TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi de la sinistralitat d'una cartera d'automòbil

Autor: Anna Navarro Catarineu

Director: Salvador Torra Porras

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2023

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum

Els sinistres, en el ram dels automòbils, són uns fets que es produeixen diàriament. Cada dia es produeixen avaries als vehicles o accidents. Tothom té contractada una assegurança obligatòria per cobrir els danys que produeixen a les persones de fora. El grup Catalana Occident és una empresa pionera en el món de les assegurances. Aquesta recopila tots els sinistres que han passat en unes bases de dades mensual. En aquestes bases hi ha una variable que ens indica la fase en la qual es troba el sinistre. Però a vegades ens trobem amb problemes de classificació.

Per aquest motiu, l'ús del Machine learning o aprenentatge automàtic amb models de classificació pot ser un recurs per poder solucionar el problema de mala classificació de la fase en la qual es troba el sinistre. Aquests models ens permeten entrenar les dades per aprendre d'elles i obtenir classificacions i prediccions de les bases amb un nivell més alt de precisió.

Aquest treball té com a objectiu aplicar tècniques de classificació mitjançant el Machine Learning per a abordar el problema de la mala classificació de les fases del sinistre. Per això, utilitzem tres mètodes de classificació que avaluaran el seu rendiment. Aquests són: Arbres de decisió, Naive Bayes i Regressió Logística. Compararem aquests mètodes i seleccionarem el millor model de classificació per les dades del treball.

Paraules clau: Sinistre, Automòbils, Machine learning, models de classificació, Arbre de decisió, Naive Bayes i Regressió Logística.

Abstract

Accidents, within the field of automobiles, are events that occur daily. Vehicle breakdowns or accidents happen every day. Everyone has mandatory insurance coverage to compensate for damages caused to people outside the vehicle. Catalana Occident Group is a pioneering company in the insurance industry. They collect all the claims that occur in monthly databases. These databases include a variable indicating the phase in which the claim is. However, sometimes we encounter classification problems.

For this reason, the use of Machine Learning or automated learning with classification models can be a resource to solve the problem of misclassification of the claim phase. These models allow us to train the data to learn from them and obtain classifications and predictions from the databases with a higher level of accuracy.

This work aims to apply classification techniques using Machine Learning to address the issue of misclassification of claim phases. To do this, we use three classification methods that will evaluate their performance. These methods are Decision Trees, Naive Bayes, and Logistic Regression. We will compare these methods and select the best classification model for the data in the study.

Keywords: Claim, Automobiles, Machine learning, Classification models, Decision Tree, Naive Bayes, and Logistic Regression.

Classificació AMS

62-XX STATISTICS

- 62-07 Data analysis
- 62Q05 Statistical tables

68Txx Artificial intelligence

- 68Wxx Algorithms
- 68W40 Analysis of algorithms

Sumari

1. Introducció 10
1.1. Objectius i motivació 11
1.2. Hipòtesis 11
2. Metodologia 12
3. Apartat teòric 13
3.1. Assegurances d'automòbil 13
3.1.1. Què és una empresa asseguradora? 13
3.1.2. La línia de negoci d'automòbils 14
3.1.3. Grup Catalana Occident 16
3.2. Machine learning 17
3.2.1. Què és? 17
3.2.2. Àmbits on el podem aplicar 17
3.2.3. Com funciona? 18
3.2.4. Anàlisi de dades amb *machine learning* 18
3.2.5. Tipus de models 19
3.2.6. Algoritmes usats en aquest treball 20
4. Base de dades i anàlisi inicial 21
4.1. Base de dades 21
4.1.1. Obtenció de la base de dades 21
4.1.2. Descripció inicial de les dades 21
4.1.3. Resum inicial segons el tipus de variable 23
4.1.4. Creació de variables noves 25
4.2. Software 29
4.2.1. Paquets de Rstudio 29
5. Anàlisi de la base de dades 30
5.1. Preprocessament de les dades 30
5.1.1. Tractament dels outliers 30
5.1.2. Tractament dels missings 31
5.1.3. Anàlisi de les variables numèriques 34
5.1.4. Anàlisi de les variables categòriques 38
6. Aplicació dels models 50
6.1. Arbres de decisió 51
6.1.1. Arbres de decisió amb les dades d'entrenament 51
6.1.2. Arbres de decisió amb les dades de prova 54

6.1.3. Comparació dels resultats 57
6.2. Naive Bayes 58
6.2.1. Naive Bayes amb les dades d'entrenament 58
6.2.2. Naive Bayes amb les dades de prova 62
6.2.3. Comparació dels resultats 66
6.3. Regressió logística 67
6.3.1. Regressió logística amb les dades d'entrenament 67
6.3.2. Regressió logística amb les dades de prova 70
6.3.3. Comparació dels resultats 72
7. Comparació dels models 73
7.1. Models aplicats a les dades d'entrenament 73
7.2. Models aplicats a les dades de prova 74
7.3. Comparació de les dues mostres 75
8. Conclusions 76
9. Materials utilitzats 78
9.1. Webgrafia 78
9.2. Apunts classe 79
9.3. Bibliografia 79
Annex 80
Codi Rstudio 80

Índex de taules

Taula 1: Descripció de les variables de la base de dades 22
Taula 2: Tipus de variables de la base de dades 23
Taula 3: Sumari de la variable numèrica de la provisió inicial (PROVINI) 23
Taula 4: Sumari de la variable numèrica de la provisió en el moment actual (PROVACT) 23
Taula 5: Sumari de la variable numèrica de l'import base de pagaments de l'exercici actual (PAGB) 24
Taula 6: Sumari de la variable numèrica de l'import IVA de pagaments de l'exercici actual (PAGI) 24
Taula 7: Nombre de nivells de les variables categòriques 24
Taula 8: Nombre i freqüència de vehicles possibles 25
Taula 9: Nivells de la variable NUMVEH 26
Taula 10: Nivells i freqüència de la variable resposta 28
Taula 11: Paquets Rstudio utilitzats 29
Taula 12: Classificació dels missings 31
Taula 13: Tants per u dels missings 32
Taula 14: Freqüència de la variable INDSITU 39
Taula 15: Freqüència de la variable resposta 39
Taula 16: Freqüència de la variable INDRECO 40
Taula 17: Freqüència de la variable INDCONS 41
Taula 18: Freqüència de la variable INDCULP 42
Taula 19: Freqüència de la variable INDDASE 43
Taula 20: Freqüència de la variable NUMVEH 43
Taula 21: Freqüència de la variable INDCOLD 44
Taula 22: Freqüència de la variable IMDCORP 44
Taula 23: Freqüència de la variable INDCIDR 45
Taula 24: Freqüència de la variable INDINCE 45
Taula 25: Freqüència de la variable INDROBO 46
Taula 26: Freqüència de la variable DANOSMAT 46
Taula 27: Freqüència de la variable INDDAPR 47
Taula 28: Freqüència de la variable INDASV 47
Taula 29: Freqüència de la variable TIPTECD 48
Taula 30: Freqüència de la variable PERDTOT 49
Taula 31: Freqüència de la variable TALPREF 49
Taula 32: Paràmetres calculats segons les dades utilitzades 57
Taula 33: Paràmetres calculats segons les dades utilitzades 66
Taula 34: Paràmetres calculats segons les dades utilitzades 72
Taula 35: Models aplicats a les dades d'entrenament 73
Taula 36: Models aplicats a les dades de prova 74

Índex de figures

Figura 1: Grup Catalana Occident 16
Figura 2: Gràfica de la distribució del nombre de vehicles externs involucrats 25
Figura 3: Gràfica de la distribució del nombre de vehicles externs dels valors 0, 1 i 2 26
Figura 4: Gràfica de la distribució del nombre de vehicles externs dels valors superiors a 2 .. 26
Figura 5: Distribució de la nova variable 27
Figura 6: Distribució de la situació actual del sinistre 27
Figura 7: Distribució de la nova variable resposta 28
Figura 8: Distàncies de Cook de la variable de la provisió inicial 30
Figura 9: Distàncies de Cook de les variables de la provisió en el moment actual 30
Figura 10: Distàncies de Cook de la variable de l'import base de pagaments de l'exercici actual 31
Figura 11: Distàncies de Cook de la variable de l'import IVA de pagaments de l'exercici actual 31
Figura 12: Estructura de les dades abans del preprocessament 32
Figura 13: Gràfiques del patró que segueixen els valors missing 33
Figura 14: Sumari després del preprcessament de les dades 34
Figura 15: Histograma i boxplot inicials de PROVINI 34
Figura 16: Histograma i boxplot finals de PROVINI 35
Figura 17: Histograma i boxplot inicials de PROVACT 35
Figura 18: Histograma i boxplot finals de PROVACT 35
Figura 19: Histograma i boxplot inicials de PAGB 36
Figura 20: Histograma i boxplot finals de PAGB 36
Figura 21: Histograma i boxplot inicials de PAGI 37
Figura 22: Histograma i boxplot finals de PAGI 37
Figura 23: Gràfica de sectors de la variable RAMEMIS 38
Figura 24: Gràfica de sectors de la variable RAMITO 38
Figura 25: Gràfica de sectors de la variable INDSITU 39
Figura 26: Gràfica de sectors de la variable Y 39
Figura 27: Gràfica de sectors de la variable INDTIVE 40
Figura 28: Gràfica de sectors de la variable INDRECO 40
Figura 29: Gràfica de sectors de la variable INDCONS 41
Figura 30: Gràfica de sectors de la variable TOTVEHC 41
Figura 31: Gràfica de sectors de la variable CNATSIN 42
Figura 32: Gràfica de sectors de la variable INDCULP 42
Figura 33: Gràfica de sectors de la variable INDDASE 43
Figura 34: Gràfica de sectors de la variable NUMVEH 43
Figura 35: Gràfica de sectors de la variable INDCOLD 44
Figura 36: Gràfica de sectors de la variable IMDCORP 44
Figura 37: Gràfica de sectors de la variable INDVIDR 45
Figura 38: Gràfica de sectors de la variable INDINCE 45
Figura 39: Gràfica de sectors de la variable INDROBO 46
Figura 40: Gràfica de sectors de la variable DANOSMAT 46
Figura 41: Gràfica de sectors de la variable INDDAPR 47
Figura 42: Gràfica de sectors de la variable INDASV 47
Figura 43: Gràfica de sectors de la variable TIPTECD 48

Figura 44: Gràfica de sectors de la variable TIPUSUD 48
Figura 45: Gràfica de sectors de la variable PERDTOT 49
Figura 46: Gràfica de sectors de la variable TALPREF 49
Figura 47: Model d'Arbre de decisió amb les dades d'entrenament 51
Figura 48: Arbre de decisió amb les dades d'entrenament 52
Figura 49: Matriu de confusió amb les dades d'entrenament 52
Figura 50: Corba de ROC amb les dades d'entrenament 53
Figura 51: Model d'Arbre de decisió amb les dades de prova 54
Figura 52: Arbre de decisió amb les dades de prova 55
Figura 53: Matriu de confusió amb les dades de prova 55
Figura 54: Corba de ROC amb les dades de prova 56
Figura 55: Model Naive Bayes amb les dades d'entrenament 58
Figura 56: Variable resposta Y amb la provisió actual 58
Figura 57: Variable resposta Y amb la provisió inicial 59
Figura 58: Variable resposta Y amb l'indicador del sinistre recobrable 59
Figura 59: Variable resposta Y amb l'indicador de culpa 59
Figura 60: Variable resposta Y amb el tipus de declaració del sinistre de tecnologia 60
Figura 61: Matriu de confusió de les dades d'entrenament 60
Figura 62: Corba de ROC del model Naive Bayes amb les dades d'entrenament 61
Figura 63: Model Naive Bayes amb les dades de prova 62
Figura 64: Variable resposta Y amb la provisió actual 62
Figura 65: Variable resposta Y amb la provisió inicial 63
Figura 66: Variable resposta Y amb l'indicador del sinistre recobrable 63
Figura 67: Variable resposta Y amb l'indicador de culpa 63
Figura 68: Variable resposta Y amb el tipus de declaració del sinistre de tecnologia 64
Figura 69: Matriu de confusió de les dades de prova 64
Figura 70: Corba de ROC del model Naive Bayes amb les dades de prova 65
Figura 71: Model de regressió logística amb les dades d'entrenament 67
Figura 72: Gràfica de la variable resposta, PROVACT i PROVINI 68
Figura 73: Matriu de confusió de les dades d'entrenament 68
Figura 74: Corba de ROC del model de classificació lineal amb les dades d'entrenament 69
Figura 75: Model de regressió logística amb les dades de prova 70
Figura 76: Gràfica de la variable resposta, PROVACT i PROVINI 71
Figura 77: Matriu de confusió de les dades de prova 71
Figura 78: Corba de ROC del model de classificació lineal amb les dades de prova 72

1. Introducció

Els sinistres, en el ram dels automòbils, són uns fets que es produeixen a diàriament. Cada dia es produeixen avaries als vehicles o accidents. Tothom té contractada una assegurança obligatòria per cobrir els danys que produeixen a les persones de fora. El grup Catalana Occident[1] és una empresa pionera en el món de les assegurances. Aquesta recopila tots els sinistres que han passat en una base de dades per mesos. En el mes de desembre de cada any, tenim la base de dades completa. En aquestes dades queda registrada la fase en la qual es troba el sinistre, aquesta pot ser la 0, la 1, la 2 o la 3[2]. Però a vegades hi ha problemes de classificació, és a dir, que hi ha sinistres que es troben en fases equivocades per problemes de logística.

Per aquest motiu, l'ús del Machine learning o aprenentatge automàtic amb models de classificació pot ser un recurs per poder solucionar el problema de mala classificació de la fase en la qual es troba el sinistre. Aquests models ens permeten entrenar les dades per aprendre d'elles i obtenir classificacions i prediccions de les bases amb un nivell de precisió més alt.

L'origen d'aquest estudi comença amb les meves pràctiques a l'empresa asseguradora Catalana Occident o com ara es coneix com a Occident. Durant aquestes pràctiques he tingut un clar interès pel ram dels automòbils, ja que és el portafolis en el qual he començat a involucrar-me. Quan fas divuit anys et surt l'oportunitat de treure't el carnet de cotxe i guanyar independència al no dependre de ningú per moure't. Però quan comences a conduir necessites una assegurança per allò que poguessis ocasionar als altres mentre estàs a la carretera.

La sinistralitat de l'automòbil no només es redueix als accidents. Qualsevol cosa que cobreixi la pòlissa contractada també s'hi considera. Un exemple seria si hi ha hagut un robatori en el vehicle o quan es necessita assistència a la carretera.

L'any 2020, la població espanyola es va veure afectada per una pandèmia mundial produïda pel virus de la Covid-19. Des de llavors, les empreses, a poc a poc, s'estan recuperant i la circulació s'ha normalitzat. Per dur a terme aquest treball, hem triat la sinistralitat de l'any 2019, ja que va ser l'últim any que podem garantir la normalitat en la circulació.

1. Grup Catalana Occident: *https://www.seguroscatalanaoccidente.com/cat*

2. Fases en la que es troba el sinistre: Fase 0: Notificació; Fase 1: Investigació i avaluació; Fase 2: Processament de la reclamació; Fase 3: Liquidació del sinistre

1.1. Objectius i motivació

La motivació inicial d'aquest estudi és poder veure les tècniques d'aprenentatge automàtic vistes a classe aplicades a la realitat. L'objectiu principal del treball serà veure la mala classificació de les fases del sinistre utilitzant diferents algoritmes i veure'n la diferència entre ells.

Començarem veient el món de les assegurances i el portafolis de l'automòbil. Veurem tots els factors implicats i així tindrem una àmplia visió del tema de què estem parlant. A continuació proposarem una sèrie de models de classificació per veure si el fenomen mencionat és cert i compararem els diferents models emprats a través de diferents paràmetres per veure la seva capacitat de predicció. L'objectiu final serà trobar el millor model, és a dir, aquell que tingui la millor capacitat de predicció per les nostres dades.

1.2. Hipòtesis

Un altre punt important a veure són les hipòtesis formulades abans de començar l'estudi. La primera hipòtesi és que es creu que no hi haurà molta diferència significativa en la comparació dels diferents models de classificació. Com a segona hipòtesis, s'esperaria que hi hagués una mala classificació de la variable resposta i que, per tant, hi hagués un percentatge alt de falsos positius.

2. Metodologia

En aquest capítol veurem la metodologia utilitzada durant aquest estudi. Aquest treball el tenim dividit en dos grups principals, la part teòrica i la part pràctica.

En aquesta part teòrica aprendrem tots els coneixements necessaris del món de les assegurances, sobretot en la part dels automòbils. Veurem les diferents pòlisses i, per tant, els diferents tipus de sinistres que tindrem registrats a la base de dades.

Per altra banda, també veurem que és l'aprenentatge automàtic i les seves funcions. A partir de la descripció dels diferents models, entendrem perquè serveixen cadascun. Hem tingut en compte els **arbres de decisió**, en què es prenen decisions a partir de l'estructura jeràrquica de regles. També hem tingut en compte **Naive Bayes**, en què necessitarem la teoria de la probabilitat per classificar i, per últim, la **Regressió Logística**, en què a partir d'una funció logística estimarem la probabilitat de què un sinistre estigui a una fase en concret.

En la part pràctica veurem els resultats un cop s'han aplicat aquests algoritmes però, també, com s'ha preparat la base de dades.

Començarem aquest apartat amb el preprocessament i una anàlisi de cada variable depenent del tipus en el qual pertanyi. Veurem que tenim variables categòriques i numèriques. El preprocessament de les dades consisteix a veure el comportament de les dades buides i dels valors punta i, després, decidir. Netejarem la base de dades i crearem les variables necessàries per dur a terme el projecte.

A continuació implementarem els diferents models, on a cada un d'ells calcularem els paràmetres necessaris per poder, després, fer una comparativa. Utilitzarem el programa de Rstudio amb unes funcions específiques per poder aplicar-los.

Finalment, valorarem quin dels models aplicats és el que classifica de millor manera la nostra variable resposta i, per tant, té més precisió.

3. Apartat teòric

En aquest capítol mostrarem els conceptes bàsics dels quals parlarem en aquest treball. Començarem introduint el camp on es mou l'empresa, d'on hem extret les dades, i descrivint la línia de negoci en la que en centrarem. A més a més, farem un resum del concepte de Machine learning i tot el que és necessari saber, com per exemple, els diferents tipus de tècniques que hi ha.

3.1. Assegurances d'automòbil

En aquest apartat explicarem què és una empresa asseguradora i una pòlissa d'assegurances. A més a més, mostrarem els diferents tipus d'assegurances que hi ha quan ens centrem en el portafolis d'Automòbils.

3.1.1. Què és una empresa asseguradora?

El diccionari de l'institut d'estudis catalans ens defineix les assegurances com "contracte mitjançant el qual, a canvi d'una prima, pagada d'un cop o anualment, mensualment, etc., és garantit algú contra un risc"[3]. En unes altres paraules, diríem que una empresa asseguradora és aquella que es fa càrrec del que acabarà costant un sinistre ocorregut als seus clients, fent ús de la quota que han pagat a priori.

L'import de la quota que els clients pagaran, normalment, està determinat en funció del risc que la companyia calculi sobre els danys, lesions o pèrdues. Tant l'empresa asseguradora com els assegurats firmen un contracte amb les seves clàusules i condicions pactades sobre aquests riscos mencionats anteriorment. Aquest document s'anomena **pòlissa d'assegurances**. Quan es produeix un sinistre, l'empresa arregla els possibles danys o paga la quantitat proporcional de diners a l'altra part afectada no culpable en el cas dels accidents o altres situacions.

Perquè l'asseguradora pugui firmar un gran nombre de contractes i, a la vegada, protegir a un gran nombre de clients, és necessari que tingui una reserva i tenir un alt marge de solvència davant dels pagaments futurs propis de la seva activitat econòmica. Aquesta és la raó per la qual les companyies d'assegurances tenen molta importància i són un pilar del mercat financer juntament amb les companyies bancàries.

Dins de cada empresa asseguradora, tenim diferents tipus d'assegurances, depenent del que estiguin assegurant. A continuació mostrem tres exemples d'aquests tipus:

- **Assegurances de persones:** són aquelles que cobreixen els riscs que afecten a la vida, a la salut o a la integritat d'una o més persones.
- **Assegurances de danys o sobre coses:** són aquelles que cobreixen els riscs ocorreguts sobre els béns que han assegurat.
- **Assegurances de patrimoni o de responsabilitat:** són aquelles que cobreixen els riscos que produeixin obligacions patrimonials a les persones que l'han contractat.

3. Diccionari de l'Institut d'Estudis Catalans: *https://dlc.iec.cat/*

3.1.2. La línia de negoci d'automòbils

La definició tècnica de les assegurances d'automòbils diu que és un contracte firmat entre una entitat asseguradora i una persona que és conductora. Aquest contracte cobrirà els riscs generats per la conducció d'un automòbil i la possibilitat de causar algun accident.

En la majoria dels països, l'assegurança en automòbils és obligatòria, sempre que la modalitat obligada per l'estat sigui la bàsica. La modalitat bàsica és la que cobreix la responsabilitat civil del conductor i propietari de l'automòbil. Encara que el contracte pot incloure altres modalitats. Poden cobrir els riscs d'un accident sense danys a tercers i altres sèries de complements voluntaris que fan que s'ampliï les cobertures de l'assegurança.

Dins d'aquesta línia de negoci trobem diferents tipus en funció de la cobertura dels danys i riscs. A continuació en distingim uns exemples[4]:

Assegurança a tercers

Aquesta és l'assegurança bàsica, la qual es limita a cobrir la responsabilitat civil i és obligatòria per l'Estat.

En el grup Catalana Occident l'assegurança a tercers cobreix els punts següents:

- **Responsabilitat civil obligatòria**: és la cobertura principal i bàsica en qualsevol assegurança. S'encarrega de fer front als danys materials i personals que causis durant la conducció.
- **Responsabilitat civil per incendi del vehicle**: en cas que hi hagi un incendi o una explosió del vehicle que està assegurat, l'assegurança es fa càrrec dels danys causats.
- **Trencament de vidres**: l'assegurança es fa càrrec de reparar les llunes del cotxe, tant en un taller com al domicili.
- **Assistència en viatge**: inclou una indemnització per rescat i km 0 en ajuda tècnica en carretera.
- **Defenses i reclamació**: l'empresa s'encarrega de recórrer les multes i dels cursos de la recuperació dels punts del carnet
- **Danys causats per animals cinegètics**: tots aquells danys causats per senglars, cérvols i altres animals de caça que poden causar greus destrosses en cas de col·lisió.

4. Assegurances a contractar del grup: *https://www.seguroscatalanaoccidente.com/cat/assegurances-cotxe*

Assegurança a tercers completa

Aquesta és la mateixa que l'anterior però, en aquest cas s'hi afegeixen els següents avantatges:

- **Furt i robatori**: l'assegurança actua en cas de robatori i, també, inclou els desperfectes ocasionats per l'intent de treure una peça del cotxe assegurat.
- **Incendis**: cobreix el mateix que l'anterior però, en aquest cas també cobreix els danys que es produeixin en el vehicle assegurat.

Assegurança a tot risc sense franquícia

Aquesta és la que té la cobertura més amplia. Garanteix que tots els danys del vehicle, incloent-hi el robatori, estiguin inclosos. També inclou els danys causats per fenòmens atmosfèrics o danys en el teu automòbil.

Assegurança a tot risc amb franquícia

Aquesta és semblant a l'anterior, però en aquest cas s'implementen pagaments complementaris per la prestació d'alguns serveis en la reparació del vehicle.

Com hem dit anteriorment, les diferents modalitats es fixen segons les cobertures. A continuació mostrarem els tres tipus de cobertura, ordenats segons el cost de les assegurances de menor a major:

- **Cobertura de responsabilitat civil:** aquesta només cobreix aquells danys que es produeixen a tercers.
- **Cobertura limitada:** aquesta cobreix el mateix que l'anterior, però, a més a més, s'afegeixen altres cobertures com el robatori del vehicle o arreglar el vidre de davant quan s'ha trencat.
- **Cobertura ampliada:** aquesta és la més gran, cobreix pràcticament tot el que li passa al vehicle, independentment de qui hagi causat l'accident.

3.1.3. Grup Catalana Occident

El Grup Catalana Occident (GCO) és una empresa asseguradora líder en el seu sector i en el de l'assegurança de crèdit al món. En l'actualitat, s'estan fusionant diverses companyies en una nova anomenada Occident. Les tres companyies que el grup ha comprat són *NorteHispana*, *SegurosBilbao* i *PlusUltra*.

Amb un creixement constant i una gran implantació, compta amb més de 7.300 empleats, té presència en més de 50 països i dona servei a més de 4.000.000 assegurats. La seva xarxa consta de 1.600 oficines i de prop de 19.000 mediadors. Actualment, ocupa la sisena posició en el mercat espanyol assegurances generals i la segona a escala mundial en l'assegurança de crèdit.

Avui dia, el Grup Catalana Occident és una de les empreses d'assegurances més grans del país gràcies a la seva experiència, una solvència històrica contrastada i una gestió coherent i rendible. A més a més, cotitza en borsa i presenta un *holding* amb empreses tan importants com *Atradius*, que opera a gairebé 40 països.

Figura 1: Grup Catalana Occident

Origen de la imatge: https://www.gco.com/cat/marques/assegurances-catalana-occident

3.2. Machine learning

En aquest apartat introduirem l'aprenentatge automàtic. Veurem què és, com funciona i quins tipus de mètodes es troben dins d'aquest camp.

3.2.1. Què és?

Els mètodes de *machine learning* o aprenentatge automàtic són un dels camps de les ciències de la computació i de la intel·ligència artificial. L'objectiu principal d'aquests mètodes és poder identificar patrons recurrents en un conjunt de dades. Aquestes dades poden ser números, paraules, imatges, estadístiques, entre altres. Qualsevol informació que es pugui guardar de manera digital pot servir. Gràcies a la detecció de patrons en la base, els algoritmes aprenen i milloren el rendiment amb el temps. Quan l'algoritme està entrenat es podran trobar els patrons.

3.2.2. Àmbits on el podem aplicar

L'aprenentatge automàtic té molts àmbits d'aplicació. A mesura que la tecnologia es desenvolupa i apareixen nous descobriments, aquests àmbits amplien. Alguns dels exemples[5] són:

- **Detectar el rostre**: Molts dels dispositius que utilitzem dàriament necessiten el nostre rostre reconeixent-lo com a imatge. S'usa en les xarxes socials, per exemple.
- **Correu electrònic**: En el correu moltes vegades en rebem de no desitjat. Gràcies als mètodes de Machine learning, el sistema aprèn d'exemples passats, és a dir, aquells correus que no són desitjats, i pren decisions en el futur per evitar-los.
- **Antivirus**: Aprèn a detectar aquells softwares dolents basats en les dades que s'han introduït prèviament i en prediccions possibles.
- **Finances**: Crea algoritmes per poder aprendre els patrons que segueix una inversió, de manera que, compra i ven de manera eficient.
- **Vehicles autònoms**: es fan ús algunes tècniques pels cotxes que es condueixen sols, reconeixent la ruta, tenint en compte els cotxes i l'entorn en el qual es troba. Aprenen dels seus errors i milloren el comportament.
- **Generació de textos**: interpreta els textos proporcionats a través d'un algoritme i d'un etiquetat concret que ens permet que el dispositiu generi els seus textos propis.

5. Exemples dels àmbits del Machine learning: *https://www.tokioschool.com/noticias/aplicaciones-machine-learning/*

3.2.3. Com funciona?

Per poder desenvolupar un model de *machine learning* tenim quatre etapes principals:

PRIMERA ETAPA

Aquesta **primera etapa** consisteix a seleccionar, preparar, organitzar i netejar les dades d'entrenament. Si no es fes, el model podria sortir esbiaixat i les prediccions estaran afectades. Les dades d'entrenament les podem etiquetar per així, indicar-li al model, quines són les característiques o patrons que hem de buscar. En el cas de no fer-ho, el model haurà d'extreure-les per ell mateix.

SEGONA ETAPA

En la **segona etapa** s'ha de triar l'algoritme en funció del tipus i del volum de dades d'entrenament i el tipus de problema que hem de resoldre.

TERCERA ETAPA

Durant la **tercera etapa** entrenem l'algoritme. Aquest és un procés amb moltes repeticions. Executem l'algoritme sobre les dades d'entrenament i els resultats extrets els comparem amb els que s'haurien d'haver produït. Per poder augmentar la precisió dels resultats podem ajustar els pesos i el biaix. Executem les variables fins que l'algoritme ens doni el resultat correcte a la majoria dels casos. Aquest algoritme entrat és el model que utilitzarem per resoldre el problema.

QUARTA ETAPA

En la quarta i **última etapa**, usarem i millorarem el model generat en l'etapa anterior sobre les dades noves.

3.2.4. Anàlisi de dades amb *machine learning*

El *machine learning* s'utilitza de forma massiva per l'anàlisi de dades i la ciència de dades. Ens permet desenvolupar, fer proves i aplicar algoritmes per poder fer una anàlisi predictiva sobre els diferents tipus de dades i poder predir el futur.

Gràcies al *machine learning* podem accelerar l'anàlisi de dades i aconseguir que sigui més precís automatitzant el desenvolupament del model analític. Ajuda a assignar als algoritmes les tasques centrals com, per exemple la classificació, el *clustering* o la detecció de dades anòmales.

Els algoritmes fan servir les dades per tornar inferències estadístiques i s'automilloren amb el temps. Són capaços de prendre les dedicions necessàries per resoldre el problema sense intervencions externes quan detecten algun canvi en les dades. Igualment, l'ésser humà encara ha de revisar els resultats de les dades, per donar-los un sentit i s'assegurar-se que les dades tractades no estan esbiaixades ni alterades.

3.2.5. Tipus de models

Els mètodes de *machine learning* els classifiquem en quatre grups d'aprenentatge: el supervisat, el no supervisat, *semisupervisat* i el reforçat.

Models supervisats

Aquest és el tipus més comú, en el qual, les dades s'etiqueten per indicar al software quins patrons ha de buscar. La finalitat d'aquests models és predir un esdeveniment o estimar el valor d'una variable numèrica contínua. Un factor que s'ha de tenir en compte és que el model pot estar esbiaixat per culpa de les dades d'entrenament i pot afectar en el rendiment quan processem les dades noves. Aquests models tenen variables descriptives (predictores) i, també, tenen una variable objectiu (resposta) que està associada als predictors a través d'una funció generada pel model. Inclouen models de classificació i models d'estimació. En el cas d'aquests models, les dades d'entrenament poden ser menors que en els altres tipus, ja que facilita el procés d'entrenament.

Models no supervisats

En aquest tipus de models les dades no tenen etiquetes. La finalitat és que l'algoritme reconeixi patrons i estructures dins de les dades, les quals no estan guiades per una variable objectiu específica ni hi ha variable resposta. Utilitza molta quantitat de dades i extreu les característiques rellevants per poder etiquetar, ordenar i classificar les dades en temps real sense que hi hagi hagut interacció humana. Inclou models de clúster, models d'associació i models de reducció de dimensió.

Models semisupervisats

Aquest tipus de model es troba entre els supervisats i els no supervisats, un punt mig. Fem servir un conjunt de dades etiquetat més petit durant l'entrenament per poder guiar la classificació i l'extracció de les característiques de la base de dades, la qual no està etiquetat i és més gran. Aquest model s'usarà quan no tinguem suficients dades etiquetades per poder entrenar l'algoritme i solucionar el problema.

Models reforçats

Aquest model consisteix a deixar que l'algoritme aprengui dels errors per ell mateix. Aquest anirà provant diferents maneres per intentar aconseguir l'objectiu. Depenent de com rendeixi, es recompensarà o penalitzarà, per així, animar-lo a continuar o provar una altra manera d'enfocar el problema.

3.2.6. Algoritmes usats en aquest treball

Arbres de decisió

L'arbre de decisió és un algoritme supervisat que utilitzem per a problemes de classificació i regressió. Es basa, principalment, en la creació d'una estructura en forma d'arbre construïda a partir d'uns nodes i branques que ens ensenyen les decisions i els resultats possibles. Cada un dels nodes d'un arbre representa una característica i cada branca una condició.

En el procés de construcció de l'arbre dividim les dades d'entrenament. S'intenta buscar les divisions que maximitzin el guany d'informació i, així, hi ha una reducció de la incertesa després de la divisió.

Amb l'arbre construït podem fer prediccions de nous exemples. Seguirem el camí, a través de l'arbre, basant-nos en les condicions dels nodes interns fins arribar a l'últim node.

En l'apartat 6.1. podrem veure com apliquem l'algoritme d'arbres de decisió en les nostres dades i els resultats que ens dóna.

Naive Bayes

Naive Bayes és un algoritme supervisat basat en el teorema de Bayes. L'objectiu és assignar una classificació en unes dades d'entrada quan hi ha problemes de classificació.

En aquest model se suposa la independència condicional, és a dir, que assumeix que les variables predictores són independents entre si, donat el valor de la classe objectiu. Això ens permet que sigui eficient i fàcil d'aplicar.

Per un altre costat, el nom Naive es deu a la suposició d'aquesta independència condicional, ja que, en la majoria dels casos, les característiques no són completament independents en la vida real.

En aquest algoritme s'utilitza el teorema de Bayes per calcular la probabilitat condicional de què un registre pertanyi a una determinada classe segons les seves característiques. Aquesta probabilitat la calcularem usant la probabilitat a priori de cada una de les classes i les probabilitats de les característiques condicionals de les classes.

En l'apartat 6.2. podrem veure com apliquem el model Naive Bayes en les nostres dades i els resultats que ens dóna.

Regressió logística

La regressió logística és un algoritme supervisat utilitzat per problemes de classificació. Aquest algoritme es basa en una regressió lineal que estima les probabilitats de que un exemple pertanyi a una classe concreta. S'assumeix que la relació entre les variables d'entrada i la probabilitat de pertànyer a la classe és lineal i en la funció es comprimeixen els valors en un rand entre 0 i 1. Aquests valors s'interpreten com la probabilitat de que estigui a la classe positiva.

Quan tenim el model entrenat, l'utilitzem per fer prediccions de com es classificarien nous exemples. Calcularem la probabilitat amb els coeficients ja apresos i prendrem una decisió de classificació.

En l'apartat 6.3. podrem veure com apliquem el model Regressió logística en les nostres dades i els resultats que ens dóna.

4. Base de dades i anàlisi inicial

En aquest capítol mostrarem l'estructura de la base de dades i explicarem la metodologia del disseny del model. També inclourem la descripció d'aquesta, les seves variables i farem l'anàlisi descriptiva. Finalment, veurem el software i els paquets utilitzats durant aquest estudi.

4.1. Base de dades

4.1.1. Obtenció de la base de dades

Les bases de dades utilitzades per dur a terme aquest estudi han estat cedides per la companyia asseguradora Catalana Occident. Ens han donat la base de dades de l'any 2019.

La base original consta de 66 variables, de les quals ens en quedem 32 de moment, durant l'estudi anirem eliminant algunes altres. Això és degut al fet que algunes de les variables són confidencials o no ens interessen per aquest estudi. Per altra banda, tenim un total de 755062 registres en la base.

4.1.2. Descripció inicial de les dades

De manera mensual, l'empresa recull els sinistres ocorreguts als seus clients. Aquesta sinistralitat és acumulativa, de manera que en el recull de dades del mes de desembre tindrem la fase on es troben els sinistres declarats durant aquell any.

La nostra base de dades és la que correspon al mes de desembre de l'any 2019. En aquesta tenim els diferents rams de la línia d'automòbils, de manera que ens donen la informació de la situació de cada un dels sinistres al final any.

La matriu de la base té 755062 files i 38 columnes. En la taula següent veurem la descripció de les diferents variables, agrupades en quatre classificadors:

Taula 1: Descripció de les variables de la base de dades

CLASSIFICADOR	VARIABLE	DESCRIPCIÓ
DADES DEL SINISTRE	RAMEMIS	Ram GCO
	RAMITO	Ram de gestió
	INDSITU	Situació actual del sinistre
TIPOLOGIA I INDICACIONS	TIPSINI	Tipologia del sinistre
	CODAGTI	Subtipologia del sinistre
	INDTIVE	Categoria
	INDRECO	Indicador del sinistre recobrable
	INDCONS	Indicador del sinistre consorciable
	TIPVEHI	Classe de vehicle (només autos)
	CNATSIN	Naturalesa del sinistre
	INDCIDE	Tipus de conveni (valor per CIDE)
	INDSDMM	Indicador SDM
	INDCULP	Indicador de culpa
	INDDASE	Indicador de danys del vehicle
	TOTVEHC	Número total de vehicles externs involucrats
	INDCOLD	Indicador de col·lisió
	IMDCORP	Indicador de lesionats
	INDVIDR	Indicador de vidre
	INDINCE	Indicador d'incendi
	INDROBO	Indicador de robatori
	DANOSMAT	Indicador de danys materials
	INDDAPR	Indicador de danys propis
DADES ECONÒMIQUES (RESERVES)	PROVINI	Provisió inicial
	PROVACT	Provisió en el moment actual
DADES ECONÒMIQUES (PAGAMENTS)	PAGB	Import base de pagaments d'exercici actual
	PAGI	Import IVA de pagaments d'exercici actual
ALTRES	INDASV	Indicador d'assistència de viatge
	TIPTECD	Tipus de declaració del sinistre de tecnologia
	TIPUSUD	Tipus de declaració del sinistre d'usuari
	CULPOBJ	Culpa objectiva (només DC)
	PERDTOT	Pèrdua total del vehicle assegurat
	TALPREF	Vehicle assegurat reparat pel taller preferent

A continuació, trobem la classificació en forma de taula de cada variable depenent del tipus que és. Aquesta ens serà útil a l'hora de fer l'anàlisi descriptiva de la base de dades.

Taula 2: Tipus de variables de la base de dades

VARIABLE	TIPUS DE VARIABLE	VARIABLE	TIPUS DE VARIABLE
RAMEMIS	Categòrica	IMDCORP	Categòrica dicotòmica
RAMITO	Categòrica	INDVIDR	Categòrica dicotòmica
INDSITU	Categòrica	INDINCE	Categòrica dicotòmica
TIPSINI	Categòrica	INDROBO	Categòrica dicotòmica
CODAGTI	Categòrica	DANOSMAT	Categòrica dicotòmica
INDTIVE	Categòrica	INDDAPR	Categòrica dicotòmica
INDRECO	Categòrica dicotòmica	PROVINI	Numèrica
INDCONS	Categòrica dicotòmica	PROVACT	Numèrica
TIPVEHI	Categòrica	PAGB	Numèrica
CNATSIN	Categòrica	PAGI	Numèrica
INDCIDE	Categòrica	INDASV	Categòrica dicotòmica
INDSDMM	Categòrica	TIPTECD	Categòrica
INDCULP	Categòrica	TIPUSUD	Categòrica
INDDASE	Categòrica dicotòmica	CULPOBJ	Categòrica
TOTVEHC	Categòrica	PERDTOT	Categòrica dicotòmica
INDCOLD	Categòrica dicotòmica	TALPREF	Categòrica dicotòmica

En aquesta taula podem veure que tenim quatre variables numèriques i vint-i-vuit variables categòriques, entre elles, tretze dicotòmiques.

4.1.3. Resum inicial segons el tipus de variable

Seguidament, tenim un resum inicial de les nostres dades depenent del tipus de variable.

Per una banda, per les variables numèriques realitzem un sumari amb el software R-studio. Aquest el podem veure a les taules següents:

Taula 3: Sumari de la variable numèrica de la provisió inicial (PROVINI)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-405119,6	150	305	378,1	350	1248171,5

Podem veure que el valor mínim de la provisió inicial és de -405119,6. Aquest valor es pot considerar un valor punta, ja que la mitjana en val 378,1. El mateix ens passa amb el valor màxim.

Taula 4: Sumari de la variable numèrica de la provisió en el moment actual (PROVACT)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2888	0	0	432	0	7063689

Observant el resum de la variable PROVACT, intuïm que tenim molts registres amb valor 0. Donat que, tant el primer quadrant com el segon (mediana) com el tercer, val 0. Considerarem que tant el valor mínim com el valor màxim són valors punta.

Taula 5: Sumari de la variable numèrica de l'import base de pagaments de l'exercici actual (PAGB)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-859981	33	78	407	285	3459787

En la variable PAGB també podem intuir que hi ha valors punta tant per dalt com per baix, ja que el valor del màxim és massa alt i el del mínim és massa baix.

Taula 6: Sumari de la variable numèrica de l'import IVA de pagaments de l'exercici actual (PAGI)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-586,01	2,29	11,95	28,70	22,94	11538,67

En aquest resum podem deduir que el valor màxim de la variable és un valor punta. En canvi, el valor mínim no l'haurem d'eliminar durant el preprocessament.

Per altra banda, per les variables categòriques numerem el nombre de nivells de cada una d'elles i les llistarem en una taula. Aquesta la podem veure a continuació:

Taula 7: Nombre de nivells de les variables categòriques

VARIABLE	NIVELLS	VARIABLE	NIVELLS
RAMEMIS	90	TOTVEHC	14
RAMITO	27	INDCOLD	2
INDSITU	4	IMDCORP	2
TIPSINI	106	INDVIDR	2
CODAGTI	287	INDINCE	2
INDTIVE	25	INDROBO	2
INDRECO	2	DANOSMAT	2
INDCONS	2	INDDAPR	2
TIPVEHI	14	INDASV	2
CNATSIN	40	TIPTECD	4
INDCIDE	4	TIPUSUD	14
INDSDMM	5	CULPOBJ	6
INDCULP	5	PERDTOT	2
INDDASE	2	TALPREF	2

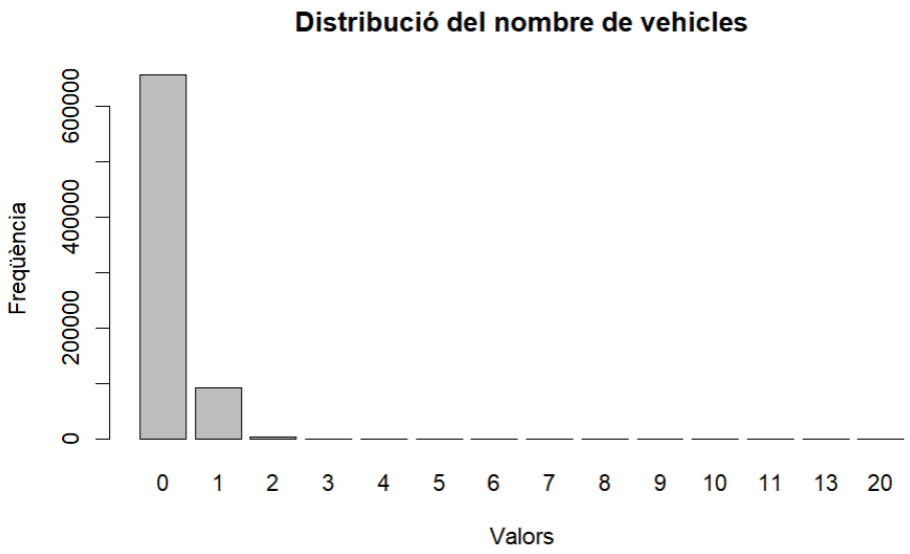
4.1.4. Creació de variables noves

A continuació creem noves variables a partir de les ja existents.

NUMVEH

La primera variable que creem ens fa la classificació del nombre de vehicles externs involucrats en intervals. Comencem mirant la gràfica que tenim a continuació per poder veure la distribució de la variable inicial i decidir el nombre d'intervals.

Figura 2: Gràfica de la distribució del nombre de vehicles externs involucrats



Un cop vista la gràfica, podem dir que no es veu del tot bé com es distribueix la variable. A continuació, farem un resum per poder veure quins són els valors que predominen.

Taula 8: Nombre i freqüència de vehicles possibles

0	**1**	**2**	**3**	**4**	**5**	**6**
656125	93664	4098	816	219	67	42
86,9%	12,4%	0,54%	0,11%	0,03%	0,009%	0,006%
7	**8**	**9**	**10**	**11**	**13**	**20**
19	5	2	1	1	2	1
0,003%	0,0007%	0,0003%	0,0001%	0,0001%	0,0003%	0,0001%

Com podem veure, la majoria dels valors es troben entre cap vehicle i 2 vehicles. A continuació farem una gràfica que només ens mostra el valor 0, 1 i 2. També en farem una de la resta de valors per veure com es distribueixen.

Figura 3: Gràfica de la distribució del nombre de vehicles externs dels valors 0, 1 i 2

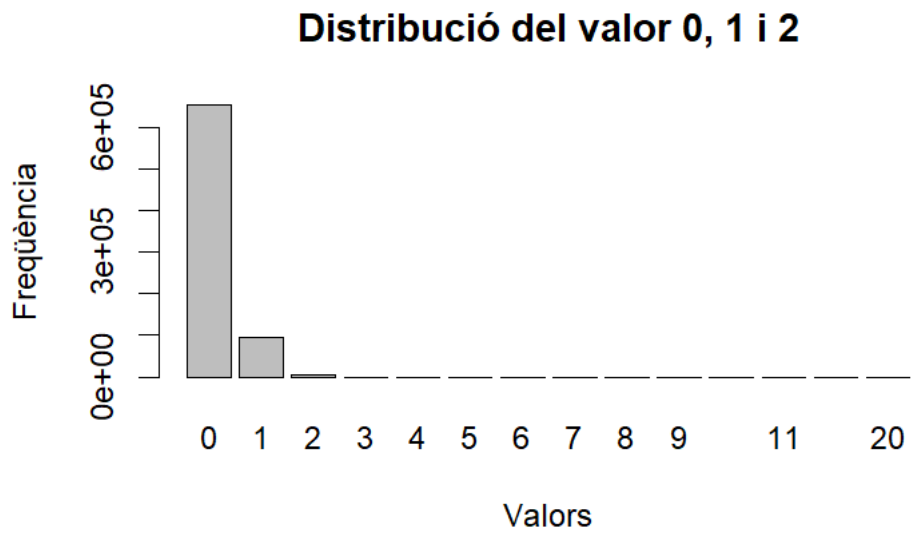
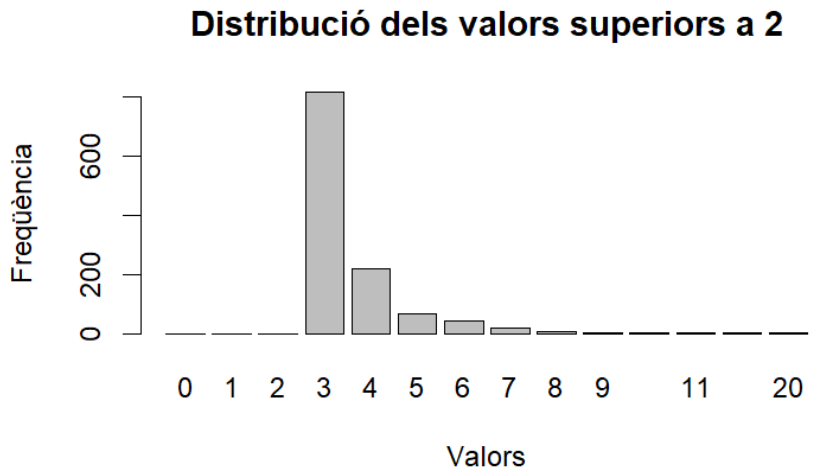


Figura 4: Gràfica de la distribució del nombre de vehicles externs dels valors superiors a 2

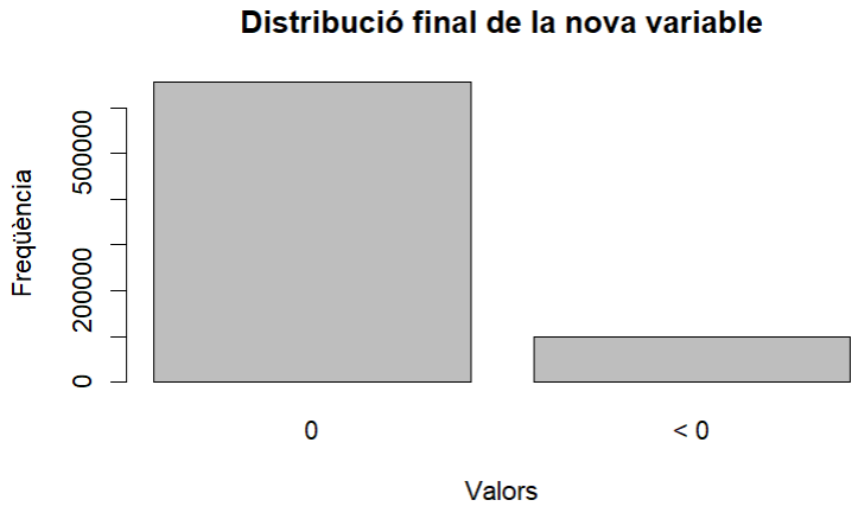


Veient les tres gràfiques hem decidit que farem 2 intervals diferents. El primer interval serà només amb el valor 0 i el segon interval quan és superior. Aquesta variable l'anomenarem NUMVEH i a continuació veurem una taula de com es distribueixen els diferents intervals.

Taula 9: Nivells de la variable NUMVEH

0	**< 0**
656125	98937
86,9%	13,1%

Figura 5: Distribució de la nova variable



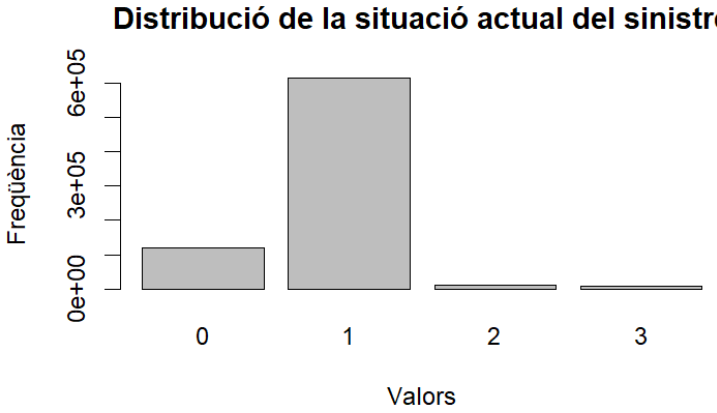
En aquesta primera gràfica podem veure la distribució final de la nova variable. Destaquem que encara predomina el valor 0.

Variable resposta (Y)

La variable INDSITU té quatre nivells que ens descriuen la fase en la qual es troba el sinistre. Si la variable val 0, ens indica que s'està notificant el sinistre. Si la variable val 1, ens diu que s'està fent la investigació i avaluació pròpia. Si la variable val 2, significa que estem fent el processament de la reclamació. I finalment, si la variable val 3, ens indica que estem a la liquidació del sinistre.

Quan ens fixem en la variable que ens dona la fase en la qual es troba el sinistre, veiem que no hi ha una proporció semblant. Les fases que realment ens interessen són la fase 0 i la fase 1, ja que són les fases en què es fan les provisions i es fa un estudi dels possibles danys. La fase 0 és la notificació del sinistre i la fase 1 és la investigació i avaluació del sinistre.

Figura 6: Distribució de la situació actual del sinistre



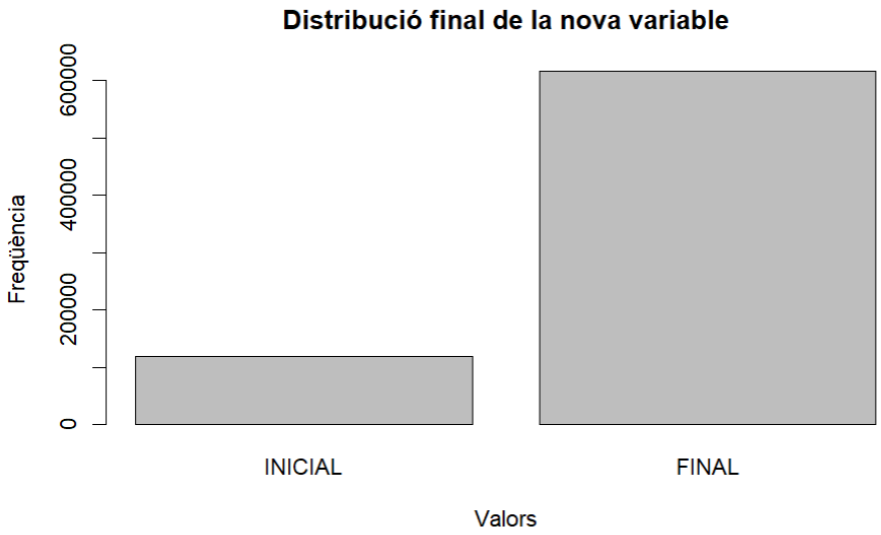
Aquesta segona variable que creem ens farà la classificació de la situació actual del sinistre en dos grups. Aquesta serà la nostra variable resposta. Com que volem que aquesta sigui binària, dividirem la variable. Amb l'ajuda de la gràfica anterior, on veiem la distribució d'aquesta variable, decidim els grups.

La fase predominant és la 1. Distribuirem la nova variable de manera que quan estigui en la fase 0, és a dir, la fase INICIAL, la variable valdrà 0 i quan estigui en la fase 1, la variable també valdrà 1. La resta de nivells els posarem com a valors nuls i els tractarem en el preprocessament, ja que no ens interessen per l'estudi. Per veure com ha quedat la nova variable veiem la taula de la distribució dels dos nivells d'aquesta.

Taula 10: Nivells i freqüència de la variable resposta

INICIAL	FINAL
119235	616740
16,2%	83,8%

Figura 7: Distribució de la nova variable resposta



A partir d'aquesta primera gràfica, podem veure la distribució final de la nova variable. Destaquem que continua predominant la fase inicial.

4.2. Software

El programa que utilitzarem durant aquest treball és el **R-studio**, ja que gràcies a aquest podem veure una àmplia varietat de tècniques estadístiques per la computació estadística, l'anàlisi de dades i el *machine learning*.

4.2.1. Paquets de Rstudio

Els paquets de necessaris per al desenvolupament són els següents:

Taula 11: Paquets Rstudio utilitzats

PAQUETS DE RSTUDIO			
arules	e1071	kableExtra	rattle
arulesViz	factoextra	lessR	rpart
caret	FactoMineR	MASS	rpart.plot
class	fpc	matrix	scales
cluster	ggplot2	mice	tidyverse
dbscan	ggpubr	naivebayes	vcd
descr	Hmisc	purrr	VIM
dplyr	ISLR	randomForest	

5. Anàlisi de la base de dades

En aquest apartat trobarem l'anàlisi de la base de dades. Començarem fent el preprocessament per poder fer l'anàlisi de la base arreglada i ens servirà per treure els valors punta i els valors buits de la base. Aquesta anàlisi dependrà del tipus de variables que tinguem. En el cas de les numèriques tindrem una anàlisi inicial i una final per veure si s'han extret els valors punta i fer la comparació. En canvi, en el cas de les variables categòriques, tindrem l'anàlisi final que és el que realment ens interessa.

Per depurar les diferents variables farem servir la distància de Cook. La distància de Cook és mètode d'anàlisi de regressió per trobar els valors atípics influents d'una base de dades. És a dir, ens ajuda a identificar els punts que afecten de manera negativa el model. Aquest model permet classificar i treure els valors punta i, finalment, eliminar tots aquells registres que tenen valors *missing*.

5.1. Preprocessament de les dades

5.1.1. Tractament dels outliers

En les diferents gràfiques que veiem a continuació veiem les distàncies de Cook de les variables numèriques per poder tractar els outliers.

Figura 8: Distàncies de Cook de la variable de la provisió inicial

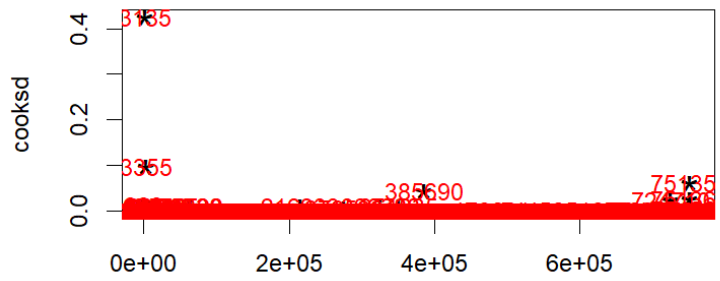


Figura 9: Distàncies de Cook de les variables de la provisió en el moment actual

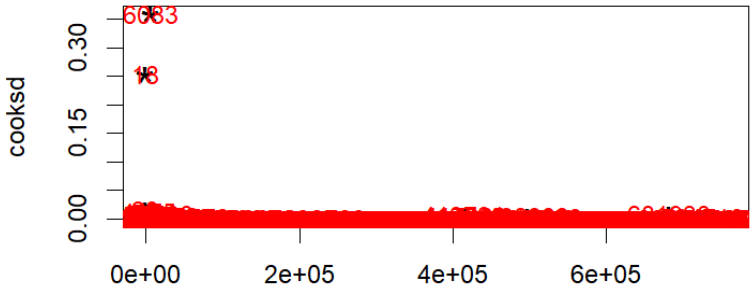


Figura 10: Distàncies de Cook de la variable de l'import base de pagaments de l'exercici actual

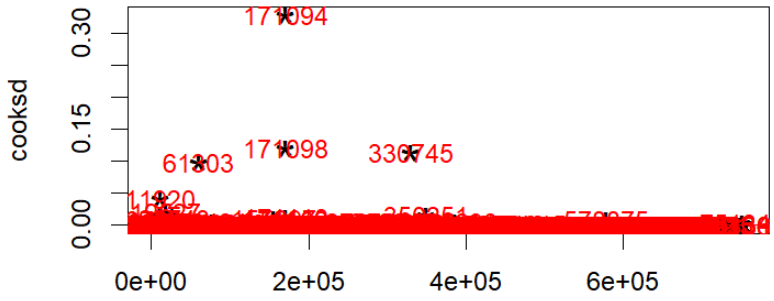
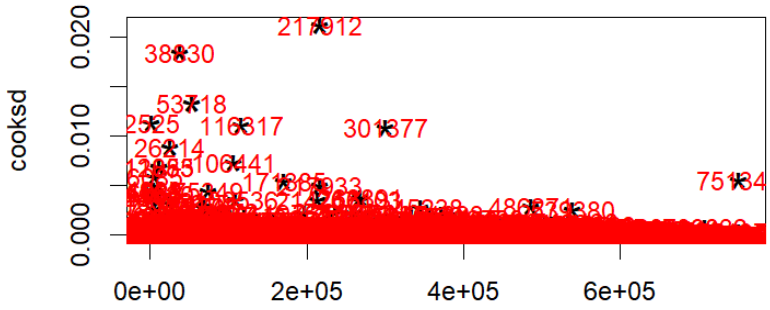


Figura 11: Distàncies de Cook de la variable de l'import IVA de pagaments de l'exercici actual



A partir d'aquestes gràfiques veiem alguns valors punta en les variables numèriques. Eliminem aquells innecessaris per l'estudi i els convertim en valors buits.

5.1.2. Tractament dels missings

Per fer un tractament dels valors nuls comencem fent una classificació dels valors *missing*. Amb aquesta classificació descartarem les variables que tinguin un percentatge molt alt de valors mancants.

Taula 12: Classificació dels missings

RAMEMIS	RAMITO	INDSITU	TIPSINI	CODAGTI	INDTIVE	INDRECO
0,00%	0,00%	0,00%	39,85%	39,85%	0,00%	0,00%
INDCONS	**TIPVEHI**	**CNATSIN**	**INDCIDE**	**INDSDMM**	**INDCULP**	**INDDASE**
0,00%	60,73%	0,00%	60,16%	89,12%	0,00%	0,00%
TOTVEHC	**INDCOLD**	**IMDCORP**	**INDVIDR**	**INDINCE**	**INDROBO**	**DANOSMAT**
0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
INDDAPR	**PROVINI**	**PROVACT**	**PAGB**	**PAGI**	**INDASV**	**TIPTECD**
0,00%	21,26%	14,73%	19,80%	15,92%	0,00%	0,85%
TIPUSUD	**CULPOBJ**	**PERDTOT**	**TALPREF**	**NUMVEH**	**Y**	
0,00%	96,72%	0,00%	0,00%	0,00%	2,53%	

Veiem que les variables amb el percentatge més alt són les següents:

- Tipologia del sinistre (TIPSINI)
- Subtipologia del sinistre (CODAGTI)
- Indicador SDM (INDSDMM)
- Culpa objectiva, només DC (CULPOBJ)
- Classe de vehicle, només autos (TIPVEHI)
- Tipus de conveni, valor per CIDE (INDCIDE)

Eliminem aquestes, i un cop ho estan, mirem l'estructura de les dades. Veiem que només tenim dos tipus de variables, les categòriques (com a factors) i les numèriques.

Figura 12: Estructura de les dades abans del preprocessament

```
'data.frame':   755062 obs. of  28 variables:
 $ RAMEMIS : Factor w/ 90 levels "100","101","102",..: 17 57 63 67 53 78 78 60 57 69 ...
 $ RAMITO  : Factor w/ 27 levels "100","110","140",..: 6 12 16 26 15 27 27 13 12 19 ...
 $ INDSITU : Factor w/ 4 levels "0","1","2","3": 3 1 3 3 1 2 1 1 1 3 ...
 $ INDTIVE : Factor w/ 25 levels "0","1","2","3",..: 1 1 1 1 1 2 3 1 1 1 ...
 $ INDRECO : Factor w/ 2 levels "N","S": 2 1 1 2 1 1 1 1 1 1 ...
 $ INDCONS : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ CNATSIN : Factor w/ 40 levels "0","1","2","3",..: 1 1 1 1 1 9 11 1 1 1 ...
 $ INDCULP : Factor w/ 4 levels "A","C","D","I": NA 1 NA 1 1 3 4 1 1 1 ...
 $ INDDASE : Factor w/ 2 levels "N","S": 1 1 1 1 1 2 1 1 1 1 ...
 $ TOTVEHC : Factor w/ 14 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ INDCOLD : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ IMDCORP : Factor w/ 2 levels "N","S": 1 1 2 2 1 2 1 1 1 1 ...
 $ INDVIDR : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 2 1 1 1 ...
 $ INDINCE : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ INDROBO : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ DANOSMAT: Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ INDDAPR : Factor w/ 2 levels "N","S": 1 1 1 1 1 2 1 1 1 1 ...
 $ PROVINI : num  0 189 0 NA NA ...
 $ PROVACT : num  NA NA 20 NA NA 0 NA NA NA NA ...
 $ PAGB    : num  NA 0 0 0 141 ...
 $ PAGI    : num  0 0 0 0 29 ...
 $ INDASV  : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ TIPTECD : Factor w/ 3 levels "BCH","INT","WEB": NA 2 NA NA NA NA 2 2 2 NA ...
 $ TIPUSUD : Factor w/ 14 levels "   ","AGE","ASI",..: 1 2 1 1 1 1 3 2 7 1 ...
 $ PERDTOT : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ TALPREF : Factor w/ 2 levels "N","S": 1 1 1 1 1 1 1 1 1 1 ...
 $ NUMVEH  : Factor w/ 2 levels "0","< 0": 1 1 1 1 1 1 1 1 1 1 ...
 $ Y       : Factor w/ 2 levels "INICIAL","FINAL": NA 1 NA NA 1 2 1 1 1 NA ...
```

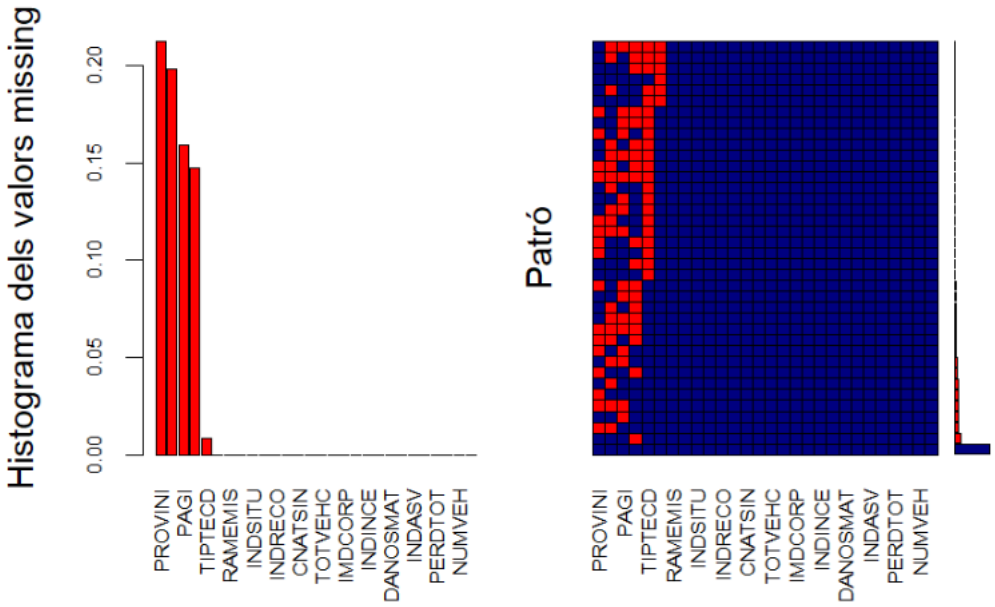
A continuació tenim la taula que ens mostra quins tants per u de valors buits hi ha en cada variable. Podem veure que les variables amb més dades mancants és PROVACT amb un 2,579% i, a continuació, PAGI amb un 1,807%. La resta de variables que no veiem és perquè el percentatge és 0.

Taula 13: Tants per u dels missings

VARIABLE	COUNT
PROVINI	0,21257327213
PAGB	0,19802877115
PAGI	0,15917765693
PROVACT	0,14725413277
TIPTECD	0,00852512774
INDCULP	0,00001191955

Aquesta taula, la representem gràficament. Veiem que les barres més llargues corresponen a les variables amb el percentatge més alt de la taula. En les dues gràfiques podem veure on hi ha més i menys *missings* en totes les diferents variables.

Figura 13: Gràfiques del patró que segueixen els valors missing



Un cop hem vist quines variables hem de tractar, sabem que els percentatges de les dades mancants és petit en tots els casos, a més a més, de tenir una base de dades amb molts registres. Per tant, eliminant tots els registres que continguin valors buits en tindrem prou per tractar-los. Així, tindrem la base de dades preparada.

Un cop els tenim eliminats, fem un sumari de les dades per veure que ja estan processades. En aquest resum podem veure els diferents nivells de cada variable categòrica amb el nombre de registres que té cada un. Per les variables numèriques veiem el resum numèric amb el càlcul del mínim, mitjana i els quatre quadrants (primer, mediana, tercer i màxim).

Figura 14: Sumari després del preprcessament de les dades

```
   RAMEMIS           RAMITO       INDSITU       INDTIVE       INDRECO    INDCONS       CNATSIN
685    :156564   691    :156718   0:  9826   0      :287392   N:407249   N:415756   0      :287394
800    :128492   800    :128492   1:406058   1      : 73314   S:  8635   S:   128   23     :110558
690    : 94234   693    : 94234   2:     0   2      : 14031                         16     : 14739
689    : 12833   696    : 12836   3:     0   3      : 11894                         10     :  1535
688    :  5841   695    :  5967              6      :  9668                         1      :   728
280    :  5540   252    :  5573              7      :  8864                         35     :   442
(Other): 12380   (Other): 12064              (Other): 10721                         (Other):   488
INDCULP     INDDASE       TOTVEHC       INDCOLD    INDCORP    INDVIDR    INDINCE    INDROBO    DANOSMAT
A:287827   N:413889   0      :414821   N:415214   N:408690   N:404671   N:415884   N:415867   N:414247
C:  1967   S:  1995   1      :   995   S:   670   S:  7194   S: 11213   S:     0   S:    17   S:  1637
D:    57              2      :    60
I:126033              3      :     4
                      4      :     3
                      5      :     1
                      (Other):     0
INDDAPR       PROVINI          PROVACT             PAGB              PAGI         INDASV     TIPTECD
N:395836   Min.   :-144.4   Min.   :-14.9300   Min.   :-225.00   Min.   :-16.87   N:304581   BCH:120888
S: 20048   1st Qu.:  91.4   1st Qu.:  0.0000   1st Qu.:  42.93   1st Qu.:  6.41   S:111303   INT:294973
           Median : 250.0   Median :  0.0000   Median :  66.26   Median : 11.59              WEB:    23
           Mean   : 234.0   Mean   :  0.2745   Mean   : 103.96   Mean   : 11.82
           3rd Qu.: 336.0   3rd Qu.:  0.0000   3rd Qu.: 134.09   3rd Qu.: 15.02
           Max.   : 650.0   Max.   : 50.0000   Max.   : 489.46   Max.   : 39.49

   TIPUSUD       PERDTOT    TALPREF    NUMVEH             Y
AGE    :186428   N:415799   N:415576   0  :414821   INICIAL:  9826
ASV    :111134   S:    85   S:   308   < 0:  1063   FINAL  :406058
CAT    : 99849
ASL    :  9639
SUC    :  3940
CST    :  3415
(Other):  1479
```

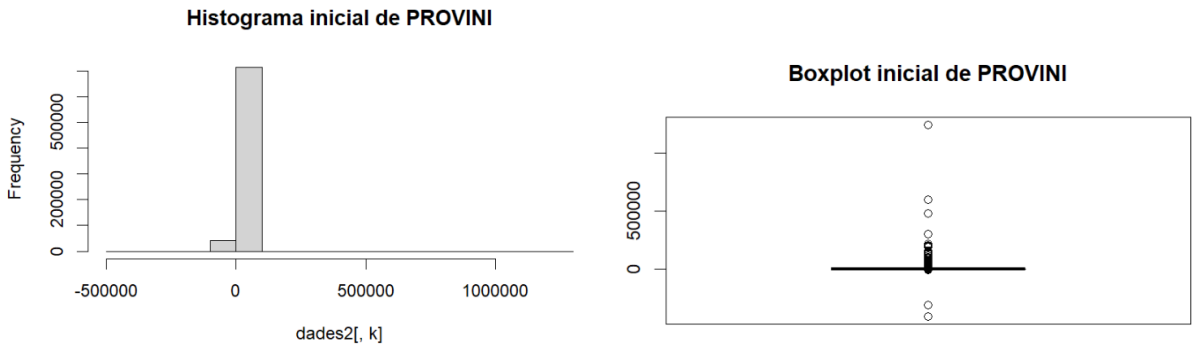
Amb aquest resum podem veure que no tenim valors buits i afirmar que tenim la base de dades preparada per fer el nostre estudi.

5.1.3. Anàlisi de les variables numèriques

En aquest subapartat analitzem les variables numèriques. Per cada una d'elles tenim un histograma i un *boxplot* d'abans i de després de depurar les dades.

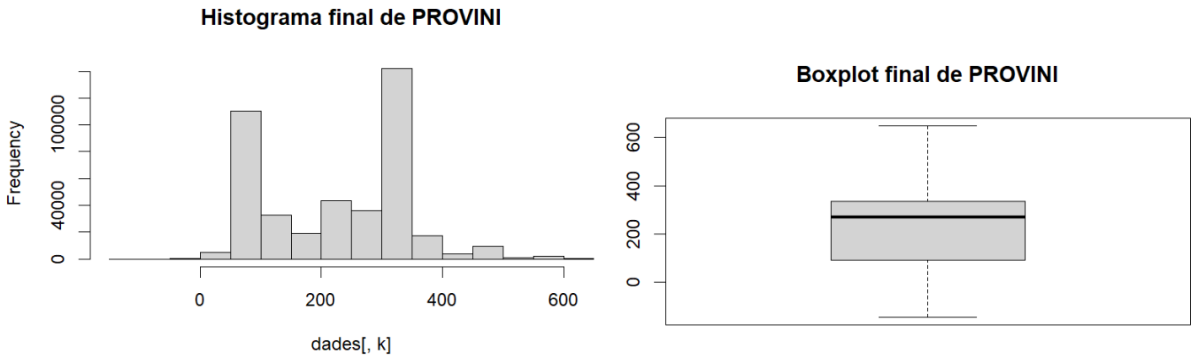
PROVINI: Provisió inicial

Figura 15: Histograma i boxplot inicials de PROVINI



Per començar, ens fixem en l'histograma el qual no ens mostra bé la distribució que segueix aquesta variable. Veiem que només tenim dues columnes i aquestes es troben al centre. Per altra banda, gràcies al *boxplot* podem veure que aquesta conté molts valors punta, tant per sobre com per sota, els hem descartat amb les distàncies de Cook. A continuació veiem aquestes mateixes gràfiques en el moment actual, és a dir, amb les dades tractades.

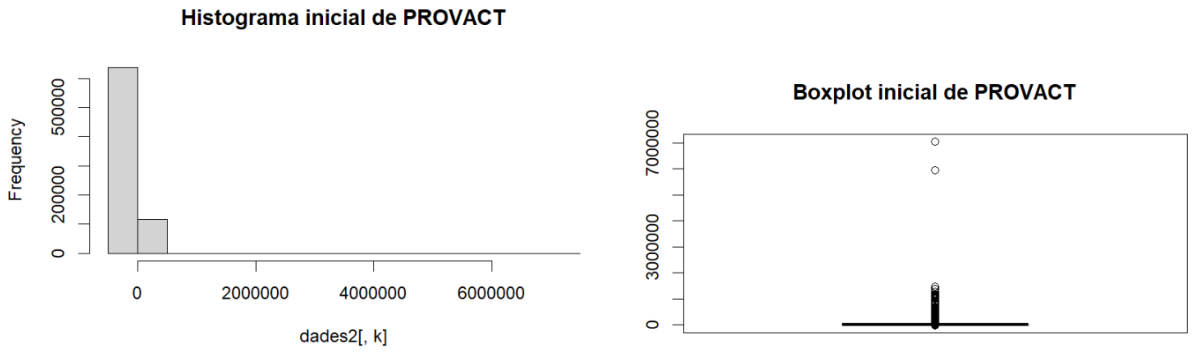
Figura 16: Histograma i boxplot finals de PROVINI



En fixar-nos en l'histograma veiem que ha millorat, ja que ara ens apareixen més barres que abans. Això ens diu que els valors punta ja són fora. Amb el *boxplot* podem veure que no tenim ni un valor punta i que tenim ben representada la variable.

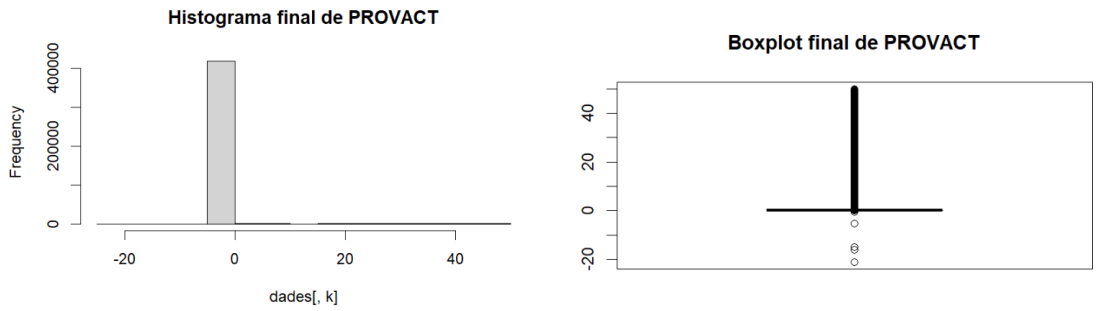
PROVACT: Provisió actual

Figura 17: Histograma i boxplot inicials de PROVACT



Ens fixem en l'histograma, el qual, no ens mostra bé la distribució que segueix aquesta variable. Veiem que, només tenim dues columnes i, aquestes, es troben al costat esquerre. Per altra banda, gràcies al *boxplot* podem veure que aquesta conté molts valors punta, tant per sobre com per sota, els quals els podrem descartar. A continuació veiem aquestes mateixes gràfiques un cop s'hagin tractat les dades.

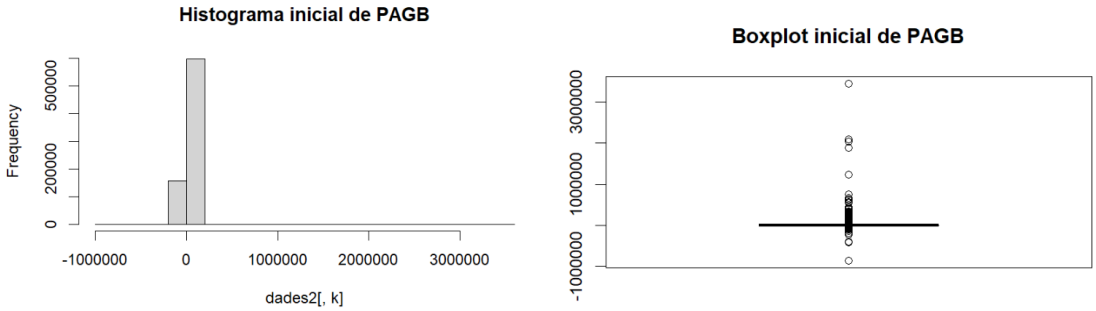
Figura 18: Histograma i boxplot finals de PROVACT



Ara veiem que l'histograma ha millorat una mica, ja que ens apareixen més barres que abans i no les tenim només a l'esquerra. Això ens diu que els valors punta innecessaris ja són fora i, els que queden, hem considerat que tenien sentit per l'estudi. Aquests valors els podem veure representats de manera més clara en la gràfica de caixa. Veiem que en tenim més per sobre que per sota.

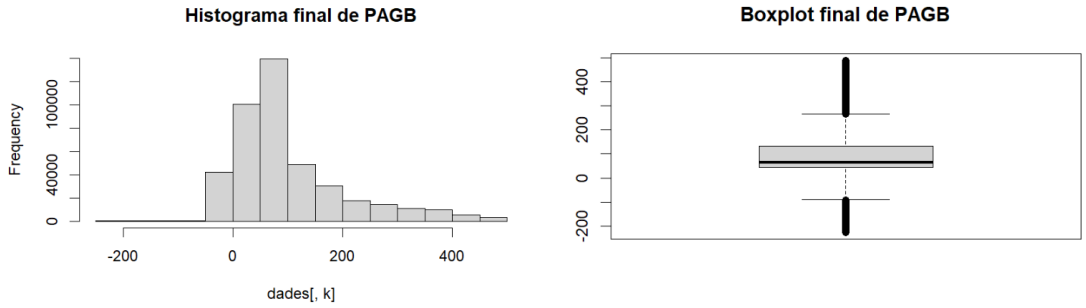
PAGB: Import base de pagaments d'exercici actual

Figura 19: Histograma i boxplot inicials de PAGB



En l'histograma d'aquesta variable podem observar que no ens mostra bé la distribució que segueix. Veiem que només tenim dues columnes i aquestes es troben al centre. Per altra banda, gràcies al *boxplot* podem veure que aquesta conté molts valors outliers, tant per sobre com per sota, els quals els podrem descartar. A continuació veiem aquestes mateixes gràfiques un cop s'hagi fet el preprocessament.

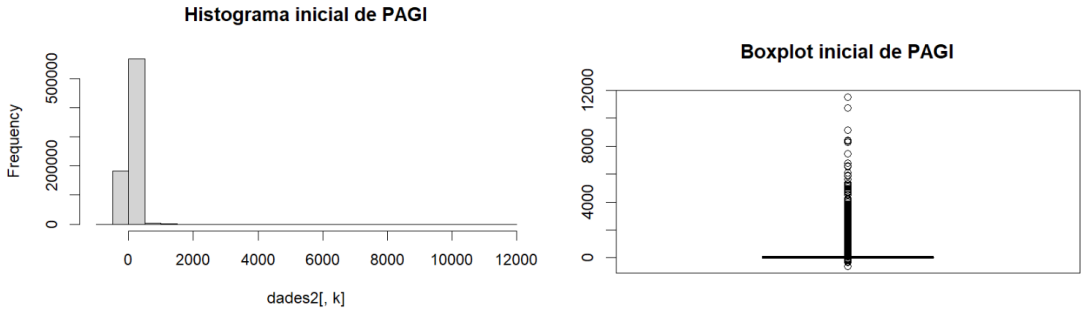
Figura 20: Histograma i boxplot finals de PAGB



Ara veiem que l'histograma ha millorat, ja que ara ens apareixen més barres que abans i no les tenim només a l'esquerra. Això ens diu que els valors punta innecessaris ja són fora i, els que queden, hem considerat que tenien sentit per l'estudi. Aquests valors els podem veure representats de manera més clara en la gràfica de caixa. Veiem que en tenim més per sobre que per sota.

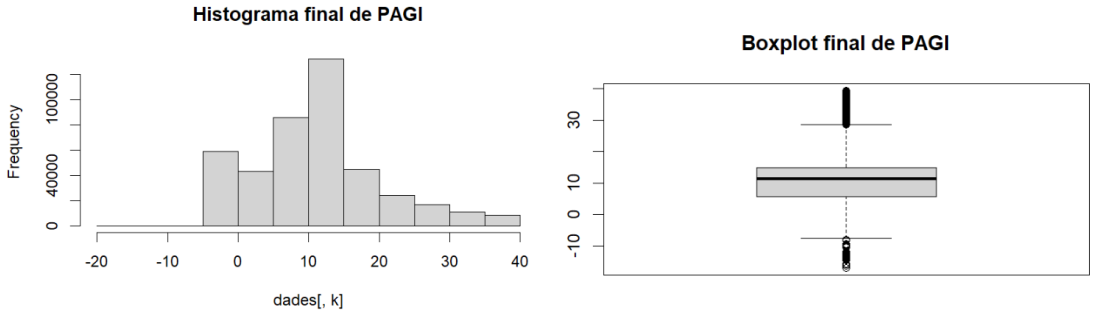
PAGI: Import IVA de pagaments d'exercici actual

Figura 21: Histograma i boxplot inicials de PAGI



Mirem l'histograma que no ens mostra bé la distribució que segueix aquesta variable. Veiem que només tenim dues columnes i aquestes es troben al costat esquerre. Per altra banda, gràcies al *boxplot* podem veure que aquesta conté molts valors punta per sobre els quals els podrem descartar. A continuació veiem aquestes mateixes gràfiques un cop s'hagin tractat els valors punta.

Figura 22: Histograma i boxplot finals de PAGI



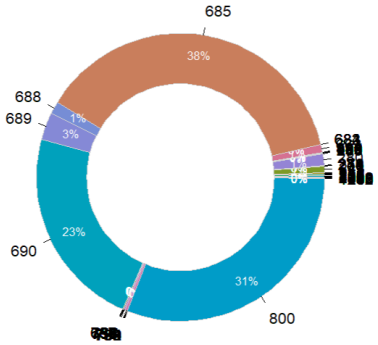
A continuació, veiem una millora en l'histograma perquè ara tenim més barres que abans i no només a l'esquerra. Això ens indica que els valors innecessaris han desaparegut, i els valors restants creiem que són significatius per a l'estudi. Podem veure que aquests són clarament representats en el diagrama de caixa. Veiem que en tenim tant per dalt com per baix.

5.1.4. Anàlisi de les variables categòriques

En aquest subapartat analitzem les variables categòriques de la base de dades. Per cada una d'aquestes, mostrem una gràfica de sectors de després de depurar les dades. No farem l'abans, ja que no ens aportarà res. En les variables que tinguem menys de 10 categories també hi haurà una taula de freqüències.

RAMEMIS: Ram del Grup Catalana Occident

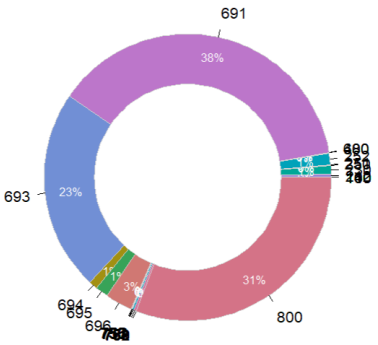
Figura 23: Gràfica de sectors de la variable RAMEMIS



En aquesta gràfica tenim la variable dels rams de la companyia. Veiem que predomina el ram 685 i a continuació el de 800. En aquest cas no es veuen tots els nivells, ja que en tenim 90 de diferents.

RAMITO: Ram de gestió

Figura 24: Gràfica de sectors de la variable RAMITO



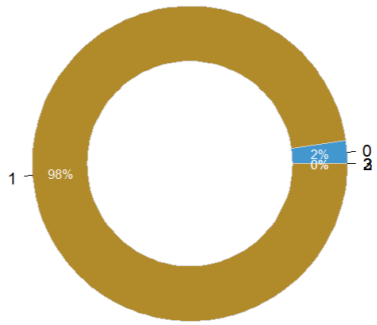
En aquesta gràfica de sectors trobem la variable dels rams de gestió. Veiem que predomina el ram de gestió 691 i a continuació el de 800. En aquest cas no es veuen tots els nivells, ja que en tenim 27 de diferents.

INDSITU: Situació actual del sinistre

Taula 14: Freqüència de la variable INDSITU

	FREQÜENCIA	PERCENTATGE
0	9826	2,362678%
1	406058	97,637322%
2	0	0%
3	0	0%
TOTAL	415884	100%

Figura 25: Gràfica de sectors de la variable INDSITU



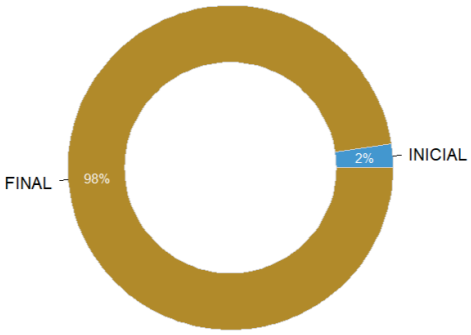
En aquesta taula trobem els quatre nivells de la variable, que ens diu la situació del sinistre. Veiem que una mica més del 97% dels sinistres es troben a la fase 1. Els percentatges d'aquesta variable els tenim de manera més visual en la gràfica de sectors. La fase amb menor registres és el 2. A continuació veiem la variable resposta que hem construït a partir d'aquesta.

Y: Variable resposta (Situació actual del sinistre)

Taula 15: Freqüència de la variable resposta

	FREQÜENCIA	PRECENTATGE
INICIAL	9826	2,362678%
FINAL	406058	97,637322%
TOTAL	415884	100%

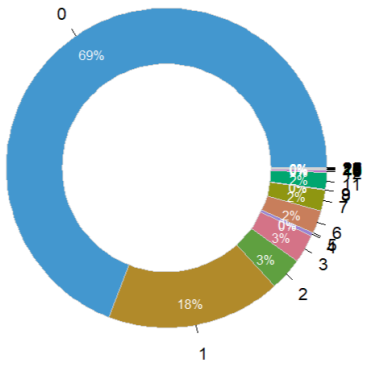
Figura 26: Gràfica de sectors de la variable Y



A l'haver agrupat la variable anterior, trobem dos nivells: l'inicial i el final. Veiem, tant a la taula com a la gràfica de sectors, que predomina la fase final.

INDTIVE: Categoria

Figura 27: Gràfica de sectors de la variable INDTIVE



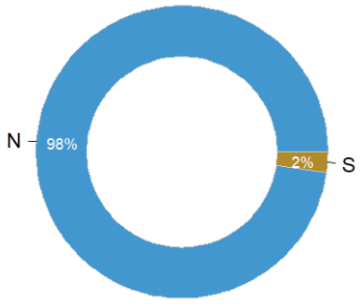
En aquesta gràfica de sectors trobem les diferents categoria. Podem veure que la zero és la que predomina amb un 69% i continua la u amb un 18%. La resta de categories segueixen amb percentatges més baixos com podem veure a la gràfica.

INDRECO: Indicador del sinistre recobrable

Taula 16: Freqüència de la variable INDRECO

	FREQÜÈNCIA	PERCENTATGE
N	407249	97,9237%
S	8635	2,0763%
TOTAL	415884	100%

Figura 28: Gràfica de sectors de la variable INDRECO



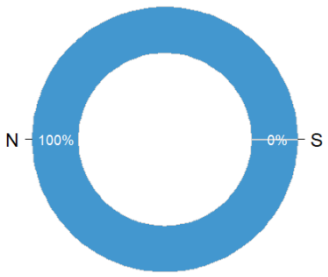
En aquesta gràfica de pastís tenim una variable binària. Aquesta ens indica si el sinistre és recobrable o no. Podem veure que la majoria dels sinistres no són recobrables, un 98% d'elles.

INDCONS: Indicador del sinistre consorciable

Taula 17: Freqüència de la variable INDCONS

	FREQÜÈNCIA	PERCENTATGE
N	415756	99,9692222%
S	128	0,0307778%
TOTAL	415884	100%

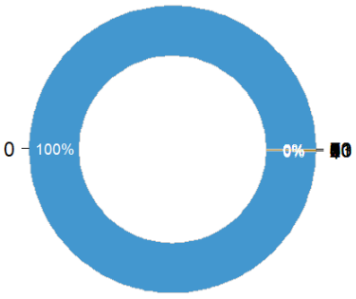
Figura 29: Gràfica de sectors de la variable INDCONS



Aquesta variable és un indicador de si el sinistre és consorciable, també és binària. Podem veure que pràcticament tots els sinistres tenen l'indicador negatiu de manera que no hi ha indicador de què el sinistre sigui consorciable.

TOTVEHC: Nombre total de vehicles externs involucrats

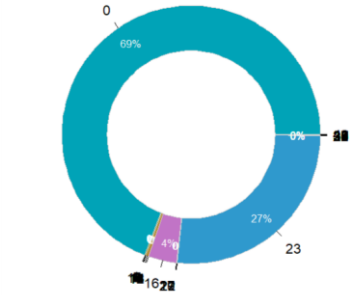
Figura 30: Gràfica de sectors de la variable TOTVEHC



En la gràfica de sectors podem veure que a tots els sinistres pràcticament no hi ha cap vehicle extern involucrat. Els altres nivells tenen percentatges molt baixos i, per tant, es veuen representants amb un valor arrodonit al 0%.

CNATSIN: Naturalesa del sinistre

Figura 31: Gràfica de sectors de la variable CNATSIN



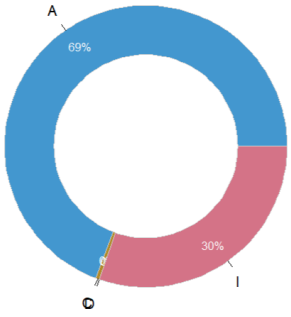
En la variable de la naturalesa del sinistre tenim 40 nivells. El que predomina és 0 amb un 69% dels registres i continua amb el 23 amb un 27% dels registres. La resta de sinistres tenen uns percentatges molt baixos i, per tant, no tenen tant pes.

INDCULP: Indicador de culpa

Taula 18: Freqüència de la variable INDCULP

	FREQÜÈNCIA	PERCENTATGE
A	287827	69,2084812%
C	1967	0,4729684%
D	57	0,0137057%
I	126033	30,3048446%
TOTAL	415884	100%

Figura 32: Gràfica de sectors de la variable INDCULP



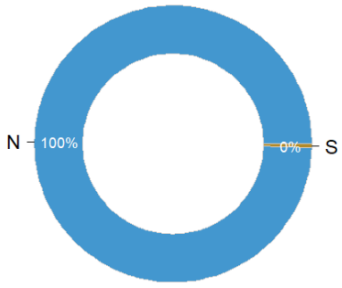
En aquesta gràfica es mostra la distribució de l'indicador de culpa. Podem veure que el 69% dels registres està dins de la categoria A. El percentatge de les categories C i D veiem que són molt petites, de manera que no arriben ni a l'1%. Per últim, la categoria I té un 30% dels registres. Aquests percentatges els podem comprovar en la taula, ja que els de la gràfica estan arrodonits a les unitats.

INDDASE: Indicador de danys del vehicle

Taula 19: Freqüència de la variable INDDASE

	FREQÜÈNCIA	PERCENTATGE
N	413889	99,5202989%
S	1995	0,4797011%
TOTAL	415884	100%

Figura 33: Gràfica de sectors de la variable INDDASE



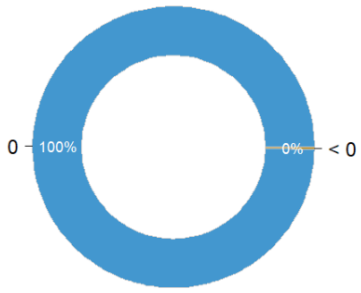
La gràfica de sectors ens mostra que tots els sinistres no tenen cap indicador de danys del vehicle. Però quan mirem la taula veiem que només un 0,48% dels sinistres tenen aquest indicador.

NUMVEH: Nombre de vehicles

Taula 20: Freqüència de la variable NUMVEH

	FREQÜÈNCIA	PERCENTATGE
0	414821	99, 7443999%
< 0	1063	0,2556001%
TOTAL	415884	100%

Figura 34: Gràfica de sectors de la variable NUMVEH



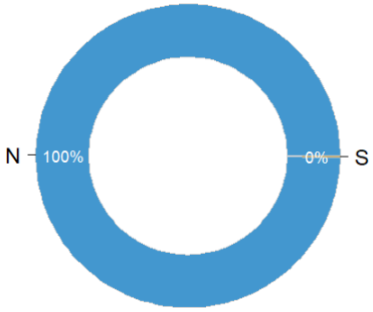
En la taula podem veure que pràcticament en tots els sinistres no hi ha altres vehicles involucrats i només un 0,256% d'aquests hi havia un o més vehicles involucrats. Aquesta taula la podem veure de manera més clara en la gràfica de sectors. Veiem com es distribueix la variable i podem afirmar que pràcticament tots els sinistres no tenen cap altre vehicle involucrat.

INDCOLD: Indicador de col·lisió

Taula 21: Freqüència de la variable INDCOLD

	FREQÜÈNCIA	PERCENTATGE
N	415214	99,8388974%
S	670	0,1611026%
TOTAL	415884	100%

Figura 35: Gràfica de sectors de la variable INDCOLD



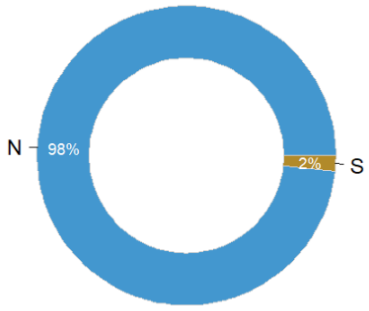
En aquesta taula podem veure que pràcticament en tots els sinistres no hi ha indicador de col·lisió i només un 0,161% d'aquests en té. Aquesta taula la podem veure de manera més clara en la gràfica de sectors. Observem com es distribueix la variable i afirmem que pràcticament tots els sinistres no tenen indicador de col·lisió.

IMDCORP: Indicador de lesionats

Taula 22: Freqüència de la variable IMDCORP

	FREQÜÈNCIA	PERCENTATGE
N	408690	98,270191%
S	7194	1,729809%
TOTAL	415884	100%

Figura 36: Gràfica de sectors de la variable IMDCORP



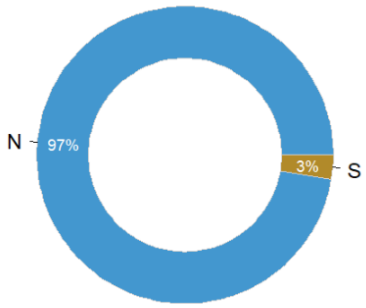
Gràcies a la taula i la gràfica podem veure que en la majoria dels sinistres no hi ha indicador de lesionats. Encara que l'indicador de què n'hi ha no val 0. Aquesta variable té un 1,73% de sinistres amb l'indicador de què hi ha hagut lesionats.

INDVIDR: Indicador de vidre

Taula 23: Freqüència de la variable INDCIDR

	FREQÜÈNCIA	PERCENTATGE
N	404671	97,303815%
S	11213	2,696185%
TOTAL	415884	100%

Figura 37: Gràfica de sectors de la variable INDVIDR



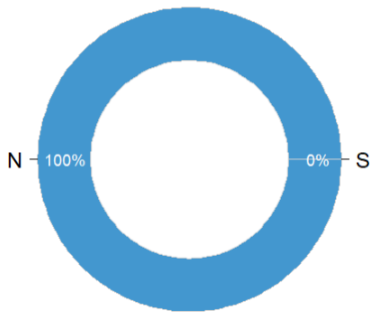
Tant amb la taula com amb la gràfica podem afirmar que la gran majoria dels sinistres no tenien indicador de vidre, és a dir, que no hi ha hagut trencament d'aquest. Encara que s'ha de tenir en compte que un petit percentatge de 2,696% si ha tingut indicador i, per tant, sí que hi ha hagut trencament de vidre.

INDINCE: Indicador d'incendi

Taula 24: Freqüència de la variable INDINCE

	FREQÜÈNCIA	PERCENTATGE
N	415884	100%
S	0	0%
TOTAL	415884	100%

Figura 38: Gràfica de sectors de la variable INDINCE



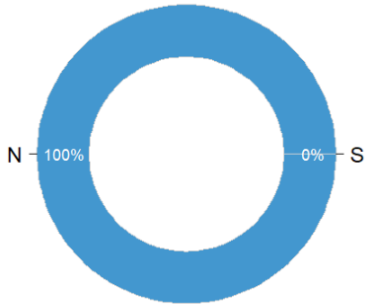
En el cas de l'indicador d'un incendi podem veure que en el 100% dels casos no n'hi ha hagut, per tant, cap sinistre té aquest indicador. Ho podem veure tant en la taula com en la gràfica de sectors.

INDROBO: Indicador de robatori

Taula 25: Freqüència de la variable INDROBO

	FREQÜÈNCIA	PERCENTATGE
N	415867	99,9959123%
S	17	0,0040877%
TOTAL	415884	100%

Figura 39: Gràfica de sectors de la variable INDROBO



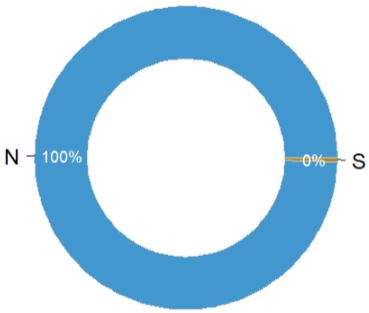
Amb la gràfica de sectors podríem pensar que aquest torna a ser un indicador negatiu amb el 100% dels sinistres però, quan ens fixem en la taula, veiem que això no és així, ja que hi ha 17 sinistres en què l'indicador era positiu i, per tant, sí que hi ha hagut un incendi.

DANOSMAT: Indicador de danys materials

Taula 26: Freqüència de la variable DANOSMAT

	FREQÜÈNCIA	PERCENTATGE
N	414247	99,6063806%
S	1637	0,3936194%
TOTAL	415884	100%

Figura 40: Gràfica de sectors de la variable DANOSMAT



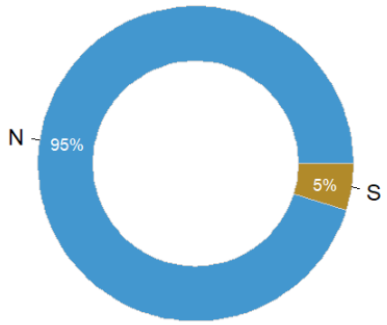
Gràcies a la taula i a la gràfica podem dir que la gran majoria dels sinistres no tenien indicador de danys materials. Encara que s'ha de tenir en compte que un petit percentatge de 0,394% sí que ha tingut aquest indicador i, per tant, sí que hi ha hagut danys materials en alguns dels sinistres.

INDDAPR: Indicador de danys propis

Taula 27: Freqüència de la variable INDDAPR

	FREQÜÈNCIA	PERCENTATGE
N	395836	95,179425%
S	20048	4,820575%
TOTAL	415884	100%

Figura 41: Gràfica de sectors de la variable INDDAPR



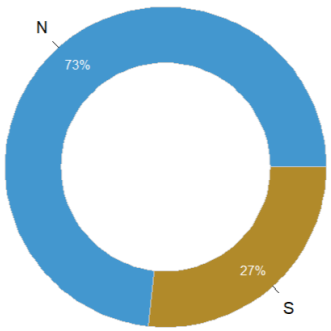
Amb la informació que ens dóna aquesta taula, juntament amb la gràfica, podem veure que, en la majoria dels sinistres, no hi ha indicador de què s'hagin produït danys propis. Encara que hem de tenir en compte que l'indicador de què sí que s'hagi produït no val 0. Aquesta variable té un 4,82% de sinistres amb l'indicador de què s'han produït danys propis.

INDASV: Indicador d'assistència de *viatge*

Taula 28: Freqüència de la variable INDASV

	FREQÜÈNCIA	PERCENTATGE
N	304581	73,23701%
S	111303	26,76299%
TOTAL	415884	100%

Figura 42: Gràfica de sectors de la variable INDASV



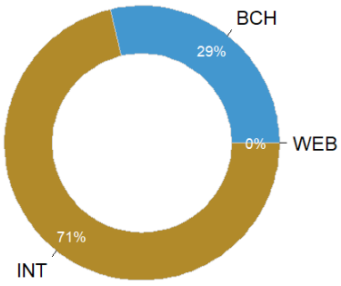
Quan ens fixem en l'indicador d'assistència de viatge veiem que, tant amb la gràfica com amb la taula, un gran percentatge dels sinistres tenen l'indicador negatiu però, no el 100%. Observem que el 26,763% dels sinistres han necessitat assistència de viatge.

TIPTECD: Tipus de declaració del sinistre de tecnologia

Taula 29: Freqüència de la variable TIPTECD

	FREQÜÈNCIA	PERCENTATGE
BCH	120888	29,0677208%
INT	294973	70,9267488%
WEB	23	0,0055304%
TOTAL	415884	100%

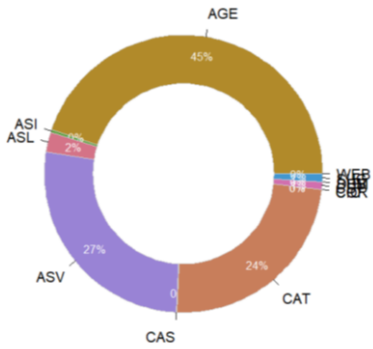
Figura 43: Gràfica de sectors de la variable TIPTECD



Aquesta variable ens informa de com ha sigut la declaració del sinistre, és a dir, quina tecnologia s'ha utilitzat. Observem que el 70,927% dels sinistres han sigut declarats a través de la intranet i que el 29,068% s'han declarat a través de BATCH. Només un percentatge molt petit dels sinistres, 0,0055% per ser exactes, han sigut declarats a partir de l'internet.

TIPUSUD: Tipus de declaració del sinistre d'usuari

Figura 44: Gràfica de sectors de la variable TIPUSUD



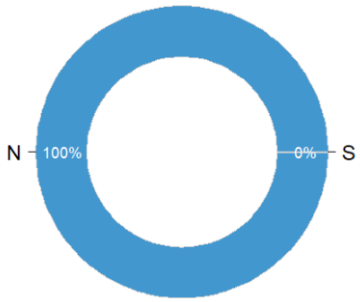
En la gràfica de sectors podem observar com predomina la declaració a través d'un agent amb un 45%, a continuació trobem els que s'han declarat a través de l'assistència asitur amb un 27% i, després, els que s'<han declarat contactant amb el centre amb un 24%. La resta de categories les trobem representades amb percentatges iguals o inferiors al 2%, per tant pocs sinistres s'han declarat a través d'aquests mitjans.

PERDTOT: Pèrdua total del vehicle assegurat

Taula 30: Freqüència de la variable PERDTOT

	FREQÜÈNCIA	PERCENTATGE
N	415799	99,9795616%
S	85	0,0204384%
TOTAL	415884	100%

Figura 45: Gràfica de sectors de la variable PERDTOT



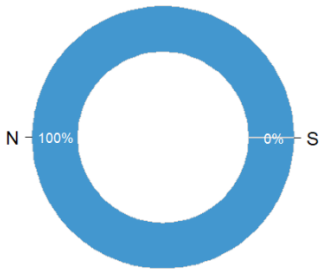
Gràcies a la informació que ens dóna la taula, juntament amb la gràfica, podem veure que en la majoria dels sinistres no hi ha hagut una pèrdua total del vehicle que s'ha assegurat. Encara que, hem de tenir en compte que, en alguns casos, sí que s'ha produït la pèrdua. Aquesta variable té un 0,02% de sinistres amb pèrdua total del vehicle. Aquest percentatge és tan petit que només 85 dels casos és positiu.

TALPREF: Vehicle assegurat reparat pel taller preferent

Taula 31: Freqüència de la variable TALPREF

	FREQÜÈNCIA	PERCENTATGE
N	415576	99,9259409%
S	308	0,0740591%
TOTAL	415884	100%

Figura 46: Gràfica de sectors de la variable TALPREF



Amb la informació que ens dóna aquesta taula i la gràfica podem veure que, en la majoria dels sinistres, no s'ha reparat el vehicle en el taller preferent per la companyia. Encara que, hem de tenir en compte que, en alguns casos, sí que s'ha acudit en aquest. Aquesta variable té un 0,07406% de sinistres que sí que han reparat el vehicle en aquest taller.

6. Aplicació dels models

En aquest apartat aplicarem els models de classificació. Avaluarem el rendiment d'aquests models comparant els diferents paràmetres. Aquests seran l'exactitud, la precisió, la sensibilitat, l'especificitat i mirarem l'espai sota la corba de ROC. Estudiarem els següents models:

- Arbres de decisió
- Naive Bayes
- Regressió logística

Aquests models han sigut escollits per les similituds que hi ha entre ells. Els tres són algoritmes de classificació supervisada. Això, significa que, necessiten unes dades d'entrenament etiquetats prèviament per aprendre a com realitzar prediccions. A més a més, els tres són models escalables, és a dir, que podem afrontar grans quantitats de dades i són relativament ràpids d'entrenar i de fer prediccions.

Hem començat dividint la base de dades en dos conjunts. El primer d'aquests és el d'entrenament que utilitza una mostra aleatòria d'aproximadament el 66,67% dels registres. Farem servir aquesta mostra pels diferents models de classificació. Aquest conjunt tindrà 277256 registres. El segon conjunt que tindrem és el de prova que té el 33,33% restant. Aquest conjunt tindrà 138628 registres.

Abans d'aplicar el model, hem triat una sèrie de variables que podrien tenir a veure amb la mala classificació de la nostra variable resposta. Aquestes seran les següents:

- Fase en la que es troba el sinistre (Y)
- Provisió actual del sinistre (PROVACT)
- Provisió inicial del sinistre (PROVINI)
- Indicador del sinistre recobrable (INDRECO)
- Indicador de lesionats (IMDCORP)
- Indicador de culpa (INDCULP)
- Indicador de vidre (INDVIDR)
- Tipus de declaració del sinistre de tecnologia (TIPTECD)

Aquestes variables les farem per tots els algoritmes, de manera que, podrem comparar els diferents models i triar el que millor s'adapti a les dades.

6.1. Arbres de decisió

El primer model que hem dut a terme és el dels arbres de decisió. Amb la funció *rpart()* agafem la variable resposta, és a dir, la fase en la qual es troba el sinistre, i set variables predictores. Les variables predictores són PROVACT, PROVINI, INDRECO, IMDCORP, INDCULP, INDVIDR i TIPTECD. Apliquem el model i obtenim els resultats amb les dades d'entrenament i amb les dades de prova.

6.1.1. Arbres de decisió amb les dades d'entrenament

Comencem aplicant el model d'arbres de decisió amb les dades de prova fixant-nos en els paràmetres de complexitat o el CP. Aquest paràmetre ens dirà la complexitat que representa l'arbre. Veiem que els dos paràmetres valen 0,9037041 i 0,01. També veiem que la variable més important per construir l'arbre és PROVACT, ja que és la que millor defineixi els nivells de la variable resposta i ens definirà els nodes fills.

Figura 47: Model d'Arbre de decisió amb les dades d'entrenament

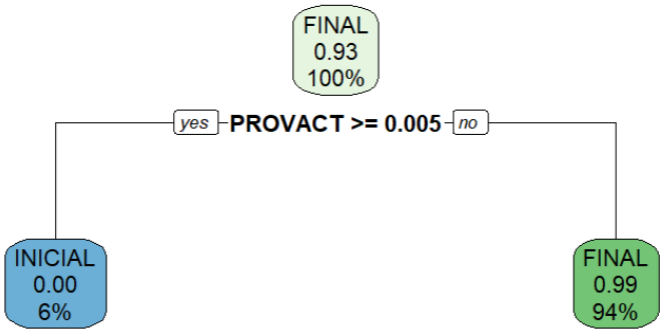
```
Call:
rpart(formula = Y ~ PROVACT + PROVINI + INDRECO + IMDCORP + INDCULP +
    INDVIDR + TIPTECD, data = dataTrain2, method = "class", parms = list(split = "information"))
  n=269386 (7870 observations deleted due to missingness)

         CP nsplit rel error    xerror        xstd
1 0.9037041      0 1.0000000 1.0000000 0.007085544
2 0.0100000      1 0.0962959 0.0962959 0.002271029

Variable importance
PROVACT
    100

Node number 1: 269386 observations,    complexity param=0.9037041
  predicted class=FINAL    expected loss=0.06884916  P(node) =1
    class counts: 18547 250839
   probabilities: 0.069 0.931
  left son=2 (16765 obs) right son=3 (252621 obs)
  Primary splits:
      PROVACT < 0.005    to the right, improve=56888.140, (0 missing)
      INDRECO splits as  RL,          improve= 5113.894, (0 missing)
      IMDCORP splits as  RL,          improve= 4499.165, (0 missing)
      PROVINI < 966.72   to the right, improve= 4010.464, (0 missing)
      INDCULP splits as  LLLR,        improve= 2558.520, (5 missing)
  Surrogate splits:
      PROVINI < 15013.15 to the right, agree=0.938, adj=0.001, (0 split)

Node number 2: 16765 observations
  predicted class=INICIAL  expected loss=0.0001192962  P(node) =0.06223412
    class counts: 16763     2
   probabilities: 1.000 0.000

Node number 3: 252621 observations
  predicted class=FINAL    expected loss=0.007061962  P(node) =0.9377659
    class counts:  1784 250837
   probabilities: 0.007 0.993
```

A partir del model obtindrem l'arbre de decisió. Aquest podem veure que té un total de 3 nodes. El node arrel ens representa tots els registres de la mostra. I l'últim node ens representa les classes predites.

Podem veure que el node 1 té 269386 observacions en total. I que es divideix en dos nodes fills. El node 2 que té 16765 registres i el node 3 que en té 252621.

Destaquem que en el node fill 3, les probabilitats ens mostren que, la majoria dels registres, es troba a la fase FINAL. Podem veure que aproximadament és un 99,3% dels casos. En canvi, en el node fill 2, pràcticament tots els casos estan en la fase INICIAL.

Figura 48: Arbre de decisió amb les dades d'entrenament



Per poder veure com avalua aquest model, calculem la seva matriu de confusió i els paràmetres de decisió per veure si és un bon model.

Figura 49: Matriu de confusió amb les dades d'entrenament

```
Confusion Matrix and Statistics
                Reference
Prediction    INICIAL     FINAL
  INICIAL       16763         2
    FINAL        1784    250837
```

En la matriu podem veure si van predir bé, segons les classes de la variable resposta. Veiem que ha classificat correctament 16763 sinistres com a fase INICIAL (positius de veritat) i que ha classificat correctament 250837 sinistres com a fase FINAL (negatius de veritat). Per altra banda, també veiem que no ha classificat bé 2 registres de la fase FINAL i que realment haurien de classificar-se a la fase INICIAL (falsos positius). I, també, que no s'han classificat bé 1784 registres que estan a la fase INICIAL que haurien de classificar-se a la fase FINAL (falsos negatius).

Gràcies a la matriu podem calcular els paràmetres que ens ajudaran a veure quin és el millor model de classificació.

El primer paràmetre de decisió calculat és l'exactitud o *accuracy.* Aquest ens diu la proporció d'instàncies que s'han classificat de manera correcta en les dades d'entrenament. En aquest cas, el paràmetre val 0,9933701 i, per tant, podem dir que té una precisió del 99,33701% en la classificació dels registres en aquest conjunt.

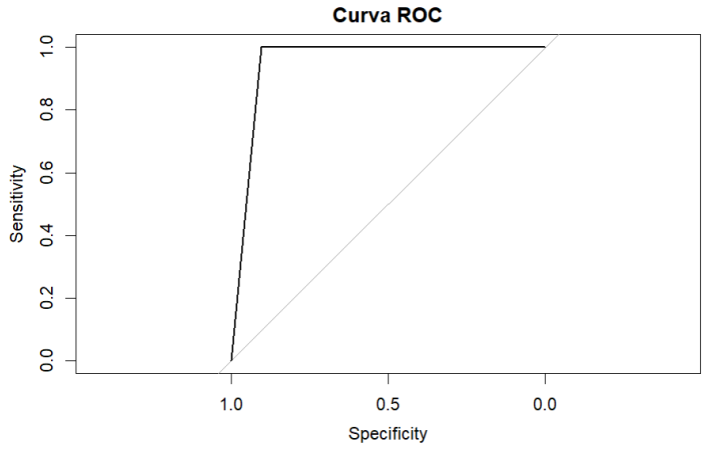
El segon paràmetre de decisió calculat és la precisió. Aquest ens diu la proporció de registres que s'han classificat correctament dins dels casos positius i, per tant, que es troben en la fase INICIAL. El valor del paràmetre en aquest cas val 0,9998807 i, així doncs, direm que té una precisió del 99,98807% en la classificació de registres dins del conjunt de positius.

El tercer paràmetre calculat s'anomena *recall* però, també és conegut com a sensibilitat. Aquest ens indica la proporció de registres positius que el model ha identificat correctament. En aquest cas, el paràmetre val 0,9038119 i, per tant, podem dir que té una precisió del 90,38119% en la classificació dels registres positius dins de la base d'entrenament.

L'últim paràmetre que calculem és l'especificitat que ens diu quina proporció de casos negatius del model s'han classificat de manera correcta. Aquest paràmetre val 0,999992 i, així que, direm que té una precisió del 99,9992% en la classificació dels registres negatius dins de la base d'entrenament.

Una altra manera de veure si el model fa bones prediccions és calcular la corba de ROC i l'àrea de sota aquesta. Veiem que aquesta àrea de sota la corba, en les dades d'entrenament, val 0,951902. Aquest valor ens diu la capacitat del model per distingir entre els nivells de la variable resposta en el conjunt de dades de prova. Una altra manera d'entendre-ho seria saber si el sinistre es troba en la fase INICIAL o la fase FINAL.

Figura 50: Corba de ROC amb les dades d'entrenament



En la corba tenim representats els falsos positius (proporció de negatius que no estan ben classificats) en l'eix X i els verdaders positius (proporció de positius que estan ben classificats) en l'eix Y. Podem veure que la corba està situada per sobre de la línia i això ens diu que hi ha una bona predicció en les dades d'entrenament.

6.1.2. Arbres de decisió amb les dades de prova

Ara ens fixem com apliquem el model d'arbres de decisió amb les dades de prova. Comencem fixant-nos en els paràmetres de complexitat o el CP. Aquest paràmetre ens dirà la complexitat que representa l'arbre. Veiem que els tres paràmetres valen 0,88864611, 0,03365843 i 0,01. També veiem que la variable més important per construir l'arbre és PROVACT. Aquesta serà la que millor defineixi els nivells de la variable resposta i ens definirà els nodes fills.

Figura 51: Model d'Arbre de decisió amb les dades de prova

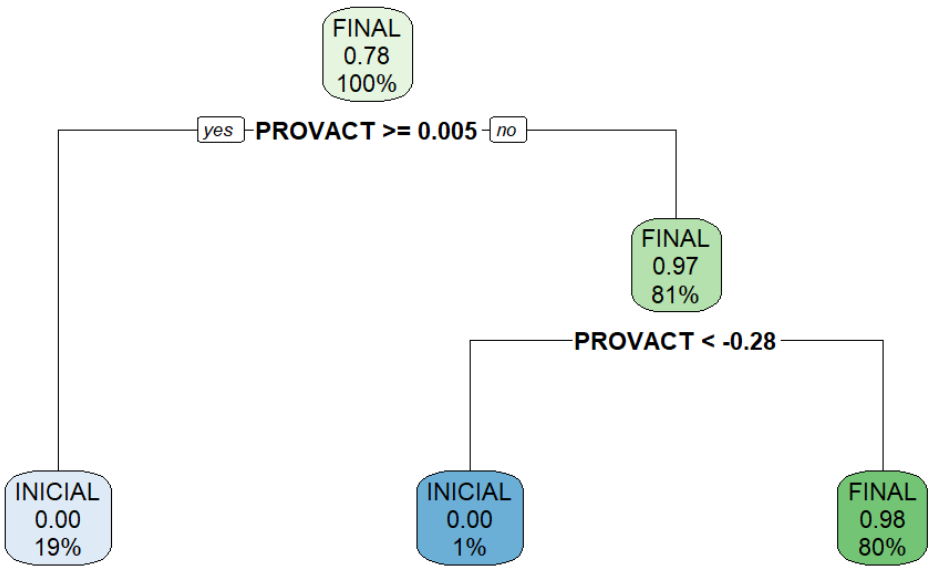
```
Call:
rpart(formula = Y ~ PROVACT + PROVINI + INDRECO + IMDCORP + INDCULP +
    INDVIDR + TIPTECD, data = dataTest2, method = "class", parms = list(split = "information"))
  n=466589 (11217 observations deleted due to missingness)

          CP nsplit  rel error     xerror         xstd
1 0.88864611      0 1.00000000 1.00000000 0.0027907806
2 0.03365843      1 0.11135389 0.11135389 0.0010389201
3 0.01000000      2 0.07769546 0.07769546 0.0008710384

Variable importance
PROVACT
    100

Node number 1: 466589 observations,    complexity param=0.8886461
  predicted class=FINAL    expected loss=0.2157959  P(node) =1
    class counts: 100688 365901
   probabilities: 0.216 0.784
  left son=2 (89488 obs) right son=3 (377101 obs)
  Primary splits:
      PROVACT < 0.005    to the right, improve=192840.100, (0 missing)
      INDRECO splits as  RL,           improve=  7246.882, (0 missing)
      IMDCORP splits as  RL,           improve=  3535.262, (0 missing)
      PROVINI < 901.9    to the right, improve=  3534.909, (0 missing)
      INDCULP splits as  LLLR,         improve=  2583.923, (0 missing)
  Surrogate splits:
      PROVINI < 2050.915 to the right, agree=0.809, adj=0.003, (0 split)

Node number 2: 89488 observations
  predicted class=INICIAL  expected loss=6.70481e-05  P(node) =0.1917919
    class counts: 89482     6
   probabilities: 1.000 0.000

Node number 3: 377101 observations,    complexity param=0.03365843
  predicted class=FINAL    expected loss=0.02971618  P(node) =0.8082081
    class counts: 11206 365895
   probabilities: 0.030 0.970
  left son=6 (3389 obs) right son=7 (373712 obs)
  Primary splits:
      PROVACT < -0.285   to the left,  improve=12474.380, (0 missing)
      INDRECO splits as  RL,           improve= 4511.352, (0 missing)
      INDCULP splits as  LLRR,         improve= 3524.448, (0 missing)
      PROVINI < -399.5   to the left,  improve= 3290.780, (0 missing)
      TIPTECD splits as  RLL,          improve= 2801.073, (1173 missing)

Node number 6: 3389 observations
  predicted class=INICIAL  expected loss=0  P(node) =0.007263352
    class counts:  3389     0
   probabilities: 1.000 0.000

Node number 7: 373712 observations
  predicted class=FINAL    expected loss=0.02091718  P(node) =0.8009447
    class counts:  7817 365895
   probabilities: 0.021 0.979
```

A partir del model obtindrem l'arbre de decisió. En ell podem veure que té un total de 7 nodes. El node arrel ens representa tots els registres de la mostra. En l'últim node ens representen les classes predites.

Podem observar que el node 1 té 466589 registres en total. I que es divideix en dos nodes fills. El node 2 que té 89488 registres i el node 3 que en té 377101. Al node 3, li surten dos nodes fills. El node 6 que té 3389 registres i el node 7 que en té 373712.

Destaquem que en el node 7, les probabilitats ens mostren que la majoria de registres es troben a la fase FINAL. Podem veure que aproximadament és en un 97,9% dels casos. En canvi, en el node 2, 3 i 6, pràcticament tots els casos estan en la fase INICIAL.

Figura 52: Arbre de decisió amb les dades de prova



Per poder veure com avalua aquest model, calculem la seva matriu de confusió i els paràmetres de decisió per veure si és un bon model.

Figura 53: Matriu de confusió amb les dades de prova

```
Confusion Matrix and Statistics
                Reference
Prediction   INICIAL    FINAL
INICIAL        92871        6
FINAL           7817   365895
```

En la matriu podem veure si van predir bé segons les classes de la variable resposta. Veiem que ha classificat correctament 92871 sinistres com a fase INICIAL (positius de veritat) i ha classificat correctament 365895 sinistres com a fase FINAL (negatius de veritat). Per altra banda, també veiem que no ha classificat bé 6 registres que estan a la fase FINAL que realment haurien de classificar-se a la fase INICIAL (falsos positius). I també que, no s'han classificat bé 7817 registres que estan a la fase INICIAL i que realment hauria de classificar-se a la fase FINAL (falsos negatius).

Gràcies a la matriu podem calcular els paràmetres que ens ajudaran a veure quin és el millor model de classificació.

El primer paràmetre de decisió calculat és l'exactitud o *accuracy*. Aquest ens diu la proporció d'instàncies que s'han classificat de manera correcta en les dades d'entrenament. En aquest cas, el paràmetre val 0,9832336 i, per tant, podem dir que té una precisió del 98,32336% en la classificació dels registres en aquest conjunt.

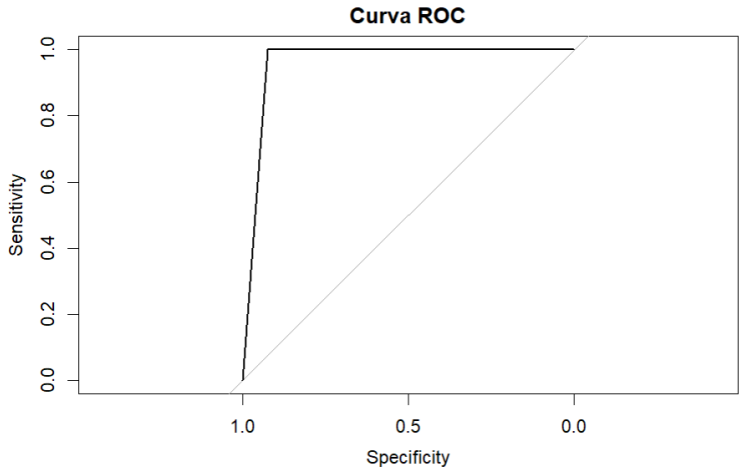
El segon paràmetre de decisió calculat és la precisió. Aquest ens diu la proporció de registres que s'han classificat correctament dins dels casos positius i, per tant, que es troben en la fase INICIAL. El valor del paràmetre en aquest cas val 0,9999354, així doncs, direm que té una precisió del 99,99354% en la classificació de registres dins del conjunt de positius.

El tercer paràmetre calculat s'anomena *recall* o sensibilitat. Aquest ens indica la proporció de registres positius que el model ha identificat correctament. En aquest cas, el paràmetre val 0,9223641 i, per tant, podem dir que té una precisió del 92,23641% en la classificació dels registres positius dins de la base d'entrenament.

L'últim paràmetre que calculem és l'especificitat que ens diu quina proporció de casos negatius del model s'han classificat de manera correcta. Aquest paràmetre val 0,9999836 i, per tant, direm que té una precisió del 99,99836% en la classificació dels registres negatius dins de la base d'entrenament.

Una altra manera de veure si el model fa bones prediccions és calcular la corba de ROC i l'àrea de sota aquesta. Veiem que aquesta àrea de sota la corba en les dades d'entrenament val 0,9611741. Aquest valor ens diu la capacitat del model per distingir entre els nivells de la variable resposta en el conjunt de dades de prova. Una altra manera d'entendre-ho seria saber si el sinistre es troba en la fase INICIAL o la fase FINAL.

Figura 54: Corba de ROC amb les dades de prova



En la corba tenim representats els falsos positius (proporció de negatius que no estan ben classificats) en l'eix X i els verdaders positius (proporció de positius que estan ben classificats) en l'eix Y. Podem veure que la corba està situada per sobre de la línia i això ens diu que hi ha una bona predicció en les dades de prova.

6.1.3. Comparació dels resultats

Taula 32: Paràmetres calculats segons les dades utilitzades

	Exactitud	**Precisió**	**Sensibilitat**	**Especificitat**	**Àrea sota la corba**
Dades d'entrenament	0,9933701	0,9998807	0,9038119	0,9999920	0,9519020
Dades de prova	0,9832336	0,9999354	0,9223641	0,9999836	0,9611741

A la taula podem veure els diferents paràmetres que hem calculat en els apartats anteriors. Amb aquests resultats podem dir que els arbres de decisió són un bon model de classificació.

6.2. Naive Bayes

En aquest apartat mostrarem els resultats després d'aplicar l'algoritme de Naive Bayes. Aquest algoritme és de classificació no lineal supervisat. Els classificadors de Bayes són uns classificadors probabilístics simples basats en l'aplicació del teorema d'aquest algoritme.

Per aplicar aquest model hem fet servir la funció *naive_bayes()* amb totes les variables de la base, tant numèriques com categòriques. Això farà que estiguem classificant la variable resposta en funció de les set variables predictores en el conjunt de dades d'entrenament i en el conjunt de dades de prova.

6.2.1. Naive Bayes amb les dades d'entrenament

En aquest apartat trobem l'anàlisi i previsió utilitzant les dades d'entrenament. Comencem aplicant la classificació de Bayes i obtenim els resultats que veiem en la següent figura.

Gràcies al model que veiem a continuació, veiem que el paràmetre *Laplace* val 0. La correcció de *Laplace* la usem per evitar que tinguem probabilitats iguals a 0 quan un nivell no està en el conjunt d'entrenament.

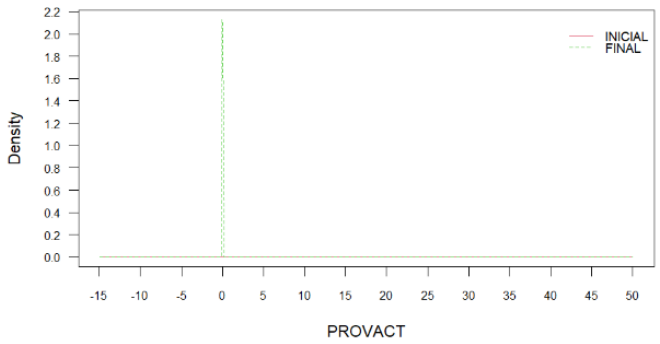
Amb aquest model també veiem les probabilitats a priori de cada un dels nivells. Hem vist que, per la classe INICIAL, la probabilitat a priori és de 0,0237 i que, per la classe FINAL, és 0,9763. Aquestes probabilitats ens donen les proporcions de cada nivell en la base d'entrenament.

Figura 55: Model Naive Bayes amb les dades d'entrenament

```
============================ Naive Bayes ============================

- Call: naive_bayes.formula(formula = Y ~ PROVACT + PROVINI +
      INDRECO + IMDCORP + INDCULP + INDVIDR + TIPTECD,
      data = dataTrain)
- Laplace: 0
- Classes: 2
- Samples: 277256
- Features: 7
- Conditional distributions:
    - Bernoulli: 3
    - Categorical: 2
    - Gaussian: 2
- Prior probabilities:
    - INICIAL: 0.0237
    - FINAL: 0.9763
```

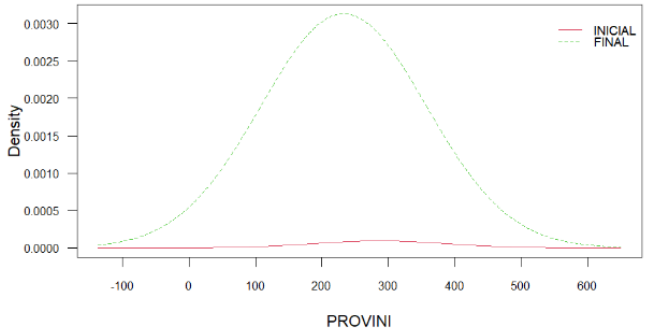
Un cop vist i analitzat el model, veurem representades les diferents variables en funció de la variable resposta (Y). De totes les variables en triarem unes quantes per poder veure com es comporten.

Figura 56: Variable resposta Y amb la provisió actual



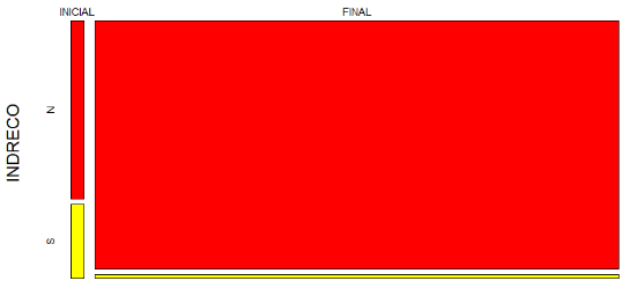
En aquesta gràfica tenim la variable resposta amb la variable que ens diu la provisió actual del sinistre. Veiem una clara diferència entre les fases del sinistre. Predomina la fase FINAL. Observem que, quan la provisió es troba al voltant de 0, és més probable que la mostra es classifiqui com a FINAL.

Figura 57: Variable resposta Y amb la provisió inicial



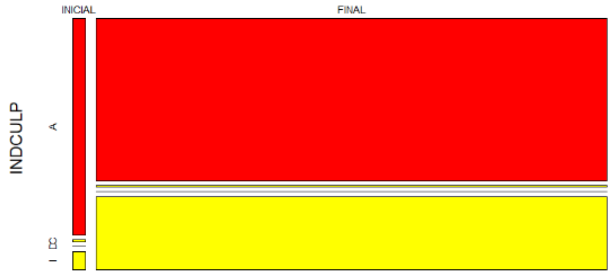
Ara ens fixem en la gràfica de la variable resposta amb la variable de la provisió inicial del sinistre. Tornem a veure una clara diferència, torna a predominar la fase FINAL. Quan la provisió es troba al voltant de 200, hi ha més probabilitats que es classifiqui com a fase FINAL.

Figura 58: Variable resposta Y amb l'indicador del sinistre recobrable



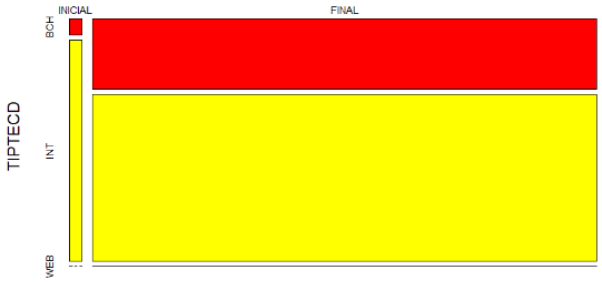
En aquesta figura podem observar la diferència en les probabilitats entre els nivells de la variable INDRECO i les de la resposta. Veiem que el nivell N (sinistre no recobrable) tenen una probabilitat major, tant en la fase INICIAL com la fase FINAL. Això ens diu que les mostres que pertanyen en aquest nivell, tenen la mateixa probabilitat d'estar classificades en la fase INICIAL com en la FINAL. En canvi, en el nivell S (sinistre sí recobrable) tenen més probabilitat que es classifiqui en la fase INICIAL que en la FINAL, com podem veure en la franja groga de la gràfica.

Figura 59: Variable resposta Y amb l'indicador de culpa



En la gràfica de sobre tenim la variable resposta amb l'indicador de culpa. Podem veure diferències en els diferents indicadors de culpa. Podem destacar que el nivell A mostra una probabilitat més alta, tant en la fase INICIAL com en la FINAL, comparat amb la resta de nivells. Per altra banda, veiem que el nivell D té la probabilitat menys baixa de totes. Això ens diu que normalment aquest indicador de culpa no sol pertànyer a cap de les fases INICIAL i FINAL.

Figura 60: Variable resposta Y amb el tipus de declaració del sinistre de tecnologia



Per últim, tenim la variable resposta amb el tipus de declaració del sinistre de tecnologia. Observem les diferències entre els tres nivells. Ens fixem que, en el nivell de l'INTRANET, hi ha una probabilitat més alta en els dos nivells de la variable resposta. I que el nivell de l'INTERNET té la menor probabilitat de pertànyer a una de les fases de la variable resposta.

Prediccions amb les dades d'entrenament

Per poder avaluar el model de classificació ens fixem en la matriu de confusió. Així podem comprovar si el model és bo i, per tant, tindrem bones prediccions. Veiem la matriu a continuació:

Figura 61: Matriu de confusió de les dades d'entrenament

```
Confusion Matrix and Statistics
                Reference
Prediction       INICIAL     FINAL
   INICIAL          1141         0
     FINAL          5437    270678
```

En la matriu podem veure si van predir bé segons les classes de la variable resposta. Podem observar que ha classificat correctament 1141 sinistres com a fase INICIAL (positius de veritat) i ha classificat correctament 270678 sinistres com a fase FINAL (negatius de veritat). Per altra banda, també veiem que no ha classificat malament cap registre que estigui a la fase FINAL i que realment hauria de classificar-se a la fase INICIAL (falsos positius). I que no s'han classificat bé 5437 registres que estan a la fase INICIAL i que realment haurien de classificar-se a la fase FINAL (falsos negatius).

Gràcies a la matriu podem calcular els paràmetres que ens ajudaran a veure quin és el millor model de classificació.

El primer paràmetre calculat és l'exactitud o *accuracy*. Aquest ens diu la proporció d'instàncies que s'han classificat de manera correcta en les dades d'entrenament. En aquest cas, el paràmetre val 0,98039 i, per tant, podem dir que té una precisió del 98,039% en la classificació dels registres en aquest conjunt.

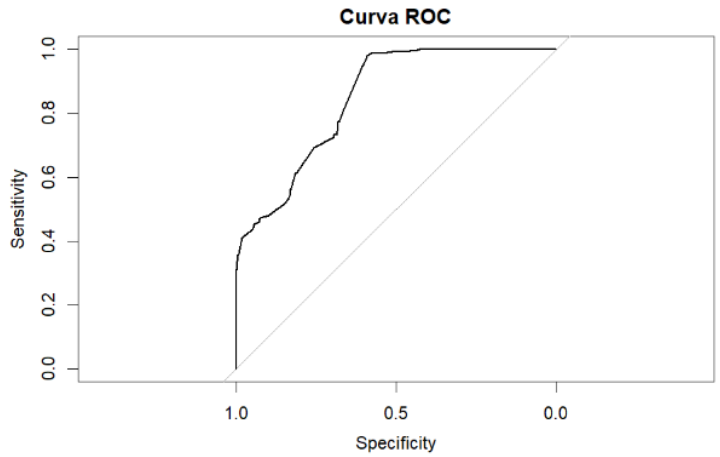
El segona paràmetre calculat és la precisió. Aquest ens diu la proporció de registres que s'han classificat correctament dins dels casos positius, així doncs, que es troben en la fase INICIAL. El valor del paràmetre en aquest cas val 1 i, en conseqüència, direm que té una precisió del 100% en la classificació de registres dins del conjunt de positius.

El tercer paràmetre calculat s'anomena *recall* o sensibilitat. Aquest ens indica la proporció de registres positius que el model ha identificat correctament. En aquest cas, el paràmetre val 0,173457 i, per tant, podem dir que té una precisió del 17,3457% en la classificació dels registres positius dins de la base d'entrenament.

L'últim paràmetre que calculem és l'especificitat, que ens diu quina proporció de casos negatius del model s'han classificat de manera correcta. Aquest paràmetre val 1 i, per tant, direm que té una precisió del 100% en la classificació dels registres negatius dins de la base d'entrenament.

Una altra manera de veure si el model fa bones prediccions és calcular la corba de ROC i l'àrea de sota aquesta. Veiem que aquesta àrea de sota la corba, en les dades d'entrenament, val 0,8473157. Aquest valor ens diu la capacitat del model per distingir entre els nivells de la variable resposta en el conjunt de dades de prova. Una altra manera d'entendre-ho seria saber si el sinistre es troba en la fase INICIAL o la fase FINAL.

Figura 62: Corba de ROC del model Naive Bayes amb les dades d'entrenament



En la corba tenim representats els falsos positius (proporció de negatius que no estan ben classificats) en l'eix X i els verdaders positius (proporció de positius que estan ben classificats) en l'eix Y. Podem veure que la corba està situada per sobre de la línia i això ens diu que hi ha una bona predicció en les dades d'entrenament.

6.2.2. Naive Bayes amb les dades de prova

En aquest apartat trobem l'anàlisi i previsió utilitzant les dades de prova. Comencem aplicant la classificació de Bayes i obtenim els resultats que veiem en la següent figura.

Gràcies al model que veiem a continuació, veiem que el paràmetre *Laplace,* val 0. La correcció de *Laplace* l'utilitzem per evitar tenir probabilitats iguals a 0 quan un nivell no està en el conjunt de prova.

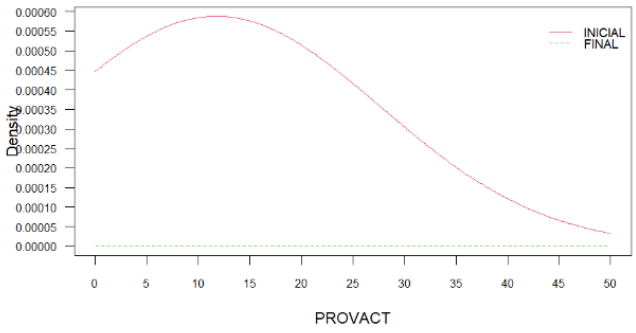
Amb aquest model també veiem les probabilitats a priori de cada un dels nivells. Hem vist que per la classe INICIAL la probabilitat a priori és de 0,1605 i que per la classe FINAL és 0,8395. Aquestes probabilitats ens donen les proporcions de cada nivell en la base de prova.

Figura 63: Model Naive Bayes amb les dades de prova

```
============================ Naive Bayes ============================
- Call: naive_bayes.formula(formula = Y ~ PROVACT + PROVINI +
      INDRECO + IMDCORP + INDCULP + INDVIDR + TIPTECD,
      data = dataTest)
- Laplace: 0
- Classes: 2
- Samples: 138628
- Features: 7
- Conditional distributions:
    - Bernoulli: 3
    - Categorical: 2
    - Gaussian: 2
- Prior probabilities:
    - INICIAL: 0.0234
    - FINAL: 0.9766
```

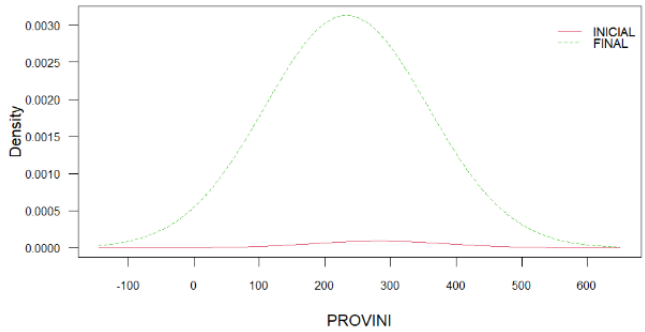
Ara que ja hem vist i analitzat el model, representem les diferents variables en funció de la variable resposta (Y) i analitzem les gràfiques. De totes les variables en triarem unes quantes per poder veure com es comporten.

Figura 64: Variable resposta Y amb la provisió actual



En aquesta gràfica tenim la variable resposta amb la variable que ens diu la provisió actual del sinistre. Veiem una clara diferència entre les fases del sinistre, predomina la fase INICIAL. Observem que quan la provisió es troba al voltant de 0, és més probable que la mostra es classifiqui com a INICIAL.

Figura 65: Variable resposta Y amb la provisió inicial

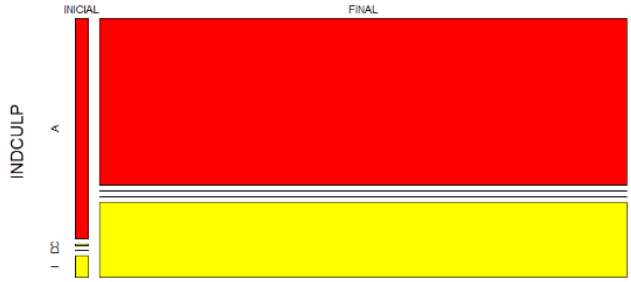


Ara ens fixem en la gràfica de la variable resposta amb la variable que ens dona la provisió inicial del sinistre. Tornem a veure una clara diferència, però aquesta vegada predomina la fase FINAL. Quan la provisió es troba al voltant de zero, hi ha més probabilitats que es classifiqui com a fase FINAL.

Figura 66: Variable resposta Y amb l'indicador del sinistre recobrable

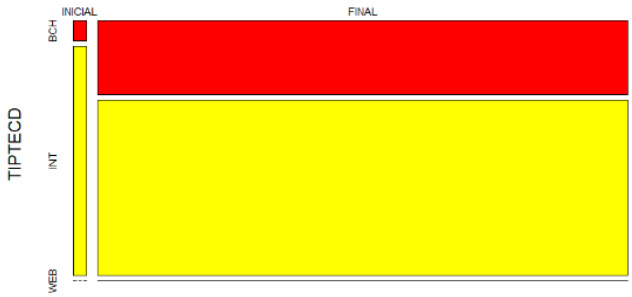
En aquesta figura podem veure la diferència en les probabilitats entre els nivells de la variable INDRECO i les de la resposta. Veiem que el nivell N (sinistre no recobrable) tenen una probabilitat major tant en la fase INICIAL com la fase FINAL. Això ens diu, que les mostres que pertanyen en aquest nivell, tenen la mateixa probabilitat d'estar classificades en la fase INICIAL com a la FINAL. En canvi, en el nivell S (sinistre sí recobrable) tenen més probabilitat de què es classifiquin en la fase INICIAL que en la FINAL, com podem veure en la franja groga de la gràfica.

Figura 67: Variable resposta Y amb l'indicador de culpa



Ara ens fixem en la gràfica de la variable resposta amb l'indicador de culpa. Podem veure diferències en els diferents indicadors, destacant que el nivell A mostra una probabilitat més alta, tant en la fase INICIAL com en la FINAL, comparat amb la resta de nivells. D'altra banda, mirant el nivell D, que té la probabilitat més baixa de totes, veiem que normalment l'indicador de culpa no sol pertànyer a les fases INICIAL i FINAL.

Figura 68: Variable resposta Y amb el tipus de declaració del sinistre de tecnologia



Per últim, tenim la variable resposta amb el tipus de declaració del sinistre de tecnologia. Observem diferències entre els tres nivells. Si ens fixem en el nivell de l'INTRANET, veiem que té una probabilitat més alta en els dos nivells de la variable resposta. En canvi, el nivell de l'INTERNET, veiem que té la menor probabilitat de pertànyer a les fases de la variable resposta.

Prediccions amb les dades de prova

Per poder avaluar el model de classificació, ens fixem en la matriu de confusió, per així, comprovar si el model és bo i, per tant, tenir bones prediccions. A continuació veiem la matriu:

Figura 69: Matriu de confusió de les dades de prova

```
Confusion Matrix and Statistics
                Reference
Prediction     INICIAL   FINAL
   INICIAL         430       0
     FINAL        2818  135380
```

En la matriu, podem veure si van predir bé, segons les classes de la variable resposta. Veiem que ha classificat correctament 430 sinistres com a fase INICIAL (positius de veritat) i ha classificat correctament 135380 sinistres com a fase FINAL (negatius de veritat). Per altra banda, també veiem que no ha classificat malament cap registre que estan a la fase FINAL i que realment hauria de classificar-se a la fase INICIAL (falsos positius) i que no s'han classificat bé 2818 registres que estan a la fase INICIAL i que realment hauria de classificar-se a la fase FINAL (falsos negatius).

Gràcies a la matriu podem calcular les mètriques que ens ajudaran a veure quin és el millor model de classificació.

El primer paràmetre calculat és l'exactitud o *accuracy*. Aquest ens diu la proporció d'instàncies que s'han classificat de manera correcta en les dades de prova. En aquest cas, el paràmetre val 0,9796722 i, per tant, podem dir que té una precisió del 97,96722% en la classificació dels registres en aquest conjunt.

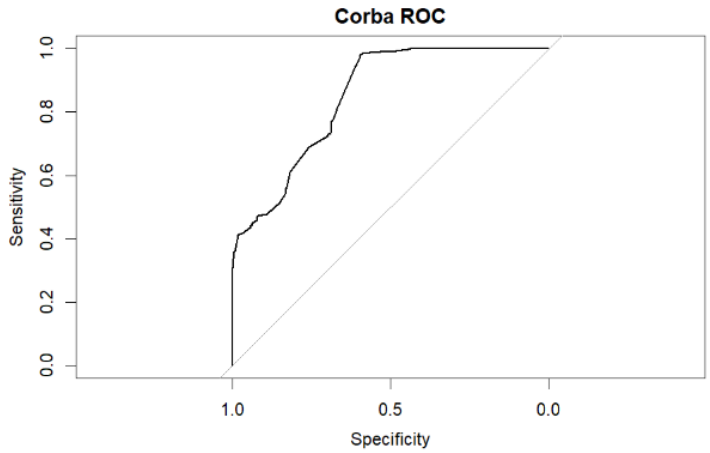
El segon paràmetre calculat és la precisió. Aquest ens diu la proporció de registres que s'han classificat correctament dins dels casos positius i, per tant, que es troben en la fase INICIAL. El valor del paràmetre en aquest cas val 1 i, en conseqüència, direm que té una precisió del 100% en la classificació de registres dins del conjunt de positius.

El tercer paràmetre calculat s'anomena *recall* o sensibilitat. Aquest ens indica la proporció de registres positius que el model ha identificat correctament. En aquest cas, el paràmetre val 0,1323892 i, per tant, podem dir que té una precisió del 13,23892% en la classificació dels registres positius dins de la base de prova.

L'últim paràmetre que calculem és l'especificitat que ens diu quina proporció de casos negatius del model que s'han classificat de manera correcta. Aquest paràmetre val 1 i, en conseqüència, direm que té una precisió del 100% en la classificació dels registres negatius dins de la base de prova.

Una altra manera de veure si el model fa bones prediccions és calcular la corba de ROC i l'àrea de sota aquesta. Veiem que, aquesta àrea de sota la corba en les dades de prova val 0,8483239. Aquest valor ens diu la capacitat del model per distingir entre els nivells de la variable resposta en el conjunt de dades de prova. Una altra manera d'entendre-ho seria saber si el sinistre es troba en la fase INICIAL o la fase FINAL.

Figura 70: Corba de ROC del model Naive Bayes amb les dades de prova



En la corba tenim representats els falsos positius (proporció de negatius que no estan ben classificats) en l'eix X i els verdaders positius (proporció de positius que estan ben classificats) en l'eix Y. Podem veure que la corba està situada per sobre de la línia i això ens diu que hi ha una bona predicció en les dades de prova.

6.2.3. Comparació dels resultats

Taula 33: Paràmetres calculats segons les dades utilitzades

	Exactitud	Precisió	Sensibilitat	Especificitat	Àrea sota la corba
Dades d'entrenament	0,98039	1	0,173457	1	0,8473157
Dades de prova	0,9796722	1	0,1323892	1	0,8483239

A la taula podem veure els diferents paràmetres que hem calculat en els apartat anteriors. Amb aquests resultats podem dir que Naive Bayes és un bon model de classificació.

6.3. Regressió logística

En aquest apartat aplicarem el model de classificació logística. Aquest model l'aplicarem gràcies a la funció *glm()*. L'objectiu del model serà predir la fase on es troba el sinistre a partir de la variable resposta (Y) en funció de les següents variables que utilitzarem com a predictores: PROVACT, PROVINI, IMDCORP, INDCULP, INDVIDR, TIPTECD i INDRECO. Hem calculat el model pel conjunt de dades d'entrenament i les de prova.

6.3.1. Regressió logística amb les dades d'entrenament

A continuació veiem els resultats del model de regressió logística amb les dades d'entrenament en la següent sortida:

Figura 71: Model de regressió logística amb les dades d'entrenament

```
Call:  glm(formula = Y ~ PROVINI + PROVACT + IMDCORP + INDRECO + INDCULP
    + INDVIDR + TIPTECD, family = binomial, data = dataTrain)

Coefficients:
(Intercept)      PROVINI      PROVACT     IMDCORPS     INDRECOS
   4.908635    -0.000236    -8.157681     0.608002    -2.906676
   INDCULPC     INDCULPD     INDCULPI     INDVIDRS   TIPTECDINT
   3.577721     0.831342     2.724718     0.407606    -0.654838
 TIPTECDWEB
   7.375488
Degrees of Freedom: 277255 Total (i.e. Null);  277254 Residual
Null Deviance:       62220
Residual Deviance: 33890    AIC: 33910
```

La classificació logística ens aporta un coeficient estimat per cada variable segons el nivell. Cada un dels coeficients li podem donar un significat.

Quan una variable té l'indicador de lesionats positius (IMDCORPS), és a dir, que hi ha hagut lesionats durant el sinistre, el coeficient val 0,608002 i, per tant, la probabilitat que ens trobem a la fase FINAL del sinistre és més alta. Per altra banda, si mirem el coeficient de la variable INDRECO quan sí hi ha l'indicador que el sinistre és recobrable, veiem que val -2,906676 i, com a resultat, la probabilitat que el sinistre es trobi a la fase FINAL disminueix.

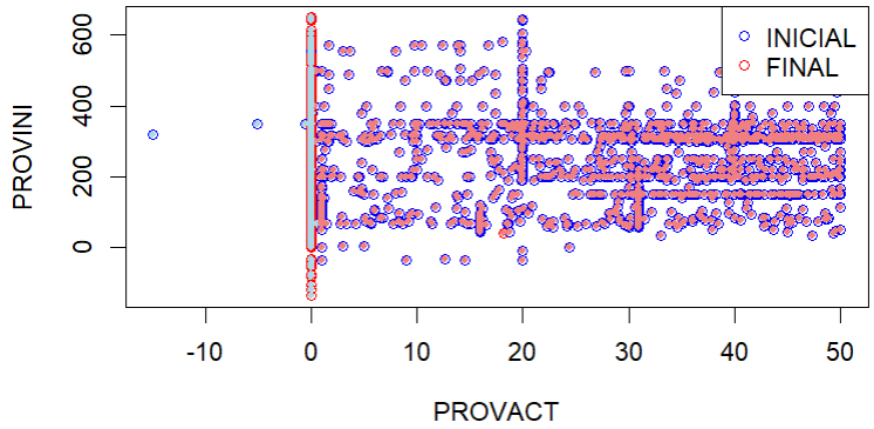
El següent coeficient que tenim és el de la variable que ens dona l'indicador de culpa (INDCULP). Quan aquest indicador és C, la probabilitat que estigui a la fase FINAL del sinistre augmenta 3,577721, quan l'indicador és I, augmenta 2,724718 i quan l'indicador és D, la probabilitat que estigui a la fase FINAL augmenta 0,831342.

Si ens fixem en el coeficient de la variable INDVIDR, veiem que, quan hi ha indicador de vidre, la probabilitat que estigui el sinistre a la fase FINAL augmenta 0,407606. Per últim, veiem la variable que ens diu amb quina tecnologia s'ha declarat el sinistre. Veiem que, quan s'ha declarat per intranet, la probabilitat disminueix 0,0654838 i, quan s'ha declarat per internet, augmenta 7,375488.

Uns altres apunts que ens aporta el model són la desviació nul·la i la residual, els graus de llibertat i el paràmetre AIC. Veiem que el model té 277255 graus de llibertat, que la desviació nul·la val 62220 i la residual 33890. Per últim, el paràmetre AIC val 33910. Totes aquestes variables ens seran útils per fer la comparació de models.

A continuació veurem una representació gràfica del model de regressió. Quan mirem els eixos veiem que, en l'eix d'abscisses, es representa la variable numèrica PROVACT, que és la provisió en el moment actual. I, en l'eix d'ordenades, es representa la variable numèrica PROVINI, que és la provisió inicial. En la llegenda observem que quan està en la fase INICIAL del sinistre, aquesta està representada de color blau, i quan està en la fase FINAL del sinistre, aquesta està representada de color vermell.

Figura 72: Gràfica de la variable resposta, PROVACT i PROVINI



Mirant la gràfica també veiem representats aquells punts que han sigut predits pel model. Aquests es mostren en vermell clar i blau clar. Sabem que quan els punts són vermells clars, la probabilitat que es preveu per la predicció és menor o igual a 0,5 i, per tant, estem a la fase FINAL. També sabem que quan els punts són blaus clars, la probabilitat que es preveu per la predicció és major a 0,5 i, per tant, estem a la fase INICIAL.

A continuació mirem la matriu de confusió per calcular els paràmetres de decisió.

Figura 73: Matriu de confusió de les dades d'entrenament

```
Confusion Matrix and Statistics
          Reference
Prediction   INICIAL     FINAL
   INICIAL      2811         1
     FINAL      3767    270677
```

En la matriu podem veure si van predir bé segons les classes de la variable resposta. Veiem que ha classificat correctament 2811 sinistres com a fase INICIAL (positius de veritat) i ha classificat correctament 270677 sinistres com a fase FINAL (negatius de veritat). Per altra banda, també veiem que no ha classificat bé un registre que està a la fase FINAL i que realment hauria de classificar-se a la fase INICIAL (falsos positius). Tampoc s'han classificat bé 3767 registres que estan a la fase INICIAL i realment hauria de classificar-se a la fase FINAL (falsos negatius). Ara calcularem els paràmetres mencionats anteriorment.

El primer paràmetre calculat és l'exactitud o *accuracy*. Aquest ens diu la proporció d'instàncies que s'han classificat de manera correcta en les dades de prova. En aquest cas, el paràmetre val 0,9864097 i, en conseqüència, podem dir que té una precisió del 98,64097% en la classificació dels registres en aquest conjunt.

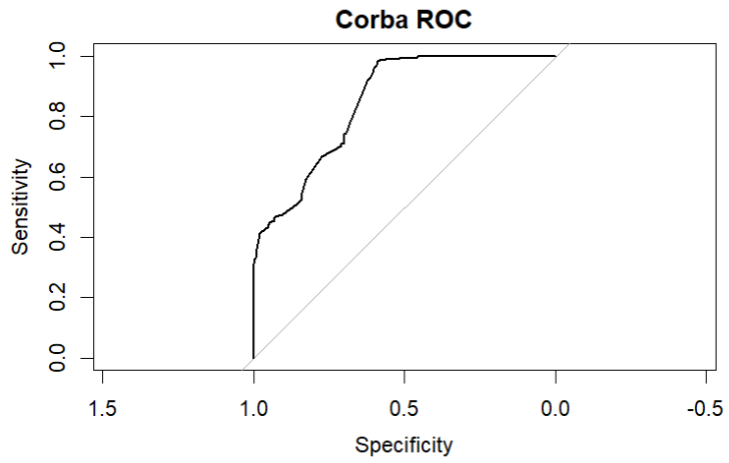
El segon paràmetre calculat és la precisió. Aquest ens diu la proporció de registres que s'han classificat correctament dins dels casos positius i, per tant, que es troben en la fase INICIAL. El valor del paràmetre en aquest cas val 0,9996444 i, com a resultat, direm que té una precisió del 99, 96444% en la classificació de registres dins del conjunt de positius.

El tercer paràmetre calculat s'anomena recall o sensibilitat. Aquest ens indica la proporció de registres positius que el model ha identificat correctament. En aquest cas, el paràmetre val 0,4273335 i, per tant, podem dir que té una precisió del 42,73335% en la classificació dels registres positius dins de la base de prova.

L'últim paràmetre que calculem és l'especificitat, el qual ens diu quina proporció de casos negatius del model que s'han classificat de manera correcta. Aquest paràmetre val 0,9999963 i, per tant, direm que té una precisió del 99, 99963% en la classificació dels registres negatius dins de la base de prova.

Una altra manera de veure si el model fa bones prediccions és calcular la corba de ROC i l'àrea de sota aquesta. Observem que aquesta àrea de sota la corba en les dades de prova val 0,8480436. Aquest valor ens diu la capacitat del model per distingir entre els nivells de la variable resposta en el conjunt de dades de prova. Una altra manera d'entendre-ho seria saber si el sinistre es troba en la fase INICIAL o la fase FINAL.

Figura 74: Corba de ROC del model de classificació lineal amb les dades d'entrenament



En la corba tenim representats els falsos positius (proporció de negatius que no estan ben classificats) en l'eix X i els verdaders positius (proporció de positius que estan ben classificats) en l'eix Y. Podem veure que la corba està situada per sobre de la línia i, això, ens diu que hi ha una bona predicció en les dades de prova.

6.3.2. Regressió logística amb les dades de prova

Ara ens fixarem en els resultats del model de regressió logística amb les dades de prova en la següent sortida:

Figura 75: Model de regressió logística amb les dades de prova

```
Call:  glm(formula = Y ~ PROVINI + PROVACT + IMDCORP + INDRECO + INDCULP
    + INDVIDR + TIPTECD, family = binomial, data = dataTest)
Coefficients:
(Intercept)      PROVINI        PROVACT     IMDCORPS     INDRECOS
 19.0314049   -0.0002646   -152.0582632    0.4249442    -2.850095
   INDCULPC     INDCULPD     INDCULPI       INDVIDRS    TIPTECDINT
 14.6994318   18.5680678   17.8145562    -15.7818865   -14.7398699
 TIPTECDWEB
  0.2681223
Degrees of Freedom: 138627 Total (i.e. Null);  138617 Residual
 Null Deviance:      30800
 Residual Deviance: 16430    AIC: 16450
```

La classificació logística ens aporta un coeficient estimat per cada variable segons el nivell. A cada un dels coeficients li podem donar un significat.

Quan una variable té l'indicador de lesionats positius (IMDCORPS), és a dir, que hi ha hagut lesionats durant el sinistre, el coeficient val 0,4249442 i la probabilitat que ens trobem a la fase FINAL del sinistre és més alta. D'altra banda, si mirem el coeficient de la variable INDRECO, quan sí hi ha l'indicador que el sinistre és recobrable, veiem que val -2,850095 i, com a resultat, la probabilitat que el sinistre es trobi a la fase FINAL disminueix.

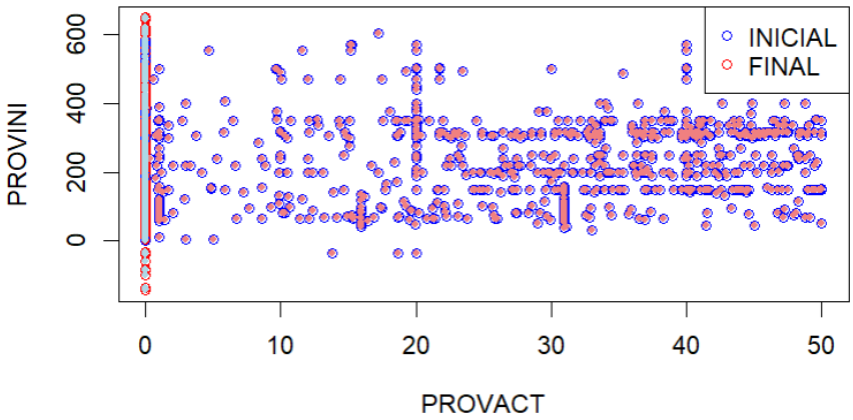
El següent coeficient que tenim és el de la variable que ens dona l'indicador de culpa (INDCULP). Quan aquest indicador és C, la probabilitat que estigui a la fase FINAL del sinistre augmenta 14,6994318, quan l'indicador és I, augmenta 17,8145562 i quan l'indicador és D, la probabilitat que estigui a la fase FINAL, augmenta 18,5680678.

Si ens fixem en el coeficient de la variable INDVIDR, veiem que quan hi ha indicador de vidre, la probabilitat que estigui el sinistre a la fase FINAL disminueix 15,7818865. Per acabar, veiem la variable que ens diu amb quina tecnologia s'ha declarat el sinistre. Observem que quan s'ha declarat per intranet, la probabilitat disminueix 14,7398699, en canvi, quan s'ha declarat per internet augmenta 3,3328683.

Uns altres apunts que ens aporta el model són la desviació nul·la i la residual, els graus de llibertat i el paràmetre AIC. Veiem que el model té 138627 graus de llibertat, que la desviació nul·la val 30800 i la residual 16430, el paràmetre AIC val 16450. Totes aquestes variables ens seran útils per fer la comparació de models.

A continuació veurem una representació gràfica del model de regressió. Quan mirem els eixos veiem que en l'eix d'abscisses es representa la variable numèrica PROVACT, que és la provisió en el moment actual. I en l'eix d'ordenades es representa la variable numèrica PROVINI, que és la provisió inicial. En la llegenda veiem que, quan està en la fase INICIAL del sinistre, aquesta, està representada de color blau, i quan està en la fase FINAL del sinistre, aquesta, està representada de color vermell.

Figura 76: Gràfica de la variable resposta, PROVACT i PROVINI



Observant la gràfica també veiem representats aquells punts que han sigut predits pel model. Aquests es mostren en vermell clar i blau clar. Sabem que quan els punts són vermells clars, la probabilitat que es preveu per la predicció és menor o igual a 0,5, així que, estem a la fase FINAL. També sabem que, quan els punts són blaus clars, la probabilitat que es preveu per la predicció és major a 0,5 i, en conseqüència, estem a la fase INICIAL.

A continuació mirem la matriu de confusió per calcular els paràmetres de decisió.

Figura 77: Matriu de confusió de les dades de prova

```
Confusion Matrix and Statistics
                Reference
   Prediction  INICIAL    FINAL
   INICIAL        1414        0
   FINAL          1834   135380
```

En la matriu podem veure si s'han predit bé segons les classes de la variable resposta. Observem que ha classificat correctament 1414 sinistres com a fase INICIAL (positius de veritat) i ha classificat correctament 135380 sinistres com a fase FINAL (negatius de veritat). Alhora, també veiem que no hi ha cap fals positius, és a dir, cap registre de la fase INICIAL s'ha classificat erròniament a la FINAL. Acabem veient que no s'han classificat bé 1834 registres que estan a la fase INICIAL i que realment hauria de classificar-se a la fase FINAL (falsos negatius). Ara calcularem els paràmetres mencionats anteriorment.

El primer paràmetre calculat és l'exactitud o *accuracy*. Aquest ens diu la proporció d'instàncies que s'han classificat de manera correcta en les dades de prova. En aquest cas, el paràmetre val 0,9867703 i, per tant, podem dir que té una precisió del 98,67703% en la classificació dels registres en aquest conjunt.

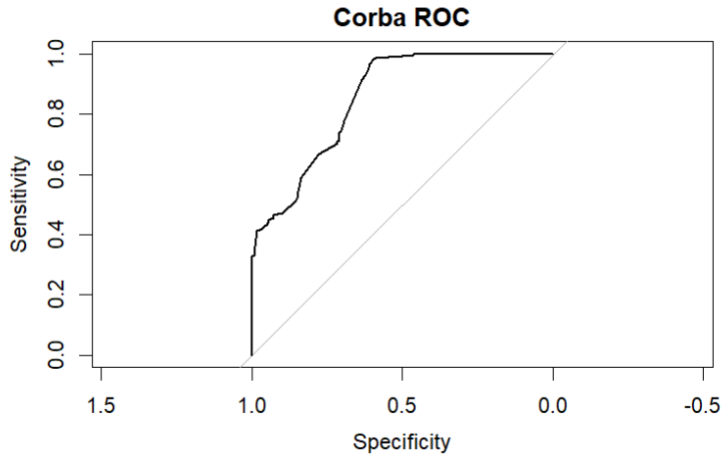
El segon paràmetre calculat és la precisió. Aquest ens diu la proporció de registres que s'han classificat correctament dins dels casos positius i, en conseqüència, es troben en la fase INICIAL. El valor del paràmetre en aquest cas val 1 i, per tant, direm que té una precisió del 100% en la classificació de registres dins del conjunt de positius.

El tercer paràmetre calculat s'anomena *recall* o sensibilitat. Aquest ens indica la proporció de registres positius que el model ha identificat correctament. En aquest cas, el paràmetre val 0,4353448 i, per tant, podem dir que té una precisió del 43,53448% en la classificació dels registres positius dins de la base de prova.

L'últim paràmetre que calculem és l'especificitat que ens diu quina proporció de casos negatius del model s'han classificat de manera correcta. Aquest paràmetre val 1 i, per tant, direm que té una precisió del 100% en la classificació dels registres negatius dins de la base de prova.

Una altra manera de veure si el model fa bones prediccions és calcular la corba de ROC i l'àrea de sota aquesta. Observem que aquesta àrea de sota la corba en les dades de prova val 0,8513982. Aquest valor ens diu la capacitat del model per distingir entre els nivells de la variable resposta en el conjunt de dades de prova. Una altra manera d'entendre-ho seria saber si el sinistre es troba en la fase INICIAL o la fase FINAL.

Figura 78: Corba de ROC del model de classificació lineal amb les dades de prova



En la corba tenim representats els falsos positius (proporció de negatius que no estan ben classificats) en l'eix X i els verdaders positius (proporció de positius que estan ben classificats) en l'eix Y. Podem veure que la corba està situada per sobre de la línia i, això, ens diu que hi ha una bona predicció en les dades de prova.

6.3.3. Comparació dels resultats

Taula 34: Paràmetres calculats segons les dades utilitzades

	Exactitud	**Precisió**	**Sensibilitat**	**Especificitat**	**Àrea sota la corba**
Dades d'entrenament	0, 9864097	0,9996444	0,4273335	0,9999963	0,8480436
Dades de prova	0,9867703	1	0,4353448	1	0,8513982

A la taula podem veure els diferents paràmetres que hem calculat en els apartats anteriors. Amb aquests resultats podem dir que la regressió logística és un bon model de classificació.

7. Comparació dels models

En aquest apartat farem la comparació dels diferents models que hem aplicat durant aquest treball. El dividirem en dos grups: un, els models aplicats a les dades d'entrenament i, dos, els models aplicats a les dades de prova.

Però, primer, recordarem el significat de cada un dels paràmetres que hem calculat per cada un dels models i mostres de les dades.

El primer paràmetre que hem calculat ha sigut l'**exactitud**, o també anomenat *accuracy*. Aquest paràmetre ens permet avaluar com classifiquen els algoritmes en tots els registres. Això vol dir que, com més alt sigui el valor, millor rendiment tindrà el model de classificació en el general dels registres.

El segon paràmetre calculat és la **precisió**, la qual ens dóna la proporció de registres que han sigut ben classificats en la classe positiva. Això vol dir que, ens dóna la probabilitat que un registre positiu sigui classificat correctament dins dels registres positius. Quan aquest paràmetre té un valor alt, significa que la mostra té una probabilitat baixa de falsos positius.

El tercer paràmetre calculat ha estat la **sensibilitat**. Aquest paràmetre ens dóna la probabilitat que tenen els models de dir que els casos positius han estat ben classificats. Quan el seu valor és alt, ens indica que el model és capaç de detectar la majoria dels casos positius de manera correcta.

El quart paràmetre que hem calculat ha sigut l'**especificitat** que ens diu la proporció de registres negatius que són realment negatius i, en conseqüencia, que s'han classificat de manera correcta.

Per acabar, hem calculat el paràmetre que ens dóna l'**àrea sota la corba de ROC**. Sabem que, quan aquest paràmetre és més gran que 0,5, el model prediu correctament la classificació de les observacions. Com més s'apropi a 1, millor serà el rendiment de classificació. En canvi, quan és menor a 0,5, el model no prediu de manera del tot correcta la classificació de les observacions. Com més s'apropi a 0, pitjor serà el rendiment de classificació.

7.1. Models aplicats a les dades d'entrenament

Taula 35: Models aplicats a les dades d'entrenament

	Exactitud	**Precisió**	**Sensibilitat**	**Especificitat**	**Àrea sota la corba**
Arbres de decisió	0,9933701	0,9998807	0,9038119	0,9999920	0,9519020
Naive Bayes	0,98039	1	0,173457	1	0,8473157
Regressió logística	0, 9864097	0,9996444	0,4273335	0,9999963	0,8480436

Quan ens fixem en la taula dels models aplicats en les dades d'entrenament veiem que quan mirem el paràmetre de l'exactitud, els arbres de decisió tenen el valor més alt. Aquest ens indica que el 99,33701% dels registres estan classificats correctament en el conjunt d'entrenament.

Passa igual amb els paràmetres de precisió i sensibilitat. En el cas de la precisió, veiem que el 99,98807% dels registres positius classificats són realment positius. En el cas de la sensibilitat, observem que el 90,38119% dels registres positius han estat correctament identificats.

En l'especificitat, el 99,9992% dels registres negatius han estat ben classificats quan ens fixem en el model de Regressió logística.

Finalment, ens fixem en l'àrea sota la corba de ROC. Veiem que el paràmetre més alt és el del model Arbres de decisió. Per tant, segons aquest paràmetre, podem afirmar que el millor model per classificar seria els Arbres de decisió amb un valor de 0,9519020.

En general, tots els resultats es podrien considerar molt bons. Tots els tres models classifiquen de manera molt bona, ja que els valors són molt alts.

Depenent del problema que vulguem resoldre, triarem un model o un altre. Tant en el cas que es vulgui una millor classificació general, com en el cas d'una classificació més concreta, veient els resultats esmentats, direm que el millor model és els Arbres de decisió. Tot i que, no té els valors més alts en tots els paràmetres, aquests continuen sent alts i pròxims entre ells.

7.2. Models aplicats a les dades de prova

Taula 36: Models aplicats a les dades de prova

	Exactitud	Precisió	Sensibilitat	Especificitat	Àrea sota la corba
Arbres de decisió	0,9832336	0,9999354	0,9223641	0,9999836	0,9611741
Naive Bayes	0,9796722	1	0,1323892	1	0,8483239
Regressió logística	0,9867703	1	0,4353448	1	0,8513982

Quan ens fixem en la taula dels models aplicats en les dades d'entrenament, veiem que quan mirem el paràmetre de l'exactitud, la regressió logística tenen el valor més alt. Aquest ens indica que el 98,67703% dels registres estan classificats correctament en el conjunt d'entrenament.

Passa el mateix amb la sensibilitat. Observem que el 92,23641% dels registres positius han estat correctament identificats i que, per tant, el millor model és els Arbres de decisió.

Fixant-nos en els paràmetres de precisió i especificitat, veiem que el 100% dels registres positius classificats són realment positius, tant en el cas de Naive Bayes com la Regressió logística.

Finalment, ens fixem en l'àrea sota la corba de ROC. Veiem que el paràmetre més alt és el del model d'Arbres de Decisió. Per tant, segons aquest, el millor model per classificar seria aquest amb un valor de 0,9611741.

En general tots els resultats es podrien considerar molt bons. Tots els tres models classifiquen de manera molt bona, ja que els valors són molt alts.

Depenent del problema que vulguem resoldre, triarem un model o un altre. En el nostre cas, i veient els resultats esmentats, direm que el millor model és l'Arbre de decisió. Tant, per quan busquem una classificació general de la variable com, quan busquem una més concreta. Tot i que no té tots valors més alts si que és la que té els paràmetres més pròxims.

7.3. Comparació de les dues mostres

Quan ens mirem els paràmetres calculats a partir de les dades d'entrenament hem de tenir en compte que, estan avaluant el rendiment del model pel seu ajust. Per molt que ens doni una idea general de com s'ajusten i quin serà el millor model classificador, no significa que sigui el millor per unes dades noves o no vistes, és a dir, que no s'han triat per la mostra.

Si mirem els paràmetres calculats a partir de les dades de prova, tindrem una avaluació del rendiment amb més confiança que les d'entrenament. Ens donarà una idea general de com s'ajusten realment les dades i quin serà el millor model classificador. El model triat per aquest conjunt de dades serà més precís per les dades reals que pel d'entrenament i, per tant, tindrà millor capacitat per fer prediccions.

En aquest cas, hem vist que hi ha els mateixos resultats tant pel conjunt de dades d'entrenament com el de prova. Doncs, els Arbres de Decisió tenen un millor rendiment en la majoria dels paràmetres calculats tant en les dades d'entrenament com en els de prova. Tenint en compte que la resta de models també tenen uns paràmetres molt bons, afirmaríem que els Arbres de Decisió són el millor model. Cal destacar, que qualsevol dels tres models que es triés obtindria molt bons resultats.

Guiant-nos pels mateixos criteris, veiem que el model Regressió Logística és el segon millor, ja que té els segons valors més grans en els quatre paràmetres principals.

8. Conclusions

Aquest estudi ens ha permès aprendre i analitzar més profundament el món de la sinistralitat dels automòbils i l'aprenentatge automàtic. S'ha començat amb una anàlisi de totes les variables necessàries pel desenvolupament d'aquest projecte. Quan ens fixem en l'anàlisi de les variables numèriques, observem que, el percentatge més alt de valors punta en aquestes és 21,26% en la variable PROVINI (provisió inicial). Aplicant el preprocessament, hem eliminat totes aquelles variables que tenen un percentatge molt alt de valors buits. Hem vist que, per exemple, la variable CULPOBJ (culpa objectiva) té un 96,72% de valors nuls.

Si ens centrem en les variables categòriques, podem veure que 99,74% dels registres no han tingut cap vehicle extern involucrat (NUMVEH). També hem vist que el 73,23% dels sinistres no han necessitat assistència a la carretera (INDASV) i que només 20048 de 415884 han tingut danys propis.

Un cop hem eliminat tots els valors buits podem afirmar que les variables numèriques ja no tenen valors punta innecessaris. Tots aquells que ens queden són perquè els hem considerat útils per l'estudi. Un exemple seria en la variable PROVINI, en el boxplot de després del preprocessament podem veure algun valor punta per la part de dalt de la caixa.

Durant aquest treball hem vist i analitzat un problema de mala classificació en les fases del sinistre gràcies als models de classificació. El nostre objectiu principal era trobar el millor model per les nostres dades.

En primer lloc, a l'aplicar els mètodes de classificació, hem obtingut uns bons resultats. Hem vist que el rendiment dels tres models és molt similar, ja que en tots ells, els paràmetres tenen valors molt alts. Destaquem el model d' Arbres de decisió quan busquem una visió més detallada de la detecció dels casos positius, casos negatius o rendiment global del model. En els Arbres de decisió tenim en els paràmetres alguns dels valors més alts tant en les dades d'entrenament com en les dades de prova. En el cas de la precisió i de l'especificitat, hem vist que Naive Bayes té un valor d'1. Però, en la sensibilitat, el valor és molt baix i, per això, no el podem triar com a millor model. Per tant, concloem que el millor model per classificar les dades és l'Arbre de decisió.

Quan necessitem una visió més general de la classificació, buscarem el paràmetre més gran de l'àrea sota la corba. Aquest paràmetre veiem que torna a ser el del model d'Arbres de decisió. Per tant, per una visió més general es repeteix l'opció d'Arbres de decisió com a millor model.

En veure els dos tipus de classificació que tenim segons els paràmetres, podem afirmar que el millor model de classificació per les dades d'entrenament i de prova i, per tant, les nostres dades, és l'**Arbre de decisió**.

Alhora, caldria destacar, que la resta de models també s'adapten molt bé a les nostres dades d'entrenament i de prova. Tots els valors dels paràmetres han resultat ser propers a 1 o iguals a 1. L'únic cas que hem trobat un valor més proper a 0 ha sigut en la sensibilitat del model Naive Bayes. Recordem que la sensibilitat és la proporció de casos positius reals que s'han classificat de manera correcta.

A través dels objectius assolits hem pogut acceptar i descartar les dues hipòtesis que ens hem plantejat a l'inici de l'estudi. La primera hipòtesi ens deia que no hi hauria diferències molt significatives en la comparació entre els diferents models. Com hem pogut anar veient durant el treball, aquesta s'ha complert, ja que les diferències entre paràmetres han sigut mínimes en la majoria de casos.

Per veure si es compleix la segona hipòtesi ens centrarem en l'especificitat. Recordem que aquest paràmetre mesura la proporció de casos negatius reals que han estat classificats correctament en el model. Aquests ens ajuden especialment en la minimització dels falsos positius.

Amb els valors de l'especificitat podem afirmar que el percentatge de falsos positius és molt baix, ja que tots els valors són molt propers o iguals a 1 i, per tant, pràcticament tots els casos negatius reals s'han classificat de forma correcta.

A partir d'aquests resultats, podem dir que, per la companyia, suposa una molt bona noticia, ja que en tots els models ens dona com a resultat una bona classificació. Per tant, podem dir que ens podem fiar de la fase que ens dona el registre i que no hi ha una mala classificació.

Per futures línies de investigació es podrien buscar llibreries de models lineals automatitzades amb molts més models. Buscar tècniques més avançades i millorar més el rendiment dels models. Aplicar tècniques d'aprenentatge profund o mètodes d'assemblatge per millorar la precisió i l'estabilitat.

A títol personal, voldria remarcar el que ha suposat fer aquest treball per a mi. Com ja he esmentat anteriorment, les dades han estat extretes de la companyia Catalana Occident, lloc on he realitzat unes pràctiques curriculars i unes no curriculars. Amb la realització d'aquest estudi he pogut apropar-me més a la realitat de les dades que diàriament he hagut de tractar durant les pràctiques.

9. Materials utilitzats

9.1. Webgrafia

Assegurances de cotxe, llar, vida, salut, estalvi - Assegurances Catalana Occident. (s. f.). https://www.seguroscatalanaoccidente.com/cat

RPubs - Machine Learning con R y caret. (s. f.). https://rpubs.com/Joaquin_AR/383283

Rédac, T. (2022, 3 agosto). *Machine Learning: definición, funcionamiento, usos*. Formation Data Science | DataScientest.com. https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

Institut d'Estudis Catalans - Diec2. (s. f.). https://dlc.iec.cat/

Galán, J. S. (2022). Aseguradora. *Economipedia*. https://economipedia.com/definiciones/aseguradora.html

Morales, F. C. (2022). Seguro del automóvil. *Economipedia*. https://economipedia.com/definiciones/seguro-del-automovil.html

School, T. (2022). Todas las aplicaciones del machine learning. *Tokio School*. https://www.tokioschool.com/noticias/aplicaciones-machine-learning/

Seguro de coche a terceros básico | Protegete mientras circulas. (s. f.). https://www.seguroscatalanaoccidente.com/seguros-coche/terceros-basico

Benites, L. (2022). Distancia de Cook / D de Cook: Definición, Interpretación. *Statologos*. https://statologos.com/distancia-de-los-cocineros/?utm_content=cmp-true

¿Qué es la regresión logística? | IBM. (s. f.). https://www.ibm.com/es-es/topics/logistic-regression

Baños, R. V. (s/f). *ARBRES DE DECISIÓ*. Diposit.ub.edu. Recuperado el 29 de junio de 2023, de https://diposit.ub.edu/dspace/bitstream/2445/22282/1/Arbres%20de%20decisi%C3%B3.pdf

Regresión logística. (s/f). Datatab.es. Recuperado el 29 de junio de 2023, de https://datatab.es/tutorial/logistic-regression

9.2. Apunts classe

- Apunts de l'assignatura Mineria de Dades
- Apunts de l'assignatura Anàlisi Mutilvariant

9.3. Bibliografia

Guillen, M., & Pesantez-Narvaez, J. (2018). MACHINE LEARNING Y MODELIZACIÓN PREDICTIVA PARA LA TARIFICACIÓN EN EL SEGURO DE AUTOMÓVILES. *Anales del Instituto de Actuarios Españoles*, *24*(2018), 123–147. https://doi.org/10.26360/2018_6

Baran, S., & Rola, P. (2022). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. En *arXiv [q-fin.ST]*. http://arxiv.org/abs/2204.06109

Hanafy, M., & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. *Risks*, *9*(2), 42. https://doi.org/10.3390/risks9020042

Alejandro, D., & Rendón, O. (s/f). *FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES*. Ucm.es. Recuperado el 28 de junio de 2023, de https://eprints.ucm.es/id/eprint/62519/1/M%C3%A9todos%20de%20aprendizaje%20autom%C3%A1tico%20aplicados%20a%20la%20industria%20aseguradora%20-%20TFM%20Diego%20Ospina.pdf

Guillen, M. (s/f). *Modelos predictivos del riesgo y aplicaciones a los seguros*. Funcas.es. Recuperado el 28 de junio de 2023, de https://www.funcas.es/wp-content/uploads/2021/05/Nuevos-m%C3%A9todos-de-predicci%C3%B3n-econ%C3%B3mica-con-datos-masivos_4.pdf

Annex

Codi Rstudio

```
# PAQUETS QUE HEM UTILITZAT
library(arules)
library(arulesViz)
library(caret)
library(class)
library(cluster)
library(corrplot)
library(dbscan)
library(descr)
library(dplyr)
library(e1071)
library(factoextra)
library(FactoMineR)
library(fpc)
library(ggplot2)
library(ggpubr)
library(Hmisc)
library(ISLR)
library(kableExtra)
library(knitr)
library(lessR)
library(MASS)
library(Matrix)
library(mice)
library(naivebayes)
library(pROC)
#library(psy)
#library(psych)
library(purrr)
library(randomForest)
library(rattle)
library(readr)
library(rpart)
library(rpart.plot)
library(scales)
library(tidyverse)
library(vcd)
library(VIM)

# BASE DE DADES
dades_d19 <- read.csv("F:/uni/tfg/Sinistres 2019 - Desembre/MENSUAL-SINIESTROS.G31D91.TXT",
sep=";")

# ELIMINAR VARIABLES SOBRANTS
for(i in 66:64){
  dades_d19 <- dades_d19[,-i]
}
# CANVI DE NOM DE LA BASE DE DADES
dades <- dades_d19

# ELIMINAR VARIABLES CONFIDENCIALS
for(i in 10:5){
  dades <- dades[,-i]
}
for(i in 12:10){
  dades <- dades[,-i]


}
for(i in 54:46){
  dades <- dades[,-i]
}
```

```
# ELIMINAR BASE DE DADES NO NECESSÀRIA
rm(dades_d19)

# CANVI DE FORMAT EN LES VARIABLES
## VARIABLES CATEGÒRIQUES
for(i in 1:4){
    dades[,i] <- as.factor(dades[,i])
    }

for(i in 10:28){
    dades[,i] <- as.factor(dades[,i])
}

for(i in 40:45){
    dades[,i] <- as.factor(dades[,i])
}

### Dates
for(i in 5:6){
    dades[,i] <- as.Date(dades[,i],format="%d.%m.%y")
}

dades[,7] <- as.factor(dades[,7])

for(i in 8:9){
    dades[,i] <- as.Date(dades[,i],format="%d.%m.%y")
}

## VARIEBLES NUMÈRIQUES

for(i in 29:39){
dades[,i] <- as.numeric(gsub(",", ".", sub("+", '', dades[,i])))

}

# ASSIGNACIÓ DELS VALORS BUITS A NA

## TIPSINI
dades$TIPSINI <- as.character(dades$TIPSINI)
dades$TIPSINI[dades$TIPSINI == "   "] <- NA
dades$TIPSINI <- as.factor(dades$TIPSINI)

## CODAGTI
dades$CODAGTI <- as.character(dades$CODAGTI)
dades$CODAGTI[dades$CODAGTI == "   "] <- NA
dades$CODAGTI <- as.factor(dades$CODAGTI)

## TIPVEHI
dades$TIPVEHI <- as.character(dades$TIPVEHI)
dades$TIPVEHI[dades$TIPVEHI == ""] <- NA
dades$TIPVEHI[dades$TIPVEHI == " "] <- NA
dades$TIPVEHI <- as.factor(dades$TIPVEHI)

## INDCIDE
dades$INDCIDE <- as.character(dades$INDCIDE)
dades$INDCIDE[dades$INDCIDE == " "] <- NA
dades$INDCIDE <- as.factor(dades$INDCIDE)

## INDSDMM
dades$INDSDMM <- as.character(dades$INDSDMM)
dades$INDSDMM[dades$INDSDMM == " "] <- NA
dades$INDSDMM <- as.factor(dades$INDSDMM)
```

```
## INDCULP
dades$INDCULP <- as.character(dades$INDCULP)
dades$INDCULP[dades$INDCULP == " "] <- NA
dades$INDCULP <- as.factor(dades$INDCULP)



## TIPTECD
dades$TIPTECD <- as.character(dades$TIPTECD)
dades$TIPTECD[dades$TIPTECD == "   "] <- NA
dades$TIPTECD <- as.factor(dades$TIPTECD)




## CULPOBJ
dades$CULPOBJ <- as.character(dades$CULPOBJ)
dades$CULPOBJ[dades$CULPOBJ == " "] <- NA
dades$CULPOBJ <- as.factor(dades$CULPOBJ)

# DESCARTAR VARIABLES DE LA SITUACIÓ DE L'ANY ANTERIOR O NO NECESSÀRIES
dades$INDSITU3112 <- NULL
dades$PROV3112    <- NULL
dades$PAGB3112    <- NULL
dades$PAGI3112    <- NULL
dades$RECOB3112   <- NULL
dades$RECOI3112   <- NULL
dades$FECCON      <- NULL
dades$DATSINI     <- NULL
dades$DATOBER     <- NULL
dades$FECCIER     <- NULL
dades$FECREAP     <- NULL
dades$RECOB       <- NULL
dades$RECOI       <- NULL

# GUARDAR BASE DE DADES ORIGINAL
dades_inicial <- dades
# BUSQUEM VARIABLES NUMÈRIQUES
varNum <-which(sapply(dades,is.numeric))

summary(dades$PROVINI)
summary(dades$PROVACT)
summary(dades$PAGB)
summary(dades$PAGI)

# Creació de noves variables

v1 <- as.factor(dades$TOTVEHC)
plot(v1, xlab = "Valors", ylab = "Freqüència", main = "Distribució del nombre de vehicles")
summary(v1)

vb <- c(0, 1, 2)
v2 <- dades$TOTVEHC[dades$TOTVEHC %in% vb]
v2 <- as.factor(v2)
levels(v2)
plot(v2, xlab = "Valors", ylab = "Freqüència", main = "Distribució del valor 0, 1 i 2")

v1 <- dades$TOTVEHC
v1 <- v1[v1 != 0]
v1 <- v1[v1 != 1]
v1 <- v1[v1 != 2]
v1 <- as.factor(v1)
plot(v1, xlab = "Valors", ylab = "Freqüència", main = "Distribució dels valors superiors a
2")
```

```
# CREACIÓ DE VARIABLES
## NOMBRE DE VEHÍCLES
dades$TOTVEHC <- as.character(dades$TOTVEHC)
dades$TOTVEHC <- as.numeric(dades$TOTVEHC)
summary(dades$TOTVEHC)
NUMVEH <- c()
NUMVEH[dades$TOTVEHC < 1] <- 0
NUMVEH[dades$TOTVEHC > 0] <- 1
dades$NUMVEH <- as.factor(NUMVEH)
dades$TOTVEHC <- as.factor(dades$TOTVEHC)
levels(dades$NUMVEH) <- c("0", "< 0")
table(dades$NUMVEH)
rm(NUMVEH)

plot(dades$NUMVEH, xlab = "Valors", ylab = "Freqüència", main = "Distribució final de la no
va variable")

# Gràfica de la distribució
v3 <- as.factor(dades$INDSITU)
plot(v3, xlab = "Valors", ylab = "Freqüència", main = "Distribució de la situació actual de
l sinistre")

## VARIABLE RESPOSTA
levels(dades$INDSITU)
Y <- c()
Y[dades$INDSITU == "0"] <- 0
Y[dades$INDSITU == "1"] <- 1
Y[dades$INDSITU == "2"] <- NA
Y[dades$INDSITU == "3"] <- NA
dades$INDSITU <- as.factor(dades$INDSITU)
dades$Y <- as.factor(Y)
levels(dades$Y) <- c("INICIAL", "FINAL")
table(dades$Y)
rm(Y)
dades2 <- dades

plot(dades$Y, xlab = "Valors", ylab = "Freqüència", main = "Distribució final de la nova va
riable")

varNum <-which(sapply(dades,is.numeric))

# DISTANCIA DE COOK DE PROVINI
k <- varNum[[1]]
x <-dades[,k]
mod <- lm(x ~ 1)
cooksd <- cooks.distance(mod)
plot(cooksd, pch="*", cex=2)
abline(h = 4*mean(cooksd, na.rm=T), col="red")
text(x=1:length(cooksd)+1, y=cooksd,labels=ifelse(cooksd>4*mean(cooksd, na.rm=T), names(coo
ksd),""), col="red")

k <- varNum[[1]]
q1 <- quantile(dades[,k], 0.25, na.rm = TRUE)
q3 <- quantile(dades[,k], 0.75, na.rm = TRUE)
iqr <- q3-q1
ati_tukey <- dades[,k] < (q1-1.5*iqr) | dades[,k] > (q3+1.5*iqr)
dades[,k][ati_tukey] <- NA

# DISTANCIA DE COOK DE PROVACT
k <- varNum[[2]]
x <-dades[,k]
mod <- lm(x ~ 1)
cooksd <- cooks.distance(mod)
plot(cooksd, pch="*", cex=2)
abline(h = 4*mean(cooksd, na.rm=T), col="red")
```

```
text(x=1:length(cooksd)+1, y=cooksd,labels=ifelse(cooksd>4*mean(cooksd, na.rm=T), names(coo
ksd),""), col="red")

k <- varNum[[2]]
q1 <- quantile(dades[,k], 0.15, na.rm = TRUE)
q3 <- quantile(dades[,k], 0.85, na.rm = TRUE)
iqr <- q3-q1
ati_tukey <- dades[,k] < (q1-1.5*iqr) | dades[,k] > (q3+1.5*iqr)
dades[,k][ati_tukey] <- NA

# DISTANCIA DE COOK DE PAGB
k <- varNum[[3]]
x <-dades[,k]
mod <- lm(x ~ 1)
cooksd <- cooks.distance(mod)
plot(cooksd, pch="*", cex=2)
abline(h = 4*mean(cooksd, na.rm=T), col="red")
text(x=1:length(cooksd)+1, y=cooksd,labels=ifelse(cooksd>4*mean(cooksd, na.rm=T), names(coo
ksd),""), col="red")

k <- varNum[[3]]
q1 <- quantile(dades[,k], 0.3, na.rm = TRUE)
q3 <- quantile(dades[,k], 0.7, na.rm = TRUE)
iqr <- q3-q1
ati_tukey <- dades[,k] < (q1-1.5*iqr) | dades[,k] > (q3+1.5*iqr)
dades[,k][ati_tukey] <- NA

# DISTANCIA DE COOK DE PAGI
k <- varNum[[4]]
x <-dades[,k]
mod <- lm(x ~ 1)
cooksd <- cooks.distance(mod)
plot(cooksd, pch="*", cex=2)
abline(h = 4*mean(cooksd, na.rm=T), col="red")
text(x=1:length(cooksd)+1, y=cooksd,labels=ifelse(cooksd>4*mean(cooksd, na.rm=T), names(coo
ksd),""), col="red")

k <- varNum[[4]]
q1 <- quantile(dades[,k], 0.3, na.rm = TRUE)
q3 <- quantile(dades[,k], 0.7, na.rm = TRUE)
iqr <- q3-q1
ati_tukey <- dades[,k] < (q1-1.5*iqr) | dades[,k] > (q3+1.5*iqr)
dades[,k][ati_tukey] <- NA

summary(dades)

# CLASSIFICACIÓ DELS MISSINGS
dMiss <- function(x){sum(is.na(x))/length(x)*100}
round(apply(dades,2,dMiss),2)

# DESCARTAR VARIABLES AMB MÉS DEL 20% DELS VALORS MISSINGS
dades$CODAGTI     <- NULL
dades$TIPSINI     <- NULL
dades$INDSDMM     <- NULL
dades$CULPOBJ     <- NULL
dades$TIPVEHI     <- NULL
dades$INDCIDE     <- NULL

str(dades)
aggr_plot <- aggr(dades, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names
(dades), cex.axis=.7, gap=3, ylab=c("Histograma dels valors missing","Patró"))

dades <- dades[complete.cases(dades), ]
summary(dades)
```

```r
#PROVINI

# INICIAL
k <- which(colnames(dades2) == "PROVINI")
hist(dades2[,k], main = paste("Histograma inicial de", names(dades2)[k]))
boxplot(dades2[,k], main = paste("Boxplot inicial de", names(dades2)[k]))

# FINAL
k <- which(colnames(dades) == "PROVINI")
hist(dades[,k], main = paste("Histograma final de", names(dades)[k]))
boxplot(dades[,k], main = paste("Boxplot final de", names(dades)[k]))

#PROVACT

# INICIAL
k <- which(colnames(dades2) == "PROVACT")
hist(dades2[,k], main = paste("Histograma inicial de", names(dades2)[k]))
boxplot(dades2[,k], main = paste("Boxplot inicial de", names(dades2)[k]))

# FINAL
k <- which(colnames(dades) == "PROVACT")
hist(dades[,k], main = paste("Histograma final de", names(dades)[k]))
boxplot(dades[,k], main = paste("Boxplot final de", names(dades)[k]))

#PAGB

# INICIAL
k <- which(colnames(dades2) == "PAGB")
hist(dades2[,k], main = paste("Histograma inicial de", names(dades2)[k]))
boxplot(dades2[,k], main = paste("Boxplot inicial de", names(dades2)[k]))

# FINAL
k <- which(colnames(dades) == "PAGB")
hist(dades[,k], main = paste("Histograma final de", names(dades)[k]))
boxplot(dades[,k], main = paste("Boxplot final de", names(dades)[k]))

#PAGI

# INICIAL
k <- which(colnames(dades2) == "PAGI")
hist(dades2[,k], main = paste("Histograma inicial de", names(dades2)[k]))
boxplot(dades2[,k], main = paste("Boxplot inicial de", names(dades2)[k]))

# FINAL
k <- which(colnames(dades) == "PAGI")
hist(dades[,k], main = paste("Histograma final de", names(dades)[k]))
boxplot(dades[,k], main = paste("Boxplot final de", names(dades)[k]))

# Anàlisi numèriques

## RAMEMIS
PieChart(RAMEMIS, main="Ram del Grup Catala Occident", data = dades)

## RAMITO
PieChart(RAMITO, main="Ram de gestió", data = dades)

## INDSITU
VC3 <- freq(dades$INDSITU, plot = FALSE)
kable(VC3)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data = dades,INDSITU, main="Situació actual del sinistre")
rm(VC3)

## Y
VC4 <- freq(dades$Y, plot = FALSE)
kable(VC4)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
```

```
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data = dades,Y, main="Variable resposta")
rm(VC4)

## INDTIVE
PieChart(data = dades,INDTIVE, main="Categoria")

## INDRECO
VC8 <- freq(dades$INDRECO, plot = FALSE)
kable(VC8)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data = dades,INDRECO, main="Indicador del sinistre recobrable")
rm(VC8)

## INDCONS
VC9 <- freq(dades$INDCONS, plot = FALSE)
kable(VC9)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data = dades,INDCONS, main="Indicador del sinistre consorciable")
rm(VC9)

## TOTVEHC
PieChart(data = dades,TOTVEHC, main="Nombre total de vehicles externs involucrats")

## CNATSIN
PieChart(data = dades,CNATSIN, main="Naturalesa del sinistre")

## INDCULP
VC11 <- freq(dades$INDCULP, plot = FALSE)
kable(VC11)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data = dades,INDCULP, main="Indicador de culpa")
rm(VC11)

## INDDASE
VC12 <- freq(dades$INDDASE, plot = FALSE)
kable(VC12)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data =dades,INDDASE, main="Indicador de danys del vehicle")
rm(VC12)

## NUMVEH
VC13 <- freq(dades$NUMVEH, plot = FALSE)
kable(VC13)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,NUMVEH, main="Número total de vehícles")
rm(VC13)

## INDCOLD
VC14 <- freq(dades$INDCOLD, plot = FALSE)
kable(VC14)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,INDCOLD, main="Indicador de col·lisió")
rm(VC14)

## IMDCORP
VC15 <- freq(dades$IMDCORP, plot = FALSE)
kable(VC15)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,IMDCORP, main="Indicador de lesionats")
rm(VC15)
```

```
## INDVIDR
VC16 <- freq(dades$INDVIDR, plot = FALSE)
kable(VC16)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,INDVIDR, main="Indicador de vidre")
rm(VC16)

## INDINCE
VC17 <- freq(dades$INDINCE, plot = FALSE)
kable(VC17)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,INDINCE, main="Indicador d'incendi")
rm(VC17)

## INDROBO
VC18 <- freq(dades$INDROBO, plot = FALSE)
kable(VC18)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,INDROBO, main="Indicador de robo")
rm(VC18)

## DANOSMAT
VC19 <- freq(dades$DANOSMAT, plot = FALSE)
kable(VC19)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,DANOSMAT, main="Indicador de danys materials")
rm(VC19)

## INDDAPR
VC20 <- freq(dades$INDDAPR, plot = FALSE)
kable(VC20)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,INDDAPR, main="Indicador de danys propis")
rm(VC20)

## INDASV
VC21 <- freq(dades$INDASV, plot = FALSE)
kable(VC21)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,INDASV, main="Indicador d'assistència de viatge")
rm(VC21)

## TIPTECD
VC22 <- freq(dades$TIPTECD, plot = FALSE)
kable(VC22)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,TIPTECD, main="Tipus declaració del sinistre de tecnologia")
rm(VC22)

## TIPUSUD
PieChart(data=dades,TIPUSUD, main="Tipus declaració del sinistre de tecnologia d'usuari")

## PERDTOT
VC24 <- freq(dades$PERDTOT, plot = FALSE)
kable(VC24)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
PieChart(data=dades,PERDTOT, main="Pèrdua total del vehicle assegurat")
rm(VC24)

## TALPREF
VC25 <- freq(dades$TALPREF, plot = FALSE)
kable(VC25)%>% kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>% column_spec(1, border_right = T)
```

```
PieChart(data=dades,TALPREF, main="Vehicle assegurat reparat pel taller preferent")
rm(VC25)

dades_original <- dades

## MODELS DE CLASSIFICACIÓ

set.seed(123)
mostra<-sample(1:nrow(dades), round(2*nrow(dades)/3))

# Fem un subconjunt per les dades de prova i d'entrenament
dades_entrenament <- dades[mostra, ]
dades_prova <- dades[-mostra, ]
# El subconjunt quan utilitzem les dades amb valors punta
dades_entrenament2 <- dades2[mostra, ]
dades_prova2  <- dades2[-mostra, ]

## ARBRES DE DECISIÓ

# Model amb dades d'entrenament
m_entrenament <- rpart(Y ~ PROVACT + PROVINI + INDRECO + IMDCORP + INDCULP + INDVIDR + TIPT
ECD, data=dades_entrenament2, method="class", parms = list(split="information"))
summary(m_entrenament)
rpart.plot::rpart.plot(m_entrenament)

# Model amb dades de prova
m_prova <- rpart(Y~ PROVACT + PROVINI + INDRECO + IMDCORP + INDCULP + INDVIDR + TIPTECD, da
ta=dades_prova2, method="class", parms = list(split="information"))
summary(m_prova)
rpart.plot::rpart.plot(m_prova)

# Matriu de decisió amb dades d'entrenament
pred_entrenament <- predict(m_entrenament, newdata = dades_entrenament2, type = "class")
matriu_entrenament <- confusionMatrix(pred_entrenament, dades_entrenament2[["Y"]])
matriu_entrenament
# Matriu de decisió amb dades de prova
pred_prova <- predict(m_prova, newdata = dades_prova2, type = "class")
matriu_prova <- confusionMatrix(pred_prova, dades_prova2[["Y"]])
matriu_prova

## PARÀMETRES DE DECISIÓ AMB DADES D'ENTRENAMENT
V_negatiu <- matriu_entrenament$table[2, 2]
V_positiu <- matriu_entrenament$table[1, 1]
F_positiu <- matriu_entrenament$table[1, 2]
F_negatiu <- matriu_entrenament$table[2, 1]
# Accuracy
accuracy_entrenament <- (V_positiu + V_negatiu)/(V_positiu + F_positiu + F_negatiu + V_nega
tiu)
# Precisio
precisio_entrenament <- V_positiu / (V_positiu + F_positiu)
# Sensibilitat
sensibilitat_entrenament <- V_positiu / (V_positiu + F_negatiu)
# Especificitat
especificitat_entrenament <- V_negatiu / (V_negatiu + F_positiu)

pred_ROC <- predict(m_entrenament, newdata = dades_entrenament2, type = "prob") # Probabili
tats de la corba de ROC

PP <- pred_ROC[, "FINAL"] # Probabilitat que sigui positiu

corbaROC <- roc(dades_entrenament2$Y, PP) # Corba de ROC de les dades d'entrenament
plot(corbaROC, main = "Corba ROC")

# Paràmetre de sota la corba
```

```
auc_entrenament <- roc(dades_entrenament2$Y, PP)$auc

# TAULA RESULTATS
nom <- c("Exactitud", "Precisió", "Sensibilitat", "Especificitat", "Àrea sota la corba de R
OC")
param <- c(accuracy_entrenament, precisio_entrenament, sensibilitat_entrenament, especifici
tat_entrenament, auc_entrenament)
data.frame(nom, param)

## PARÀMETRES DE DECISIÓ AMB DADES DE PROVA
V_negatiu <- matriu_prova$table[2, 2]
V_positiu <- matriu_prova$table[1, 1]
F_positiu <- matriu_prova$table[1, 2]
F_negatiu <- matriu_prova$table[2, 1]
# Accuracy
accuracy_prova <- (V_positiu + V_negatiu)/(V_positiu + F_positiu + F_negatiu + V_negatiu)
# Precisio
precisio_prova <- V_positiu / (V_positiu + F_positiu)
# Sensibilitat
sensibilitat_prova <- V_positiu / (V_positiu + F_negatiu)
# Especificitat
especificitat_prova <- V_negatiu / (V_negatiu + F_positiu)

pred_ROC <- predict(m_prova, newdata = dades_prova2, type = "prob")  # Probabilitats de la
corba de ROC

PP <- pred_ROC[, "FINAL"] # Probabilitat que sigui positiu

corbaROC <- roc(dades_prova2$Y, PP) # Corba de ROC de les dades de prova
plot(corbaROC, main = "Corba ROC")

# Calcular el área bajo la corba ROC
auc_prova <- roc(dades_prova2$Y, PP)$auc
# TAULA RESULTATS
nom <- c("Exactitud", "Precisió", "Sensibilitat", "Especificitat", "Àrea sota la corba de R
OC")
param <- c(accuracy_prova, precisio_prova, sensibilitat_prova, especificitat_prova, auc_pro
va)
data.frame(nom, param)

## NAIVE BAYES

dd<-dades
#test<-sample(1:nrow(dd),size = nrow(dd)/3)
#dades_entrenament<-dd[-test,]
#dades_prova<-dd[test,]
#DD_test<-sample(1:nrow(dades_prova),size = nrow(dades_prova)/3)
#DD_train<-sample(1:nrow(dades_entrenament),size = nrow(dades_entrenament)/3)

#head(dd)
xtabs(~Y+INDRECO, data = dd)
xtabs(~Y+IMDCORP, data = dd)
xtabs(~Y+INDCULP, data = dd)
xtabs(~Y+INDVIDR, data = dd)
xtabs(~Y+TIPTECD, data = dd)

# DADES ENTRENAMENT
nb_entrenament <- naive_bayes(Y ~ PROVACT + PROVINI + INDRECO + IMDCORP + INDCULP + INDVIDR
+ TIPTECD, data = dades_entrenament)
summary(nb_entrenament)

# Gràfiques
plot(nb_entrenament, which = "PROVACT", legend=T)
plot(nb_entrenament, which = "PROVINI", legend=T)
plot(nb_entrenament, which = "INDRECO", legend=T)
```

```
plot(nb_entrenament, which = "IMDCORP", legend=T)
plot(nb_entrenament, which = "INDCULP", legend=T)
plot(nb_entrenament, which = "INDVIDR", legend=T)
plot(nb_entrenament, which = "TIPTECD", legend=T)

# Matriu de confusió amb les dades d'entrenament
pred_entrenament <- predict(nb_entrenament, dades_entrenament, type = "class") #model train
k <-which(colnames(dd) == "Y")
matriu_entrenament <- confusionMatrix(pred_entrenament, dades_entrenament[["Y"]])
matriu_entrenament

## METRIQUES DADES TRAIN
V_negatiu <- matriu_entrenament$table[2, 2]
V_positiu <- matriu_entrenament$table[1, 1]
F_positiu <- matriu_entrenament$table[1, 2]
F_negatiu <- matriu_entrenament$table[2, 1]
# Accuracy
accuracy_entrenament <- (V_positiu + V_negatiu)/(V_positiu + F_positiu + F_negatiu + V_nega
tiu)
# Precisio
precisio_entrenament <- V_positiu / (V_positiu + F_positiu)
# Sensibilitat
sensibilitat_entrenament <- V_positiu / (V_positiu + F_negatiu)
# Especificitat
especificitat_entrenament <- V_negatiu / (V_negatiu + F_positiu)

pred_ROC <- predict(nb_entrenament, newdata = dades_entrenament, type = "prob") # Probabili
tat de la corba de ROC

PP <- pred_ROC[, "FINAL"] # Probabilitat que sigui positiu

corbaROC <- roc(dades_entrenament$Y, PP) # Corba de ROC de les dades d'entrenament
plot(corbaROC, main = "Corba ROC")

# Paràmetre de sota la corba
auc_entrenament <- roc(dades_entrenament$Y, PP)$auc

# TAULA RESULTATS
nom <- c("Exactitud", "Precisió", "Sensibilitat", "Especificitat", "Àrea sota la corba de R
OC")
param <- c(accuracy_entrenament, precisio_entrenament, sensibilitat_entrenament, especifici
tat_entrenament, auc_entrenament)
data.frame(nom, param)

# DADES DE PROVA
nb_prova <- naive_bayes(Y ~ PROVACT + PROVINI + INDRECO + IMDCORP + INDCULP + INDVIDR + TIP
TECD, data = dades_prova)
summary(nb_prova)
```

Figura_: Variable resposta Y amb la provisió actual

```
plot(nb_prova, which = "PROVACT", legend=T)
plot(nb_prova, which = "PROVINI", legend=T)
plot(nb_prova, which = "INDRECO", legend=T)
plot(nb_prova, which = "IMDCORP", legend=T)
plot(nb_prova, which = "INDCULP", legend=T)
plot(nb_prova, which = "INDVIDR", legend=T)
plot(nb_prova, which = "TIPTECD", legend=T)

# Matriu de confusió de les dades de prova
k <-which(colnames(dd) == "Y")
pred_prova <- predict(nb_prova, dades_prova, type = "class")#model test
matriu_prova <- confusionMatrix(pred_prova, dades_prova[["Y"]])
matriu_prova
```

```
V_negatiu <- matriu_prova$table[2, 2]
V_positiu <- matriu_prova$table[1, 1]
F_positiu <- matriu_prova$table[1, 2]
F_negatiu <- matriu_prova$table[2, 1]
# Accuracy
accuracy_prova <- (V_positiu + V_negatiu)/(V_positiu + F_positiu + F_negatiu + V_negatiu)
# Precisio
precisio_prova <- V_positiu / (V_positiu + F_positiu)
# Sensibilitat
sensibilitat_prova <- V_positiu / (V_positiu + F_negatiu)
# Especificitat
especificitat_prova <- V_negatiu / (V_negatiu + F_positiu)

pred_ROC <- predict(nb_prova, newdata = dades_prova, type = "prob") # Probabilitat de la co
rba de ROC

PP <- pred_ROC[, "FINAL"] # Probabilitat que sigui positiu

corbaROC <- roc(dades_prova$Y, PP) # Corba de ROC de les dades d'entrenament
plot(corbaROC, main = "Corba ROC")

auc_prova <- roc(dades_prova$Y, PP)$auc

# TAULA RESULTATS
nom <- c("Exactitud", "Precisió", "Sensibilitat", "Especificitat", "Àrea sota la corba de R
OC")
param <- c(accuracy_prova, precisio_prova, sensibilitat_prova, especificitat_prova, auc_pro
va)
data.frame(nom, param)

## Regressió logística

# Dades d'entrenament
model_entrenament <- glm(Y~ PROVINI + PROVACT + IMDCORP + INDRECO+ INDCULP + INDVIDR + TIPT
ECD, data = dades_entrenament, family = binomial)
print(model_entrenament)

plot(dades_entrenament$PROVACT, dades_entrenament$PROVINI, col = ifelse(dades_entrenament$Y
== "INICIAL", "blue",
"red"), xlab = "PROVACT", ylab = "PROVINI")

points(dades_entrenament$PROVACT, dades_entrenament$PROVINI, col = ifelse(predict(model_ent
renament, type = "response") > 0.5, "lightblue", "lightcoral"), pch = 20)
legend("topright", legend = levels(dades_entrenament$Y), col = c("blue", "red"), pch = 1)

# Dades de prova
model_prova <- glm(Y~ PROVINI + PROVACT + IMDCORP + INDRECO + INDCULP + INDVIDR + TIPTECD,
data = dades_prova, family = binomial)
print(model_prova)

plot(dades_prova$PROVACT, dades_prova$PROVINI, col = ifelse(dades_prova$Y== "INICIAL", "blu
e",
"red"),  xlab = "PROVACT", ylab = "PROVINI")

points(dades_prova$PROVACT, dades_prova$PROVINI, col = ifelse(predict(model_prova, type = "
response") > 0.5, "lightblue", "lightcoral"), pch = 20)
legend("topright", legend = levels(dades_prova$Y), col = c("blue", "red"), pch = 1)

## Matriu de confusió de les dades d'entrenament
pred_entrenament <- predict(model_entrenament, newdata = dades_entrenament, type = "respons
e")
classe_entrenament <- ifelse(pred_entrenament >= 0.5, "FINAL", "INICIAL")
classe_entrenament <- factor(classe_entrenament, levels = levels(dades_entrenament$Y))
```

```
matriu_entrenament <- confusionMatrix(classe_entrenament, dades_entrenament$Y)
matriu_entrenament
# Matriu de confusió de les dades de prova
pred_prova <- predict(model_prova, newdata = dades_prova, type = "response")
classe_prova <- ifelse(pred_prova >= 0.5, "FINAL", "INICIAL")
classe_prova <- factor(classe_prova, levels = levels(dades_prova$Y))
matriu_prova <- confusionMatrix(classe_prova, dades_prova$Y)
matriu_prova

# PARÀMETRES DE DECISIÓ AMB LES DADES D'ENTRENAMENT
V_negatiu <- matriu_entrenament$table[2, 2]
V_positiu <- matriu_entrenament$table[1, 1]
F_positiu <- matriu_entrenament$table[1, 2]
F_negatiu <- matriu_entrenament$table[2, 1]
# Accuracy
accuracy_entrenament <- (V_positiu + V_negatiu)/(V_positiu + F_positiu + F_negatiu + V_nega
tiu)
# Precisio
precisio_entrenament <- V_positiu / (V_positiu + F_positiu)
# Sensibilitat
sensibilitat_entrenament <- V_positiu / (V_positiu + F_negatiu)
# Especificitat
especificitat_entrenament <- V_negatiu / (V_negatiu + F_positiu)

pred_ROC <- predict(model_entrenament, newdata = dades_entrenament, type = "response") # Pr
obabilitats de la corba d'entrenament

corbaROC <- roc(dades_entrenament$Y, pred_ROC) # Corba de ROC de els dades d'entrenament
plot(corbaROC, main = "Corba ROC")

# Paràmetre de sota la corba
auc_entrenament <- roc(dades_entrenament$Y, pred_ROC)$auc

# TAULA RESULTATS
nom <- c("Exactitud", "Precisió", "Sensibilitat", "Especificitat", "Àrea sota la corba de R
OC")
param <- c(accuracy_entrenament, precisio_entrenament, sensibilitat_entrenament, especifici
tat_entrenament, auc_entrenament)
data.frame(nom, param)

# PARÀMETRES DE DECISIÓ AMB LES DADES DE PROVA
V_negatiu <- matriu_prova$table[2, 2]
V_positiu <- matriu_prova$table[1, 1]
F_positiu <- matriu_prova$table[1, 2]
F_negatiu <- matriu_prova$table[2, 1]
# Exactitud / Accuracy
accuracy_prova <- (V_positiu + V_negatiu)/(V_positiu + F_positiu + F_negatiu + V_negatiu)
# Precisio
precisio_prova <- V_positiu / (V_positiu + F_positiu)
# Sensibilitat
sensibilitat_prova <- V_positiu / (V_positiu + F_negatiu)
# Especificitat
especificitat_prova <- V_negatiu / (V_negatiu + F_positiu)

pred_ROC <- predict(model_prova, newdata = dades_prova, type = "response") # Probabilitat d
e la corba de ROC

corbaROC <- roc(dades_prova$Y, pred_ROC) # Corba de ROC de les dades de prova
plot(corbaROC, main = "Corba ROC")

# Paràmetre de sota la corba
auc_prova <- roc(dades_prova$Y, pred_ROC)$auc
```

```
# TAULA RESULTATS
nom <- c("Exactitud", "Precisió", "Sensibilitat", "Especificitat", "Àrea sota la corba de R
OC")
param <- c(accuracy_prova, precisio_prova, sensibilitat_prova, especificitat_prova, auc_pro
va)
data.frame(nom, param)
```