segons el tipus d'accident:

		Freqüència	Accident fatal/greu	Accidents lleus
Atropellament				
	Presència	3.436	199	3.237
	No presència	26.188	525	25.663
Caiguda				
	Presència	3.629	97	3.532
	No presència	25.995	627	25.368
Col·lisió				
	Presència	18.411	380	19.031
	No presència	10.213	344	9.869
Xoc				
	Presència	2.336	37	2.299
	No presència	27.288	687	26.601

Taula IV.7: Summary de les variables sobre la tipologia d'accident

Veiem que la tipologia d'accident més freqüent és la col·lisió i, també la que més accidents mortals o amb lesions greus provoca.

A continuació, es modelaran les dades amb diferents tècniques per veure quins factors contribueixen en major o menor mesura en el grau de severitat de l'accident.

V. MODELATGE DE LES DADES

Com hem comentat anteriorment, s'utilitza la variable *Severity* com a variable d'interès per dur a terme l'anàlisi. Aquesta variable representa la gravetat dels accidents i la prenem com a variable dependent.

Per tal de partir des del mateix punt d'inici per dur a terme els models de predicció, hem inclòs tota la resta de variables detallades a l'apartat anterior com a variables explicatives que consisteixen en tots els factors que poden influir en el grau de gravetat de l'accident.

Abans de començar amb la modelització hem creat un conjunt de dades d'entrenament i un altre de test, amb un 80% i 20% del conjunt total de les dades, garantint que les dades tinguin una distribució semblant de la variable *Severity*.

Al dur a terme l'anàlisi descriptiu de les variables, s'ha vist, que la variable resposta, està desbalancejada, ja que el percentatge d'accidents severs és molt inferior al d'accidents lleus. Per tal de reduir les diferencies entre les dues classes, s'han dut a terme tècniques de remostreig, en aquest cas, *over i undersampling*.

Per decidir quina tècnica de remostreig és millor, s'han provat les dues esmentades anteriorment amb el model de classificació logística i, el que millor resultats ha donat, ha estat el triat per a entrenar la resta de models.

Primerament, s'ha dut a terme amb el *undersampling* que consisteix en modificar la distribució de les dades reduint el nombre de casos de la classe majoritària, en l'estudi el nombre d'accidents lleus.

$$Under <- trainControl(method = "repeatedcv", number = 10, repeats = 5, \quad savePredictions = "final", \\ returnResamp = "final", \quad classProbs = TRUE, \\ seed = set.seed(423), summaryFunction = twoClassSummary, \\ sampling = "down")$$

Fórmula 1: Undersampling

Tot seguit, també s'ha fet l'entrenament del model, mitjançant el oversampling que té com a finalitat modificar la distribució de les dades, augmentant el nombre de casos de la classe majoritària, és a dir, augmentant el nombre de casos d'accidents severs.

$$Over <- trainControl(method = "repeatedcv", number = 10,$$
$$repeats = 5, \quad savePredictions = "final",$$
$$returnResamp = "final", classProbs = TRUE,$$
$$seed = set.seed(423),$$
$$summaryFunction = twoClassSummary, \quad sampling = "up")$$

Fórmula 2: Oversampling

Un cop entrenats els dos models amb els mètodes de balanceig corresponents, s'han calculat mètriques de rendiment del model amb les dades d'entrenament, s'ha calculat la mitja de les mètriques i s'ha representat mitjançant un *boxplot*:

	ROC	Sensibilitat	Especificitat
Undersampling	0,6763	0,6569	0,5789
Oversampling	0,6820	0,6746	0,5575

Taula V.1: Resultats del model de classificació logística amb els mètodes de balanceig

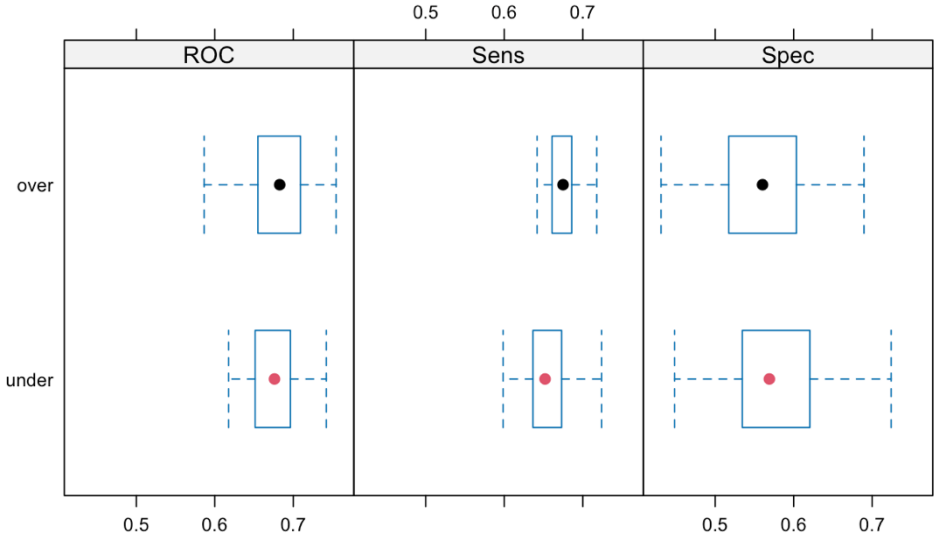

Figura V.1: Resultat de les mètriques dels mètodes de balanceig

Com s'observa, no hi ha grans diferències entre les mètriques dels mètodes, així doncs, triarem el mètode *undersampling* ja que tenen un cost computacional menor.

I. Model de classificació logística

Primerament, apliquem la funció *train* de la llibreria *caret* amb els següents paràmetres:

$$mdl\_\log <- caret :: train(Severity \sim ., data = data\_train, \quad method = \text{glm}, family = 'binomial', \quad trControl = Under)$$

Fórmula 3: Model utilitzat pel model de classificació logística

El primer paràmetre consisteix en el model que es vol ajustar, la relació entre les variables predictores i la variable resposta. Tot seguit, proporcionem les dades amb les que entrenarà el model, el conjunt de dades d'entrenament que hem creat anteriorment.

Pel que fa al mètode, hem especificat el *glm* ja que el model de regressió logística, és un tipus específic de Model Lineal Generalitzat i, li hem especificat la distribució de la variable resposta, en aquest cas, Binomial ja que la variable resposta és una variable binària i, en conseqüència, es farà servir el la funció d'enllaç *logit* per ajustar la classificació logística.

Finalment, hem especificat com s'ha de dur a terme l'entrenament i la validació del model, incloent el nombre de repeticions, el *undersampling*, entre d'altres.

II. Arbres de decisió

De la mateixa manera que en el model anterior, apliquem la funció *train* de la llibreria *caret* amb els següents paràmetres:

$$\boldsymbol{mdl\_cart <- caret::train(Severity \sim ., data = data_{train}, method = "rpart", tuneGrid = expand.grid(cp = seq(from = 0.0001, to = 0.1, length = 15)), \quad metric = "ROC", trControl = Under)}$$

Fórmula 4: Model utilitzat pels arbres de decisió

El model a ajustar i les dades que es prenen, són iguals que pel model de classificació logística. El mètode que s'usa és el "*rpart*" que consisteix en l'algoritme CART de R, és a dir, s'utilitzarà un model d'arbres de decisió. També, hem triat que la mètrica d'avaluació sigui ROC.

El *tuneGrid* especifica la seqüència de valors que provarà el paràmetre de complexitat (cp) durant el procés d'ajust. En aquest estudi, s'ha marcat una seqüència de 15 valors entre 0'0001 fins a 0'1.

El control d'entrenament, s'utilitza la tècnica de *undersampling* que ha estat la triada, a l'inici per tal de que tots els models utilitzessin el mateix mètode de balanceig.

A més dels paràmetres descrits anteriorment, és important tenir en compte la profunditat de l'arbre. Un arbre de decisió massa profund pot definir molt bé les característiques del model d'entrenament però no tindrà un bon rendiment amb unes altres dades, és a dir, es produirà sobreajustament. Per contra, un arbre de decisió massa superficial pot no capturar la suficient informació i aconseguir de la mateixa manera un baix rendiment.

Per moderar el risc de sobreajustament, s'utilitza el valor *cp* (paràmetre de complexitat) on per reduir el risc, hem d'obtenir valors més alts, ja que d'aquesta manera, tendeix a construir arbres més simples i menys propensos a sobreajustar-se. En el cas d'aquest anàlisi, ens ha sortit un paràmetre de complexitat de 0,0072, aquest valor ens farà maximitzar l'Àrea Sota la Corba de ROC.

III. Random Forest

Per dur a terme la modelització mitjançant el Random Forest, es fa servir la funció *train* del paquet *caret* amb els següents paràmetres:

$$mdl\_rf <- caret::train(Severity \sim ., data = data\_train, \quad method = \text{rf}, \quad metric = "ROC", \quad tuneGrid = expand.grid(mtry = 1:4), \quad trControl = Under)$$

Fórmula 5: Model utilitzat pel Random Forest

Per dur a terme aquest model, s'ha fet servir el mateix que a l'apartat anterior, canviant el mètode, que abans eren arbres de decisió i ara és *Random Forest*. I, també s'ha canviat el nombre de variables explicatives a considerar en cada divisió de l'arbre de decisió. En aquest cas, per *mtry* es proven valors entre 1 i 4.

Per interpretar aquest model, hem fet servir el paquet *iml*, ja que amb la llibreria *caret* no es podien identificar correctament els factors de risc en els accidents de trànsit.

IV. XGBoost

El model per dur a terme el model XGBoost amb la llibreria *caret* ha estat plantejada de la següent manera:

$$mdl\_xgb <- caret::train(Severity \sim ., data = data\_train, method = "xgbTree", metric = "ROC", tuneGrid = tuneGrid, trControl = Under)$$

Fórmula 6: Model utilitzat pel XGBoost

Per dur a terme aquest model, s'ha fet servir la funció train per la variable resposta Severity i totes les variables predictores, usant les dades d'entrenament i el mètode del XGBoost i la mètrica de ROC. Però per executar-ho, primerament, s'ha hagut de plantejar el tunegGrid que defineix els valors específics dels hiperparametres que seran provats durant el procés de cercar el millor model.

VI. RESULTATS DELS MODELS

I. Model de classificació logística

A continuació, es mostren informació sobre els factors que influeixen en la severitat dels accidents de trànsit mitjançant un model de regressió logística:

	Estimació	Std. Error	Pr(>|z|)
(Intercept)	-716,16	157,97	5,80E-06
Any	0,35	0,08	5,91E-06
DiaWeekend	-0,03	0,16	8,25E-06
MomentNit	1,23	0,24	1,94E-01
MomentTarda	0,25	0,15	9,29E-02
vehicle_bici1	-0,17	0,41	6,80E-01
vehicle_2rodes1	0,43	0,36	2,35E-01
vehicle_light1	-0,57	0,37	1,21E-01
vehicle_heavy1	0,24	0,39	5,33E-01
causa_beguda1	-0,24	0,28	3,95E-01
causa_carretera1	-1,33	0,74	7,30E-02
causa_velocitat1	2,28	0,76	2,62E-03
causa_vianant1	0,63	0,20	1,93E-03
accident_atropellament1	1,30	0,51	1,14E-02
accident_caiguda1	0,19	0,52	7,15E-01
accident_colisio1	0,12	0,50	8,14E-01
accident_xoc1	-0,29	0,55	6,02E-01
Tipus_viaCarrer	-0,02	0,14	9,00E-01
`Tipus_viaFast lane`	0,10	0,22	6,40E-01

Taula VI.1: Resultat del model de classificació logística

Com s'observa, els factors més significatius per predir la severitat són aquells que tenen un p-valor més baix, és a dir, un valor més significatiu i són:

L'any en que es produeix l'accident, ja que cada increment d'un any està relacionat amb la probabilitat de la severitat dels accidents. Conduir els caps de setmana i a la tarda , està associat amb accidents més severs. El fet de que siguin més severs els accidents a la tarda, podria estar relacionat amb la hora de sortida dels llocs de treball, dels infants de l'escola i, per tant una major concentració de trànsit superior pel que provoca una fatiga en els conductors.

Les causes que són significativament més probables de produir un accident sever, són les que fan referencia a una velocitat inadequada i les que són causades per un vianant. Això hauria de portar a dur a terme més controls de velocitat i fent pensar que cal una millora dels coneixement sobre seguretat vial a la població. Els atropellaments, també estan associats a una major severitat dels accidents.

II. Arbres de decisió

Pel mètode dels arbres de decisió, s'aconsegueix el següent arbre:

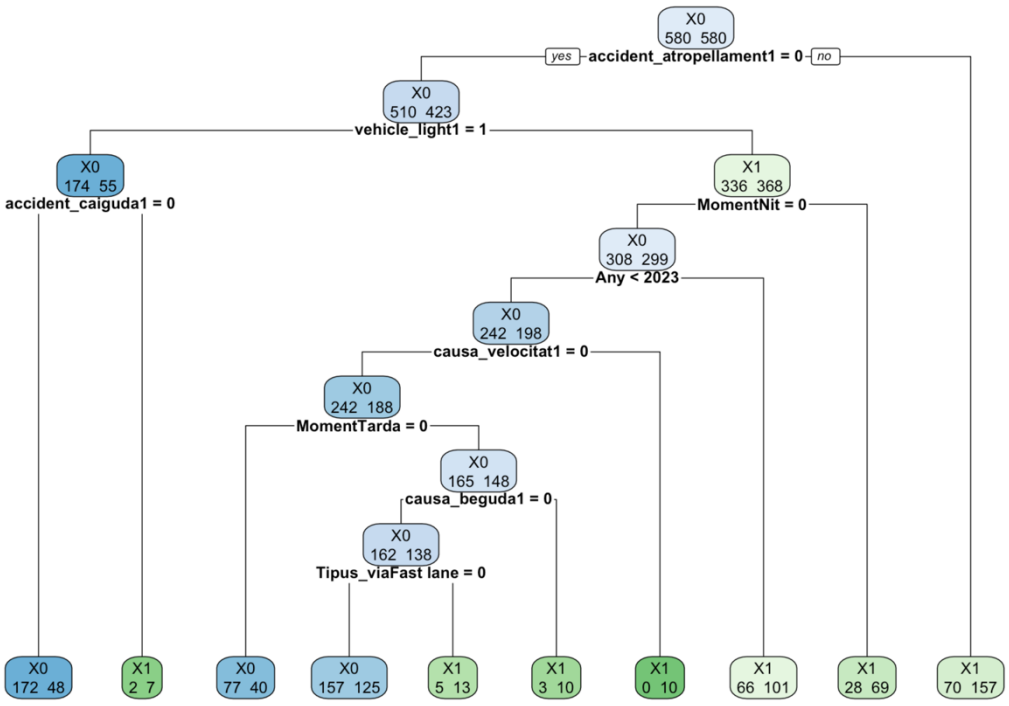

Figura VI.1: Resultat del model d'arbres de decisió

Mitjançant aquest arbre, es pot entendre de manera visual com les variables influeixen en la predicció del nivell de gravetat d'un accident. A continuació, es proporciona una explicació detallada:

A simple vista, veiem que la variable més influent en la severitat dels accidents és que sigui un atropellament. D'altra banda, el fet que sigui un vehicle *light* (turisme,

furgoneta...), que l'accident hagi estat una caiguda, que hagi sigut durant la nit, abans del 2023, causat per una velocitat inadequada, entre d'altres factors, són altres dels factors més significatius per predir la severitat dels accidents.

III. Random Forest

Per interpretar aquest model, hem fet servir el paquet iml, ja que amb la llibreria caret no es podien identificar correctament els factors de risc en els accidents de trànsit.

Així doncs, mitjançant la funció FeatureImp s'ha obtingut una representació de la influencia de les variable en la severitat dels accidents:

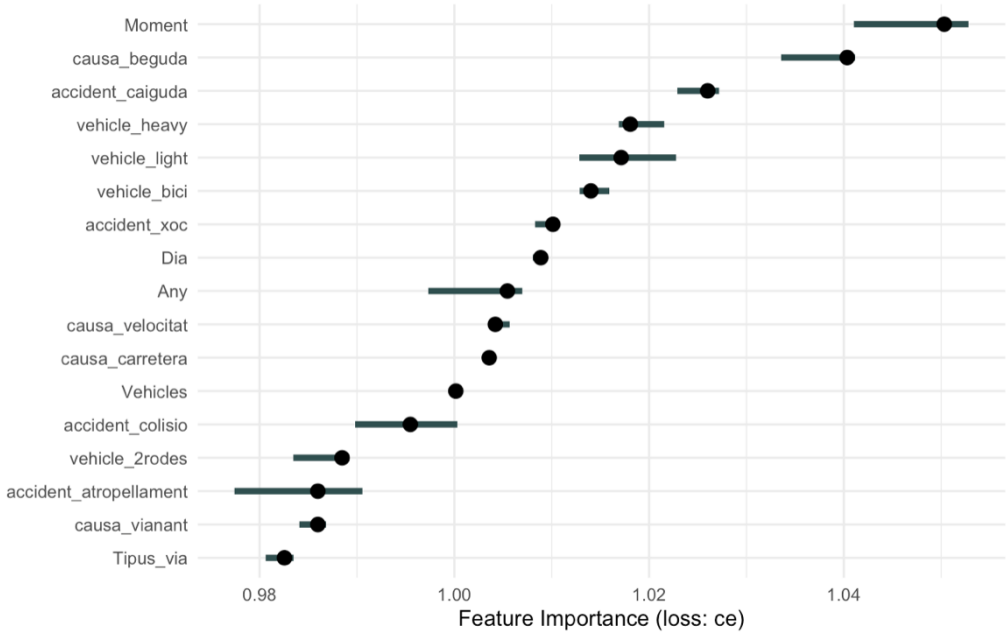

Figura VI.2: Gràfic de la influència de les variable en el model Random Forest

Sembla, que en aquest model, la variable més important és el moment del dia, indicant doncs, és a dir, que en funció de si l'accident passa al matí, a la tarda o a la nit, l'accident té més o menys probabilitats de ser sever.

Tot seguit, s'observa que el fet de que la causa sigui un consum de begudes alcohòliques inadequat és un predictor rellevant per determinar la severitat de l'accident.

Les dues variables comentades, són les més rellevants, tot i que n'hi ha d'altre com que hagi estat una caiguda, hi hagi un vehicle pesat, lleuger o una bici implicades a l'accident o que hagi estat un xoc, també són rellevants per predir si l'accident tindrà lesions greus o mortals o si les lesions seran lleugeres.

Una altra manera de fer els models més interpretables és substituir el model de “caixa negra” per un model més simple, en aquest cas, l’arbre de decisió. Es fa servir la funció *TreeSurrogate* que emula el model d’aprenentatge automàtic que s’ha creat a partir d’un arbre de decisió simple.

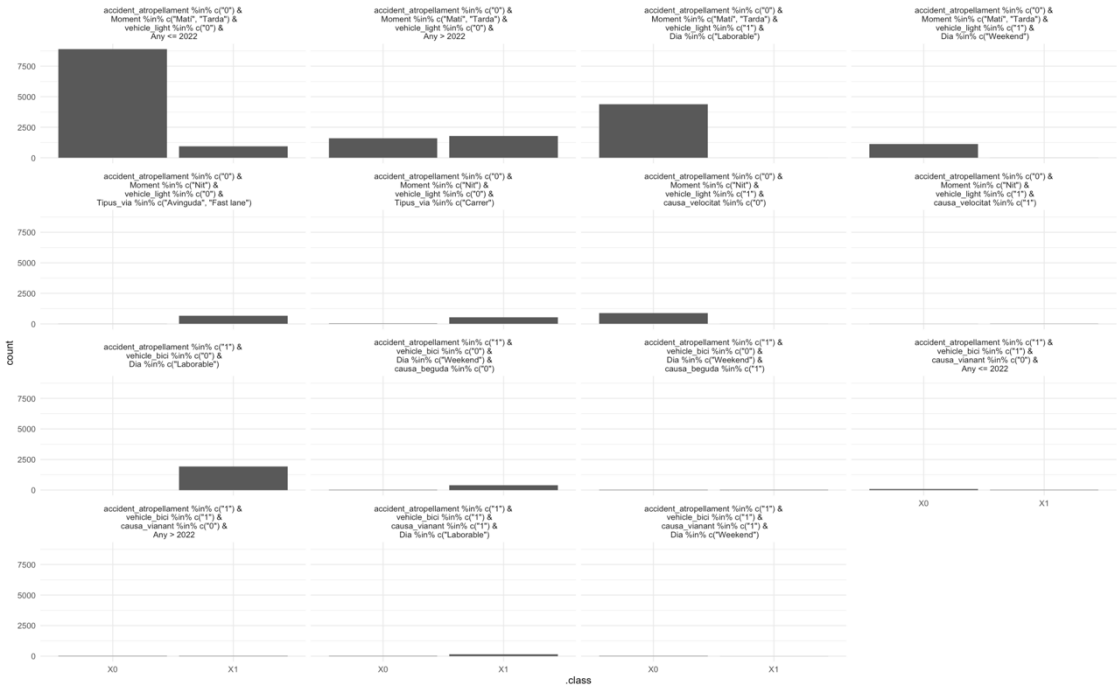

Figura VI.3: Classificació de les observacions de l'arbre de decisió - Random Forest

Tot i això, mirant el coeficient del R2 que és 0,613 ens mostra que té un ajust moderat, però no ajusta de manera adient el Random Forest.

Finalment, es calculen els valors de Shapley per poder realitzar una interpretació local d’una predicció concreta. Per fer-ho, s’escullen dues observacions de la mostra de test que hagin estat classificades correctament, per exemple, l’accident que té major probabilitat de ser sever i el que té menor probabilitat.

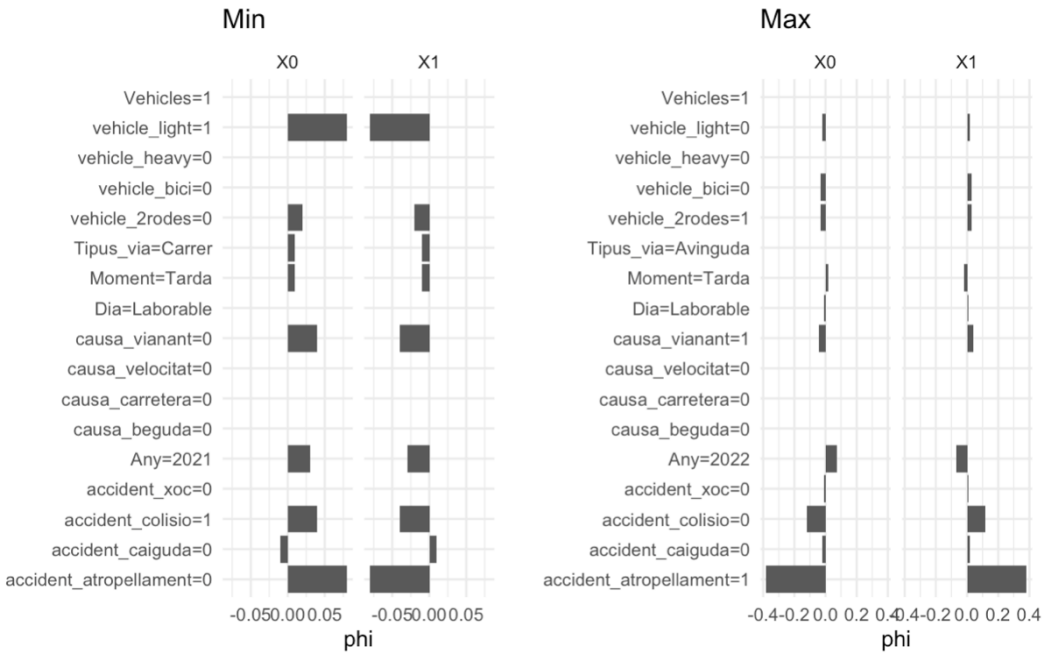

Figura VI.4: Valors de Shapley del Random Forest

En aquestes dues observacions, com es pot observar, la variable que determina en major grau la predicció en ambdós casos és quan l'accident és un atropellament.

Pels casos amb menor probabilitat de ser definits com accidents severs són també els que hi ha vehicles lleugers implicats en l'accident, quan la causa es per un vianants o l'accident és una col·lisió.

Mínim	Primer Quartil	Mediana	Mitjana	Tercer Quartil	Màxim
0,0000	0,1460	0,2460	0,2472	0,3600	0,9840

Taula VI.2: Valors de Shapley del Random Forest

Els resultats de Shapley indiquen que hi ha una variabilitat significativa amb algunes característiques tenint una contribució nul·la i altres una contribució molt alta. A més, la mediana i la mitjana són força properes suggereixen una distribució balancejada de contribucions de les característiques.

Les característiques amb valors propers al màxim són els més influents i podrien ser les més importants pel model.

Així doncs, tot i no ser la variable més important, la variable que determina el valors de la predicció en major grau és quan l'accident és un atropellament.

IV. XGBoost

Pel mètode del XGBoost, hem dut a terme el mateix procediment que pel Random Forest, degut a que entrenant el model, amb les sortides d'aquest, era complicat treure conclusions de quins factors afecten en major o menor mesura en la severitat dels accidents. Per aquest motiu, hem fet servir una llibreria xAI com és la *iml* que ens ha permès projectar en un gràfic les variables indicant la importància que tenen en la predicció de la severitat. El gràfic es mostra a continuació:

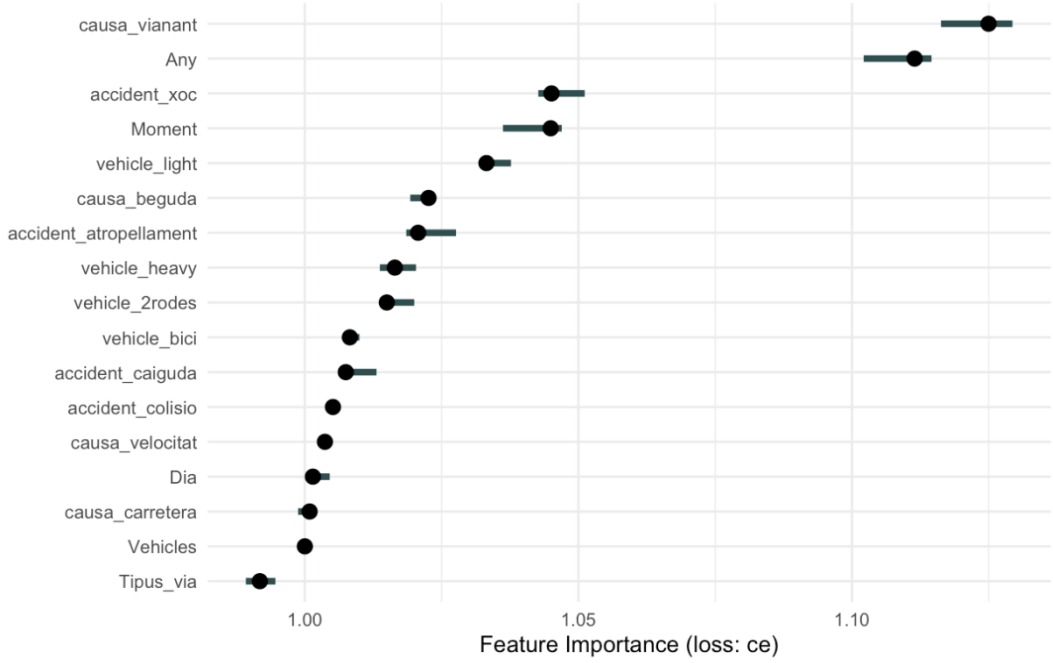

Figura VI.5: Gràfic influència de les variables amb el model XGBoost

Un comportament inadequat d'algun vianant o l'any en que s'ha produït l'accident, són els factors més significatius per predir com de greu serà un accident.

Altres variables que també són rellevants per poder intuir com de sever serà l'accident, serà si l'accident és o no un xoc, el moment de l'accident o si hi ha implicat un vehicle lleuger.

Els baixos errors de permutació reforcen la confiança en aquestes associacions i, es mostren en la següent taula:

	Error de permutació
causa_vianant	0,3232
any	0,3293
accident_xoc	0,3003
moment	0,3002
vehicle_light	0,2968
causa_beguda	0,2939

Taula VI.3: Resultat dels errors de permutació del model XGBoost

VII. AVALUACIÓ DELS MODELS

Per avaluar els models estimats, hem fet servir diferents mètriques, per tal de triar el millor model.

A partir, de la matriu de confusió hem obtingut el *Accuracy* que ens indica el grau d'exactitud, és a dir, ens diu el percentatge de prediccions correctes entre el total de prediccions. També, hem calculat el Coeficient Kappa de Cohen que mesura el grau de coincidència entre les prediccions del model i les observacions reals, ajustat per la coincidència que es podria donar per atzar.

D'altra banda, també hem estimat la corba de ROC per cada model i, l'àrea sota la corba (AUC) per tal de mesura la capacitat del model de distingir entre classes. L'objectiu, seria que el valor fos el més proper a 1 que ens indica que té una capacitat total per distingir les classes positives de les negatives.

Finalment, també s'ha tingut en compte la sensibilitat i especificitat de cada model. La sensibilitat indicarà quina és la capacitat dels models entrenats de detectar els accidents severs correctament. En canvi, la especificitat, ens indicarà la capacitat de detectar accidents severs que realment són accidents amb malalties lleus.

A continuació, es mostra la taula resultant amb els tres coeficients:

	Accuracy	Sensibilitat	Especificitat	Àrea sota la corba de ROC	Kappa
Model Logístic	0,6632	0,6667	0,5208	0,6435	0,0257
Arbres de decisió	0,8098	0,8206	0,3750	0,6134	0,0465
Random Forest	0,7221	0,7276	0,5000	0,6396	0,0374
XGBoost	0,7009	0,7055	0,5138	0,6461	0,0336

Taula VII.1: Summary de les mètriques d'avaluació dels models

El model de classificació logística mostra una capacitat moderada en totes les mètriques que hem tingut en compte. Tenim un *accuracy* del 66% aproximadament, és a dir, que encerta aproximadament dos de cada tres prediccions. La sensibilitat és bona, encerta el 66,6% dels accidents severs, en canvi, l'especificitat, és més baixa, pel que no distingeix del tot bé els accidents lleus reals. Pel que fa al AUC ens suggereix que té un rendiment moderat, que es podria millorar i el coeficient de Kappa ens indica que el model és millor del que s'esperaria per atzar ja que és lleugerament superior a 0.

Els arbres de decisió, destaquen pel seu *Accuracy* ja que és d'un 80% i, per tant, encerta una gran part de les seves prediccions. També té la major sensibilitat, per tant detecta correctament els accidents severs com a severs però, té una especificitat molt baixa, per

tant, els accidents lleus en ocasions els classifica com a severs. L'àrea sota la corba de ROC i el coeficient de Kappa, són similars als del model logístic.

En tercer lloc, hem entrenat el model *Random Forest* que té una bona precisió (72%) , la capacitat de detectar un accident greu com a greu és bo i, distingeix moderadament els accidents lleus reals. El valor del AUC és lleugerament inferior que al del model logístic, però és millor que el dels arbres de decisió, tenim la mateixa situació pel valor del Kappa.

El últim model entrenat és el XGBoost que té uns resultats similars als del *Random Forest*, lleugerament inferiors als d'aquest últim, excepte en l'especificitat que és 0,01 superior, el mateix que per l'Àrea Sota la Corba de ROC.

Un cop analitzats tots els aspectes, es pot veure que el millor model general per predir la severitat dels accidents de trànsit a Barcelona és el d'arbres de decisió. Aquest model, té un bon ajust i sensibilitat, però té una baixa especificitat que fa que detecti algun dels accidents lleus com a severs.

Aquest model, permetrà prendre decisions ràpidament i entendre quines variables són més rellevant i de major o menor risc, ja que de manera visual, podrem decidir si l'accident és sever o no.

Si el nostre objectiu fos obtenir una precisió màxima, hauriem de tenir en compte el *Random Forest* degut a que la capacitat de crear models més complexos i amb grans quantitats de dades fan que la opció sigui òptima respecte als arbres de decisió que poden tenir problemes de sobreajustament o profunditat de l'arbre.

VIII. CONCLUSIONS

Aquest estudi, té com a objectiu analitzar els factors que influeixen en els accidents de trànsit per determinar el seu grau de severitat. Per dur-ho a terme, s'han usat les dades recollides per la Guàrdia Urbana en el període comprés entre 2020 i 2023 a la ciutat de Barcelona.

La recerca, ha consistit en trobar la relació entre el grau de gravetat de l'accident, si era un accident mortal o amb lesions greus o un accident amb lesions lleus, amb els diferents factors de risc que influeixen en aquest. S'han tingut en compte, factors de moment i localització, dels vehicles implicats, les causes de l'accident i el tipus d'accident.

Per realitzar l'anàlisi , s'han dut a terme prediccions amb diferents tècniques, per fer-ho, s'han dividit les dades en dos grups. El 80% d'aquestes s'han utilitzar per l'entrenament dels models i el 20% restant s'han utilitzat per testejar. Una vegada, dividides les dades i comprovant que seguien una mateixa distribució, s'han implementat els quatre models: de classificació logística, arbres de decisió, *random forest*, *XGBoost.* Un cop entrenats els models i testejats, s'ha realitzat una comparativa entre els resultats que s'han obtingut.

El model de classificació logística, mostra un rendiment mig en totes les mètriques avaluades . Les variables que augmenten la probabilitat de que l'accident sigui sever són: quan incrementa l'any hi ha més probabilitats, que sigui cap de setmana, per la tarda, a causa d'una velocitat inadequada o provocat per un vianant.

Els arbres de decisió, en termes generals és el millor model però té una especificitat baixa. L'escenari que tenim per a que un accident sigui mortal o amb lesions greus, és que hagi estat un atropellament, i en el cas que no sigui així que hi hagi algun vehicle lleuger i hagi estat durant la nit i l'any inferior al 2023 a causa d'una velocitat inadequada.

Respecte el *random forest*, és un model amb un bon rendiment, té una bona especificitat, que l'arbre de decisió no tenia. L'escenari que presenta és el següent: depèn del moment del dia, que hagi estat causat per un consum d'alcohol, que l'accident hagi estat una caiguda on hi hagi implicats vehicles pesats, lleugers i/o bicicletes.

L'últim model que s'ha implementat, ha estat el *XGBoost*, que té un bon rendiment, però és el més lent i complicat de compilar. L'escenari que presenta amb la probabilitat més alta per a que sigui un accident sever és: que hagi estat causat per algun vianant,

depèn de l'any i el moment del dia, que hagi estat un xoc, i hi hagi algun vehicle *light* implicat.

Un cop realitzat l'estudi, el millor model general en quant a mètriques són els arbres de decisió però, si augmentes considerablement la dimensió de les dades o hi hagués algun canvi en les dades, podria suposar un sobreajustament, una profundiata de l'arbre massa gran que ens impossibilites la interpretació d'aquest.

Així doncs, s'ha seleccionat el mètode que millor prediu el grau de severitat d'un accident de trànsit a la ciutat de Barcelona i, el model triat ha estat el *random forest* per què tot i no tenir les millors mètriques, té un bon rendiment i una capacitat superior als arbres de decisió en quant a dimensió i complexitat de les dades.

Pel que fa a les variables més importants per definir el grau de severitat dels accidents amb el model triat, veiem que les dues variables més importants són moment i causa_beguda, aquestes dues, podrien disminuir amb un augment de control en el moment del dia on es produeixen la majoria dels accidents severs i, augmentant els controls d'alcoholemia per evitar que es condueixi en un estat d'embriaguesa.

Altres variables importants com el tipus de vehicles implicats o si l'accident és una caiguda són importants però no es poden aplicar mesures preventives per evitar aquest tipus d'accidents. D'altra banda, hi ha variables que inicialment podríem pensar que serien significatives per predir la severitat de l'accident com el tipus de via, finalment no han estat primordials per dir el grau de gravetat de l'accident.

Per concloure, amb la predicció duta a terme amb el *Random Forest* s'haurien d'afegir un control més exhaustiu de la circulació durant les hores de la tarda, que estaria relacionada amb la sortida de l'escola i la feina i moments de cansament i angoixa en els conductors. D'altra banda, també s'haurien d'incrementar els controls d'alcoholèmia per reduir els accidents causats per un mal consum de begudes alcohòliques del/s implicat/s en l'accident de trànsit. Tot això, haurà d'anar acompanyat d'una bona educació en seguretat vial en la població.

IX. BIBLIOGRAFIA

accidents – Infotrànsit. (s/f). Infotrànsit. Recuperado el 20 de junio de 2024, de https://infotransit.blog.gencat.cat/tag/accidents/

Bermúdez, L., & Morillo, I. (2023). Assessing the effectiveness of road safety measures in Barcelona (2013-2018). *Heliyon*, *9*(12), e23063. https://doi.org/10.1016/j.heliyon.2023.e23063

Blanco, P. (2024, abril 17). Los algoritmos random forest, una técnica muy potente de machine learning. *Educaopen.com*. https://www.educaopen.com/digital-lab/blog/inteligencia-artificial/algoritmo-random-forest

De seguretat, N. (2024, marzo 6). *Informe de la situació mundial de la seguretat viària 2023*. Notes de seguretat. https://notesdeseguretat.blog.gencat.cat/2024/03/06/informe-de-la-situacio-mundial-de-la-seguretat-viaria-2023/

Ferrero, R. (s/f). *Qué son los árboles de decisión y para qué sirven*. Máxima Formación. Recuperado el 20 de junio de 2024, de https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

Gradient Boosting. (2023, octubre 31). FOQUM; Foqum Analytics | Empowers your success. https://foqum.io/blog/termino/gradient-boosting/

Open Data Barcelona: Accidents de trànsit regristrats per la Guàrdia Urbana. (s/f). Barcelona.cat. Recuperado el 20 de junio de 2024, de https://opendata-ajuntament.barcelona.cat/data/ca/dataset?q=accidents+&sort=fecha_publicacion+desc

profesorDATA. (2020, agosto 7). *Evaluando los modelos de Clasificación en Aprendizaje Automático: La matriz de confusión*. profesordata.com. https://profesordata.com/2020/08/07/evaluando-los-modelos-de-clasificacion-en-aprendizaje-automatico-la-matriz-de-confusion-claramente-explicada/

¿Qué es un árbol de decisión? (2022, octubre 20). Ibm.com. https://www.ibm.com/es-es/topics/decision-trees

Wikipedia contributors. (s/f). *Regresión logística*. Wikipedia, The Free Encyclopedia. https://es.wikipedia.org/w/index.php?title=Regresi%C3%B3n_log%C3%ADstica&oldid=156179421

(S/f-a). Uma.es. Recuperado el 20 de junio de 2024, de https://www.bioestadistica.uma.es/apuntesMaster/regresión-log%C3%ADstica-binaria.html

(S/f-b). Amazon.com. Recuperado el 20 de junio de 2024, de https://aws.amazon.com/es/what-is/logistic-regression/

X. ANNEX. SCRIPTS D'R

```
# Llibreries:
library(sqldf)
library(tidyverse)
library(ggplot2)
library(corrr)
library(knitr)
library(xtable)
library(ISLR)
library(skimr)
library(ggpubr)
library(caret)
library(yardstick)
library(flextable)
library(plotROC)
library(iml)
library(rpart)
library(rpart.plot)
library(ROCR)

# Creem la base de dades:
data1 <- sqldf('SELECT  Nom_carrer, NK_Any as Any,
case when Descripcio_dia_setmana in ("Dissabte", "Diumenge") then "Weekend" else "Laborable" end as Dia,
case when Hora_dia between 2 and 6 then "Nit"
   when Hora_dia between 7 and 12 then "Matí"
   when Hora_dia between 13 and 22 then "Tarda" else "Nit" end as Moment,
   case when Numero_morts > 0 or Numero_lesionats_greus > 0 then 1 else 0 end as Severity,
case when Numero_vehicles_implicats > 0 then 1 else 0 end as Vehicles,
case when Descripcio_tipus_vehicle in ("Bicicleta","Tricicle")  then 1 else 0 end  as vehicle_bici,
case when Descripcio_tipus_vehicle in ("Motocicleta" ,"Ciclomotor")  then 1 else 0 end as vehicle_2rodes,
case when Descripcio_tipus_vehicle in ("Turisme", "Taxi", "Furgoneta", "Camió rígid <= 3,5 tones", "Pick-up", "Tot terreny" , "Quadricicle < 75 cc" , "Quadricicle < 75 cc" , "Autocaravana" ,"Ambulància" , "Quadricicle > 75 cc"  )  then 1 else 0 end  as vehicle_light,
case when Descripcio_tipus_vehicle in ("Autobús", "Autobús articulat" ,"Tractor camió" , "Camió rígid > 3,5 tones"  , "Microbús <= 17"  , "Tren o tramvia", "Autocar"  ) then 1 else 0 end as vehicle_heavy,
CASE WHEN Descripcio_causa_mediata in ( "Drogues o medicaments" ,"Alcoholèmia") THEN 1 ELSE 0 END AS causa_beguda,
CASE WHEN Descripcio_causa_mediata in ("Calçada en mal estat" , "Objectes o animals a la calçada", "Estat de la senyalització") THEN 1 ELSE 0 END AS causa_carretera,
CASE WHEN Descripcio_causa_mediata in ("Excés de velocitat o inadequada") THEN 1 ELSE 0 END AS causa_velocitat,
```

```
case when Descripcio_causa_vianant not in ("No és causa del  vianant") then 1 else 0 end as causa_vianant,
case when Descripcio_tipus_accident in ("Atropellament" , "Sortida de via amb atropellament", "Bolcada (més de dues rodes)", "Sortida de via amb bolcada" ) then 1 else 0 end as accident_atropellament,
case when Descripcio_tipus_accident in ("Caiguda interior vehicle" ,"Caiguda (dues rodes)") then 1 else 0 end as accident_caiguda,
case when Descripcio_tipus_accident in ("Col.lisió lateral" ,"Col.lisió fronto-lateral" ,"Col.lisió frontal","Abast", "Abast multiple", "Sortida de via amb xoc o col.lisió","Encalç" ) then 1 else 0 end as accident_colisio,
case when Descripcio_tipus_accident in ("Xoc contra element estàtic","Xoc amb animal a la calçada" ) then 1 else 0 end as accident_xoc
from data
group by Numero_expedient')

carrers <- read_excel("/Users/gemmamascaro/Downloads/categorias.xlsx")
data <- sqldf('SELECT t1.*, t2.classificacio as Tipus_via
            FROM data1 t1
              LEFT JOIN carrers T2 ON t1.Nom_carrer=t2.Carrer'
)
data <- subset(data, select = -c(Nom_carrer))

table(data$Dia)
table(data$Time)
table(data$Tipus_via)
table(data$Severity, data$Dia)
table(data$Severity, data$Time)
table(data$Severity, data$Tipus_via)
table(data$Vehicles)
table(data$Severity, data$Vehicles)
table(data$vehicle_bici)
table(data$Severity, data$vehicle_bici)
table(data$vehicle_2rodes)
table(data$Severity, data$vehicle_2rodes)
table(data$vehicle_light)
table(data$Severity, data$vehicle_light)
table(data$vehicle_heavy)
table(data$Severity, data$vehicle_heavy)
table(data$causa_vianant)
table(data$Severity, data$causa_vianant)
table(data$causa_beguda)
table(data$Severity, data$causa_beguda)
table(data$causa_carretera)
table(data$Severity, data$causa_carretera)
...

# data train i test
```

```
partition <- createDataPartition(y = data1$Severity, p = 0.8,
                         list = FALSE)
data_train <- data1[partition, ]
data_test <- data1[-partition, ]

# Under i oversampling
Under <- trainControl(method = "repeatedcv", number = 10, repeats = 5,
                savePredictions = "final",
                returnResamp =  "final",
                classProbs = TRUE,
                seed = set.seed(423),
                summaryFunction = twoClassSummary,
                sampling = "down")


Over<-trainControl(method = "repeatedcv", number = 10, repeats = 5, savePredictions
= "final",returnResamp = "final", classProbs = TRUE,seed = set.seed(423),
summaryFunction = twoClassSummary,sampling = "up")

# Model de classificació logística
mdl_log_under <- caret::train(
  Severity ~ .,
  data = data_train,
  method = "glm",
  family = 'binomial',
  trControl = Under
)

mdl_log_over <- caret::train(
  Severity ~ .,
  data = data_train,
  method = "glm",
  family = 'binomial',
  trControl = Over
)

resamples_log <- resamples(list("under" = mdl_log_under, "over" = mdl_log_over))
summary(resamples_log)
resamples_log$models <- as.factor(resamples_log$models)
bwplot(resamples_log, col = resamples_log$models,
      groups = resamples_log$models)
mdl_log <- caret::train(
  Severity ~ .,
  data = data_train,
  method = "glm",
  family = 'binomial',
  trControl = Under
```

```
)

summary_table <- summary(mdl_log)
summary_df <- as.data.frame(summary_table$coefficients)

# Especificar el formato para evitar la notación científica
summary_df[, -4] <- lapply(summary_df[, -4], function(x) {
  format(x, scientific = FALSE)
})

# Imprimir la tabla sin notación científica
print(summary_df)

### Matriu de confusió:
preds_log <- bind_cols(
  predict(mdl_log, newdata = data_test, type = "prob"),
  Predicted = predict(mdl_log, newdata = data_test, type = "raw"),
  Actual = data_test$Severity
)

levels(preds_log$Predicted)
levels(preds_log$Actual)
preds_log$Actual <- factor(preds_log$Actual, levels = c("X0", "X1"))
cm_log <- caret::confusionMatrix(data = preds_log$Predicted,
                          reference = preds_log$Actual)
fourfoldplot(as.table(cm_log), main = "Confusion Matrix Model Classificació Logística",
          color = c("#CD0000", "#00CD66"), margin = 2)

### Corba de ROC
mdl_auc <- mdl_log$results[,2]
obs <- as.factor(data_test$Severity)
pred_obj <- prediction(preds_log[, "X1"], obs)  # 'X1' es la clase positiva
perf_obj <- performance(pred_obj, measure = "tpr", x.measure = "fpr")
plot(perf_obj, main = "Corba ROC Logística", col = "#1c61b6", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "gray")  # Línea de referencia diagonal
auc_obj <- performance(pred_obj, measure = "auc")
auc_value <- auc_obj@y.values[[1]]
print(paste("AUC:", auc_value))

# Arbres de decisió

mdl_cart <- caret::train(Severity ~ .,
                  data = data_train,
                  method = "rpart",
                  tuneGrid = expand.grid(cp = seq(from = 0.0001, to = 0.1, length = 15)
                  ),
                  metric = "ROC",
```

```
                    trControl = Under)
print(mdl_cart)

rpart.plot(mdl_cart$finalModel,
        extra = 1,
        cex = 0.7,
        fallen.leaves = TRUE,
        tweak = 1.2,
        clip.right.labs = FALSE,
        clip.left.labs = FALSE,
        faclen = 0)

preds_cart <- bind_cols(
  predict(mdl_cart, newdata = data_test, type = "prob"),
  Predicted = predict(mdl_cart, newdata = data_test, type = "raw"),
  Actual = data_test$Severity
)
preds_cart$Predicted <- factor(preds_cart$Predicted, levels = c("X0", "X1"))
preds_cart$Actual <- factor(preds_cart$Actual, levels = c("X0", "X1"))
par(mfrow = c(1,2))
cm_cart <- caret::confusionMatrix(data = preds_cart$Predicted, reference =
preds_cart$Actual)

fourfoldplot(as.table(cm_cart), main = "Test data",
          color = c("#CD0000", "#00CD66"), margin = 2)

mdl_auc_2 <- mdl_cart$results[,2]
obs_2 <- as.factor(data_test$Severity)
pred_obj_2 <- prediction(preds_cart[, "X1"], obs_2)  # 'X1' es la clase positiva
perf_obj_2 <- performance(pred_obj_2, measure = "tpr", x.measure = "fpr")
plot(perf_obj_2, main = "Corbaa ROC Logística", col = "#1c61b6", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "gray")  # Línea de referencia diagonal
auc_obj <- performance(pred_obj_2, measure = "auc")
auc_value <- auc_obj@y.values[[1]]
print(paste("AUC:", auc_value))

# Random Forest

mdl_rf <- caret::train(
  Severity ~ .,
  data = data_train,
  method = "rf",
  metric = "ROC",
  tuneGrid = expand.grid(mtry = 1:4),
  trControl = Under
)
preds_rf <- bind_cols(
```

```
  predict(mdl_rf, newdata = data_test, type = "prob"),
  Predicted = predict(mdl_rf, newdata = data_test, type = "raw"),
  Actual = data_test$Severity
)
preds_rf$Predicted <- factor(preds_rf$Predicted, levels = c("X0", "X1"))
preds_rf$Actual <- factor(preds_rf$Actual, levels = c("X0", "X1"))
par(mfrow = c(1,2))
cm_rf <- caret::confusionMatrix(preds_rf$Predicted,
                          reference = preds_rf$Actual)
fourfoldplot(as.table(cm_rf), main = "Test data",
          color = c("#CD0000", "#00CD66"), margin = 2)
mdl_auc <- mdl_rf$results[which(mdl_rf$results[,1] ==
                           mdl_rf$bestTune[[1]]),2]
obs_3 <- as.factor(data_test$Severity)
pred_obj_3 <- prediction(preds_rf[, "X1"], obs_3)  # 'X1' es la clase positiva
perf_obj_3 <- performance(pred_obj_3, measure = "tpr", x.measure = "fpr")
plot(perf_obj_3, main = "Corba ROC Random Forest", col = "#1c61b6", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "gray")  # Línea de referencia diagonal
auc_obj <- performance(pred_obj_3, measure = "auc")
auc_value <- auc_obj@y.values[[1]]
print(paste("AUC:", auc_value))

### Interpretacio random forest:
X <- data_train[,-4]
predictor <- Predictor$new(mdl_rf, data = X, y = data_train$Severity)
str(predictor)
imp <- FeatureImp$new(predictor, loss = "ce")
plot(imp)
X[] <- lapply(X, function(x) {
  if(is.character(x)) as.factor(x) else x
})
predictor_rf <- Predictor$new(mdl_rf,data=X, y=data_train$Severity)
effects <- FeatureEffects$new(predictor_rf)
plot(effects)
treeS <- TreeSurrogate$new(predictor_rf, maxdepth = 4)
treeS$r.squared
plot(treeS)
fitted.rf <- predict(mdl_rf, newdata = data_test, type="prob")
data_test$prob1.rf <- fitted.rf$X1
pred.rf <- predict(mdl_rf, newdata = data_test, type = "raw")
predictions <- ifelse(pred.rf == "X1", "X1", "X0")
testing_true<-data_test[data_test$Severity==data_test$pred.rf,]
which.min(testing_true$prob1.rf)
which.max(testing_true$prob1.rf)
summary(testing_true$prob1.rf)
testing_true[c(213,1924),c("Severity","pred.rf","prob1.rf")]
columns_of_interest <- setdiff(names(testing_true)[1:18], names(testing_true)[4])
```

```
shapley.min <- Shapley$new(predictor_rf, x.interest = testing_true[1023, columns_of_interest])
shapley.max <- Shapley$new(predictor_rf, x.interest = testing_true[3314, columns_of_interest])
plot2.min <- plot(shapley.min) + ggtitle("Min")
plot2.max <- plot(shapley.max) + ggtitle("Max")
gridExtra::grid.arrange(plot2.min, plot2.max, nrow = 1)

# XGBoost

tuneGrid <- expand.grid(
  nrounds = c(50, 100),
  max_depth = c(3, 5),
  eta = c(0.01, 0.1),
  gamma = c(0, 0.1),
  colsample_bytree = c(0.8, 1),
  min_child_weight = c(1, 3),
  subsample = c(0.8, 1)          )

tuneGrid <- sample_n(tuneGrid, 10)
mdl_xgb <- caret::train(
  Severity ~ .,
  data = data_train,
  method = "xgbTree",
  metric = "ROC",
  tuneGrid = tuneGrid,
  trControl = Under
)
print(mdl_xgb$bestTune)
predictor <- Predictor$new(
  model = mdl_xgb,
  data = data_train[, -which(names(data_train) == "Severity")],
  y = data_train$Severity
)
imp <- FeatureImp$new(predictor, loss = "ce")
print(imp)
plot(imp)

preds_xgb <- bind_cols(
  predict(mdl_xgb, newdata = data_test, type = "prob"),
  Predicted = predict(mdl_xgb, newdata = data_test, type = "raw"),
  Actual = data_test$Severity
)
preds_xgb$Predicted <- factor(preds_xgb$Predicted, levels = c("X0", "X1"))
preds_xgb$Actual <- factor(preds_xgb$Actual, levels = c("X0", "X1"))
par(mfrow = c(1,2))
cm_xgb <- caret::confusionMatrix(preds_xgb$Predicted,
```

```
                        reference = preds_xgb$Actual)
fourfoldplot(as.table(cm_xgb), main = "Test data",
          color = c("#CD0000", "#00CD66"), margin = 2)

mdl_auc <- mdl_xgb$results[which(mdl_xgb$results[,1] ==
                          mdl_xgb$bestTune[[1]]),2]
obs_4 <- as.factor(data_test$Severity)
pred_obj_4 <- prediction(preds_xgb[, "X1"], obs_4)  # 'X1' es la clase positiva
perf_obj_4 <- performance(pred_obj_4, measure = "tpr", x.measure = "fpr")
plot(perf_obj_4, main = "Corbaa ROC Logística", col = "#1c61b6", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "gray")  # Línea de referencia diagonal
auc_obj <- performance(pred_obj_4, measure = "auc")
auc_value <- auc_obj@y.values[[1]]
print(paste("AUC:", auc_value))
```