TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Tècniques de classificació mitjançant *Machine Learning* aplicades al problema de la sobrequalificació.

Autor: Marina Serra Dalmau

Director: Prof. Fco Javier Sierra Martínez

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: Juny, 2023

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum

La sobrequalificació és un factor molt present en els treballadors, es produeix quan disposen d'un nivell educatiu o un conjunt d'habilitats que excedeixen els requisits necessaris per desenvolupar el seu treball actual. Aquest problema pot sorgir a causa de diversos factors, com a canvis en els requisits d'ocupació, la falta d'oportunitats laborals adequades o la cerca d'ocupació en temps de crisi econòmica.

Per altra banda, el *Machine Learning* és una branca de la intel·ligència artificial que se centra en el desenvolupament d'algorismes i models capaços d'aprendre i prendre decisions a partir de dades sense ser programats explícitament. Els algorismes de classificació són una de les principals tècniques utilitzades en l'aprenentatge automàtic per a classificar dades en diferents classes o categories.

Aquest projecte té com a objectiu aplicar tècniques de classificació mitjançant l'aprenentatge automàtic (*Machine Learning*) per a abordar el problema de la sobrequalificació laboral. Per això, s'utilitzen quatre algoritmes de classificació que requereixen tècniques estadístiques per obtenir classificacions i avaluar el seu rendiment, que són: Arbres de decisió, *Naive Bayes*, *Support Vector Machines* i Regressió Logística.

S'avaluaran els algoritmes esmentats i se seleccionarà el millor model de classificació per les dades del projecte. A partir d'aquest model es buscarà un perfil d'estudiant més probable a ser sobrequalificat.

Paraules clau: Sobrequalificació, *Machine Learning*, intel·ligència artificial, models de classificació, Arbres de decisió, *Naive Bayes*, *Support Vector Machines* i Regressió Logística.

Abstract

Overqualification is a factor that is very present in workers, occurring when they have either an educational level or set of skills that exceed the requirements necessary to carry out their current work. This problem may arise due to various factors and instances, such as changes in employment requirements, the lack of adequate employment opportunities or the search for employment in times of economic crisis.

Machine Learning is a branch of artificial intelligence that focuses on the development of algorithms and models capable of learning and making decisions from data without being explicitly programmed, classification algorithms are one of the main techniques used in machine learning to classify data in different classes or categories.

This project aims to apply grading techniques through Machine Learning to address the problem of over-qualification. To achieve these goal, four classification algorithms are used which require statistical techniques to obtain classifications and evaluations of their performance. Those algorithms are Decision trees, Naive Bayes, Support Vector Machines and Logistic Regression.

The above algorithms will be evaluated, and compared, to identify the best classification model for the project. From this model a student profile more likely to be over-qualified will be sought.

Keywords: Overqualification, Machine Learning, artificial intelligence, classification models, decision trees, Naive Bayes, Support Vector Machines and Logistic Regression.

Classificació AMS

62-XX STATISTICS

- 62-07 Data analysis
- 62Q05 Statistical tables

68Txx Artificial intelligence

- 68Wxx Algorithms
- 68W40 Analysis of algorithms

ÍNDEX

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

1. INTRODUCCIÓ

La sobrequalificació laboral és un fenomen cada vegada més rellevant en l'àmbit laboral actual. Es refereix a la situació en la qual els treballadors disposen un nivell educatiu o conjunt d'habilitats que supera els requisits necessaris per a exercir el seu treball actual. Aquesta problemàtica pot tenir repercussions tant per als treballadors com per a les organitzacions, generant insatisfacció laboral, falta de desenvolupament professional, menor productivitat i una major rotació de personal.

En aquest context, l'ús de tècniques de classificació mitjançant l'aprenentatge automàtic (*Machine Learning*) es presenta com una eina important per a abordar el problema de la sobrequalificació. El *Machine Learning* permet entrenar models capaços d'aprendre de les dades i realitzar classificacions i prediccions precises sobre la base i característiques identificades en ells.

En aquest apartat es detallarà l'origen del treball, la metodologia utilitzada, les hipòtesis i els objectius del treball, brindant una visió general de l'enfocament i l'estructura de l'estudi sobre tècniques de classificació aplicades a la sobrequalificació laboral.

1.1 Origen del treball i motivació

El present treball té el seu origen en una idea inicial relacionada amb l'àmbit educatiu. Inicialment, es tenia l'interès de dur a terme un estudi sobre l'adequació entre la formació acadèmica dels individus i les seves ocupacions laborals. No obstant això, en una interacció amb el professor, va sorgir una oportunitat interessant per abordar un tema estretament relacionat: la sobrequalificació laboral.

El professor va proposar utilitzar una base de dades que permetia calcular el factor de la sobrequalificació per a un conjunt d'estudiants recentment graduats. Aquesta base de dades contenia informació detallada sobre l'educació, l'experiència laboral i les tasques que duien a terme els individus en els seus llocs de treball, entre d'altres. Aquesta proposta va captar la meva atenció, ja que em va permetre abordar un tema actual i rellevant en l'àmbit laboral.

La motivació principal d'aquest treball és contribuir a la comprensió de la sobrequalificació laboral i explorar com les tècniques de classificació mitjançant l'aprenentatge automàtic poden ajudar a identificar, predir i entendre millor aquest fenomen.

1.2 Objectius del treball

Els objectius d'aquest treball són abordar i aprofundir en la problemàtica de la sobrequalificació laboral des de diferents perspectives.

En primer lloc, es busca analitzar i comprendre a fons el fenomen de la sobrequalificació laboral. Això implica investigar i identificar els factors que contribueixen a aquesta situació, com ara els desajustos entre les habilitats dels treballadors i els requisits dels llocs de treball. Mitjançant aquesta anàlisi, es pretén obtenir una visió clara i una comprensió en profunditat del fenomen.

En segon lloc, es proposa aplicar i avaluar diverses tècniques de classificació mitjançant l'aprenentatge automàtic per abordar el problema de la sobrequalificació. Això implica la selecció i implementació d'algoritmes de classificació adequats i la seva aplicació a conjunts de dades rellevants. Es busca avaluar l'eficàcia d'aquests models mitjançant mètriques d'avaluació per determinar la seva capacitat per predir correctament la sobrequalificació.

Finalment, un objectiu específic és trobar un perfil de treballador sobrequalificat a través del millor model de classificació escollit. Aquest perfil proporcionarà informació valuosa sobre les característiques, competències i altres factors associats als treballadors que es troben en situació de sobrequalificació.

Per a dur a terme aquest treball, es van formular les següents hipòtesis de recerca que es van buscar validar mitjançant la comprensió de la sobrequalificació, l'anàlisi i aplicació de tècniques de classificació mitjançant l'aprenentatge automàtic:

- Hipòtesi 1: Un factor que influeix en la sobrequalificació laboral serà la branca de titulació dels treballadors.
- Hipòtesi 2: S'espera que hi hagi diferències significatives en la comparació dels diferents models de classificació fets servir per predir la sobrequalificació laboral.
- Hipòtesi 3: S'espera que els treballadors que pertanyen a branca de titulació Arts i Humanitats tinguin una major probabilitat d'experimentar sobrequalificació.

2. METODOLOGIA

En aquest apartat, es descriurà la metodologia usada per dur a terme el treball de tècniques de classificació mitjançant l'aprenentatge automàtic aplicades al problema de la sobrequalificació. El projecte s'ha estructurat en dos grans blocs, una part teòrica i una altra pràctica.

En la part teòrica del treball, s'ha realitzat una revisió exhaustiva sobre el fenomen de la sobrequalificació laboral i els algorismes de classificació emprats en el camp de l'aprenentatge automàtic.

En relació amb la sobrequalificació laboral, s'ha explorat en profunditat la seva definició i els factors que poden influir en la seva aparició. S'ha analitzat com la discrepància entre el nivell educatiu o les habilitats d'un treballador i els requisits del lloc de treball pot generar una situació de sobrequalificació.

Pel que fa als algorismes de classificació usats en l'aprenentatge automàtic, s'ha proporcionat una descripció detallada de diferents tècniques. S'han explorat els arbres de decisió, que prenen decisions basades en una estructura jeràrquica de regles; *Naive Bayes*, que fa servir la teoria de la probabilitat per a fer classificacions; les màquines de vectors de suport (SVM), que troben l'hiperplà òptim per a separar les classes; i Regressió Logística, que es basa en una funció logística per estimar la probabilitat de pertànyer en una classe especifica.

En la part pràctica s'ha implementat els algoritmes esmentats. A continuació es detallen les principals etapes seguides:

1. Càlcul de la sobrequalificació: Per començar, s'ha calculat el factor de sobrequalificació de les dades obtingudes a partir d'una enquesta.
2. Tractament previ de les dades: Aquesta etapa implica la neteja i transformació de les dades per assegurar la seva qualitat i preparar-les per a l'anàlisi posterior. S'han dut a terme tasques com la gestió de valors mancants o outliers, la selecció de les característiques més rellevants i la divisió de les dades en conjunts d'entrenament i prova.
3. Implementació i entrenament dels models: A continuació, s'ha procedit a implementar i entrenar els models de classificació seleccionats. Això implica la utilització d'eines o llenguatges de programació específics per a l'aprenentatge automàtic, en aquest cas R.

4. Avaluació dels resultats: S'ha avaluat els resultats obtinguts dels models de classificació. Això implica l'ús dels conjunts de dades de prova per avaluar el rendiment i la precisió dels models. S'han utilitzat diverses mètriques d'avaluació, com ara l'exactitud, la precisió, la sensibilitat, l'especificitat i l'àrea sota la corba ROC (AUC-ROC), per mesurar l'efectivitat dels models en la classificació de la sobrequalificació laboral.
5. Perfil de treballador sobrequalificat: Finalment, a partir del millor model de classificació, s'ha buscat un perfil de treballador que sigui més probable a ser sobrequalificat.

3. ESTAT DE L'ART

En aquesta secció del treball s'aborda tota la informació teòrica i l'estat de l'art relacionats amb els temes principals d'aquest projecte, que són la sobrequalificació, la intel·ligència artificial (IA) i l'estadística. En particular, s'estudia detalladament la sobrequalificació, present en els treballadors, i el camp de l'Aprenentatge Automàtic (*Machine Learning*), que és una subdisciplina de la intel·ligència artificial.

La intel·ligència artificial i l'estadística estan estretament relacionades, encara que també presenten unes certes diferències. No obstant això, es complementen entre si i generen algorismes i mètodes complexos que resulten útils per a l'anàlisi i estudi de les dades.

3.1 Sobrequalificació

La sobrequalificació s'utilitza per descriure un treballador que adquireix un nivell d'educació i habilitats que excedeixen els requisits d'un lloc de treball específic. Això es deu al fet que el treballador pot obtenir un lloc de treball que no requereixi un títol de postgrau, que ell/ella té. Pot donar com a resultat una situació en què el treballador no està explotant al màxim la seva formació i experiència.

Existeixen nombroses investigacions sobre les conseqüències de la sobrequalificació i la seva relació amb la satisfacció del treballador. És important distingir entre sobrequalificació objectiva i percebuda. La sobrequalificació objectiva és aquella que es produeix quan la formació o experiència del treballador està objectivament per sobre de la requerida pel lloc de treball. Així i tot, en aquest cas es pot donar la situació que un empleat objectivament tingui una formació superior a la requerida pel lloc de treball i que ell no ho vegi com un aspecte negatiu, és a dir, que valori altres aspectes com la conciliació familiar o els valors de l'empresa i no el contingut estricte del lloc en si i el seu grau d'adequació. Per altra banda, la sobrequalificació percebuda és quan un treballador creu que té una formació, experiència o habilitats que superen les requerides pel lloc de treball, i ho interpreten de forma negativa (Gavilá, 2014).

La sobrequalificació és una forma de desajust del mercat laboral que ha generat nombrosos estudis que intenten analitzar el seu significat i rellevància en el mercat laboral. Autors com Berg (1970), Freeman (1976) i Smith i Welch (1978) han estudiat el fenomen de la sobrequalificació des dels seus inicis, mentre McGuinness (2006), García Montalvo (2008), Leuven i Oosterbeek (2011), Capsada-Munsech (2017) i, Nieto i Ramos (2017) ofereixen

algunes revisions de l'extensa literatura sobre la sobrequalificació i els seus factors determinants.

Nombrosos estudis han intentat determinar la importància de la sobrequalificació, així com la seva relació amb la teoria del capital humà i amb els factors socials i econòmics. Autors com Duncan i Hoffman (1981), Rumberger (1987), Verdugo i Verdugo (1989) o Alba-Ramirez (1993) han analitzat la relació de la sobrequalificació amb els salaris, com el rendiment de la inversió en el capital humà, i les característiques d'aquest capital humà. Per altra banda, estudis, segons Becker (1964) i Mincer (1974), diuen que el capital humà acumulat per les persones durant la seva educació i experiència laboral, els prepara per accedir a llocs de treball en els que poden desenvolupar un nivell de productivitat que, al mateix temps, haurien d'estar a l'altura dels seus salaris. A més a més, també hi hauria d'existir un equilibri financer entre aquests salaris i la seva inversió prèviament realitzada en capital humà per les persones empleades, sent el seu salari el retorn de la seva inversió. D'aquesta manera, no hi hauria lloc per la sobrequalificació en el lloc de treball, especialment en el cas dels graduats. (Turmo-Garuz, Bartual-Figueras, & Sierra-Martinez, 2019)

3.1.1 Factors associats a la sobrequalificació

Un conjunt d'estudis afirmen que la sobrequalificació presenta tant com un problema de frustració personal com un problema de tipus econòmic, és a dir, augmenta la taxa d'atur i disminueix la productivitat, tant en l'àmbit d'empreses com del país.

La *Teoria del Capital Humà* prediu una relació positiva entre el capital humà i els salaris percebuts pels treballadors. A mesura que augmenta el nivell educatiu d'un treballador, augmentarà la productivitat i amb això el salari. No obstant això, amb el mateix nivell de formació, caldria esperar que els treballadors que utilitzen en més proporció les habilitats adquirides siguin més productius i, per tant, obtinguin majors salaris que aquells que estan sobrequalificats. En aquest sentit, el salari més baix, podria ser causat a què la productivitat del treballador sobrequalificat és menor que els que si corresponen a un lloc de treball acord amb la seva qualificació. La mateixa teoria prediu una correlació negativa entre els salaris dels treballadors i la sobrequalificació, això significa que hi ha persones amb un nivell educatiu superior al que exigeix el seu lloc de treball; aquestes persones estan sobrequalificades pels llocs de treball que ocupen. (González & Miles, 2021)

A partir de la *Teoria del Capital Humà*, existeixen dues teories que expliquen els processos de mobilitat laboral, la *Teoria de l'Acoblament al Lloc de treball* i la *Teoria de la mobilitat Professional* (García, 1998). La *Teoria de l'Acoblament al Lloc de treball*, considera que la

mobilitat interna (dins de la mateixa empresa) i l'externa (entre empreses) respon a la cerca de l'adequació del treballador al lloc de treball. Els desajustos en la correspondència entre qualificació i lloc de treball afavoreixen la cerca d'equilibri i, per tant, la mobilitat professional. Fenòmens com la sobrequalificació i la no qualificació influeixen en la mobilitat professional, els sobrequalificats tenen més probabilitats de canviar d'empresa i de lloc de treball que els no qualificats, que a més a més tenen menor grau d'ocupabilitat. Per altra banda, la *Teoria de la mobilitat Professional* prediu que la sobrequalificació dels titulats seria un fenomen transitori, és a dir, la falta d'experiència laboral podria ser un dels factors que ocasiona l'acceptació d'un primer lloc de treball que no requereix estudis universitaris (Sicherman & Galor, 1990). Tot i això, amb experiència es podria escalar cap a llocs de treball acords amb el nivell educatiu.

Les *Teories* Credencialistes consideren que el nivell educatiu es converteix en un factor decisiu per a justificar l'accés a les posicions de major responsabilitat. Considerant el nivell educatiu com un filtre i punt de referència al moment de contractar treballadors en relació amb les tasques professionals. (Turmo-Garuz, Bartual-Figueras, & Sierra-Martinez, 2019). Segons aquestes teories, els empresaris contracten en funció del nivell educatiu o del que transmet el treballador, ja que el que transmet o nivell educatiu revela el cost de formació dels empleats; com major sigui el nivell educatiu, menor és el cost de formació. En aquest context, podem dir que les persones contractades competeixen pel lloc de treball, per la qual cosa un augment de l'oferta de mà d'obra pot provocar un excés de formació.

La *Teoria de la senyalització* desenvolupada per Spence (1973), considera que existeix una asimetria d'informació entre empresaris i treballadors a l'hora de contractar (Turmo-Garuz, Bartual-Figueras, & Sierra-Martinez, 2019). Mentre que els treballadors coneixen el seu grau de qualificació i coneixements, els empresaris, en teoria, no tenen forma de verificar aquesta informació i, per tant, es guiaran pels signes observats o senyals del mercat: el nivell d'estudis (Saturino Martínez García, 2017). Per això, les empreses hauran de contractar a aquelles persones que destaquin pel que han transmès, independentment del nivell educatiu requerit per desenvolupar les funcions del lloc de treball (Corrales Galán, 2021).

En resum, a partir de les teories, es pot identificar els principals factors generals associats a la sobrequalificació:

- Els relacionats amb la integració en el mercat laboral, com els salaris o la satisfacció en el treball.
- La mobilitat professional abans i després de la contractació.
- Credencials, el rendiment acadèmic.

3.1.2 Impacte de la sobrequalificació a Espanya

Des dels últims anys, Espanya ha experimentat un gran creixement del nombre de graduats universitaris que entren en el mercat laboral, sent aquest increment significatiu particularment en les dones. Això, l'ha convertit en un dels països amb major percentatge de titulats universitaris, però també un dels que més taxa de sobrequalificació presenta.

Tot i no ser Espanya l'únic país que pateix sobrequalificació en els treballadors, és important, ja que organismes internacionals com l'Organització per la Cooperació i el Desenvolupament Econòmics (OCDE 2007) situa el nivell de sobrequalificació a Espanya al capdavant de tots els altres països de l'organització amb un nivell superior al doble de la mitjana. Estudis tant en l'àmbit general (García Montalvo, Peiró, & Soro Bonmatí, 2003) com per als universitaris en particular (García Montalvo, 2001) (García Montalvo, Soro Bonmatí, & Peiró Silla, 2006), afirmen que el fenomen de la sobrequalificació està molt estès al mercat laboral espanyol.

Segons Eurostat, afirmen que Espanya és el país amb major taxa de sobrequalificació de la UE: el 38.8% dels joves espanyols tenen una formació superior a la necessària pel lloc de treball, mentre que la mitjana europea és del 24.1%.

En la Figura 3.1 mostra que Espanya juntament amb Grècia i Xipre són els països amb major nivell de sobrequalificació. Les dades revelen que, en mitjana, un de cada quatre joves titulats universitaris europeus està sobrequalificat. No obstant això, hi ha heterogeneïtat significativa entre països. Així, Luxemburg és el país amb una taxa de sobrequalificació menor, amb un 9.7%, mentre que a l'altre extrem hi ha Grècia amb un 48%.

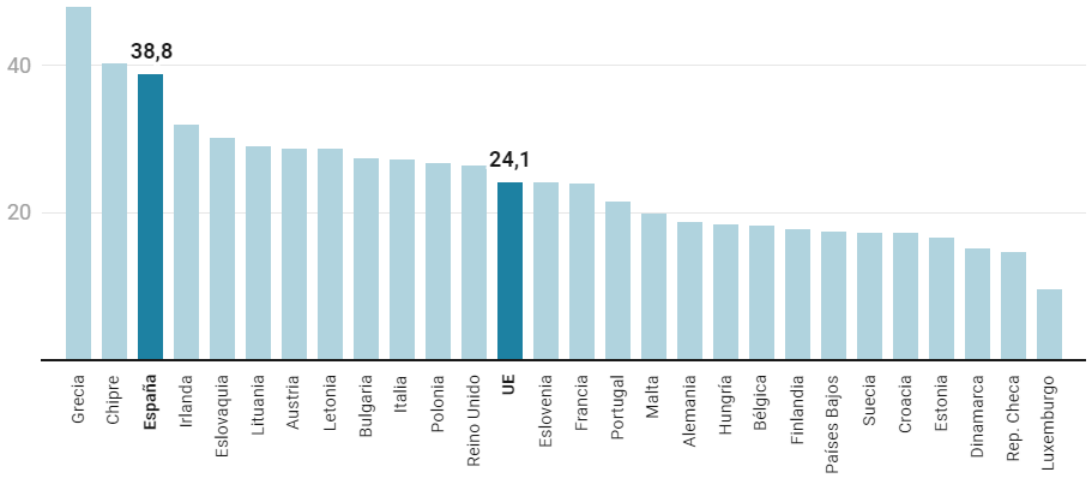


Figura 3.1: Taxa de sobrequalificació dels graduats universitaris en la UE el 2019. Font: Eurostat.

En l’article “Sobrecualificación o falta de oportunidades laborales: un análisis sectorial en España” (Montes Pineda, Garrido Yuste, & Gallo, 2019), s’argumenta que:

> "A nivell macroeconòmic, la producció nacional és potencialment menor del que podria ser si les habilitats dels treballadors sobrequalificats s'utilitzessin òptimament”.

El cas és que, des de la inversió en formació, els titulats es troben amb un gran desajust entre l'oferta i la demanda del mercat laboral. A això cal afegir unes altes taxes de desocupació juvenil superiors al 38% que desemboquen moltes vegades, a acceptar llocs de treball que no requereixen qualificació, o que estan per sota del nivell d'estudis ("és millor això que res"). (Soler, 2021)

El desajust s'observa en les dades d'ocupació. El 73% dels graduats espanyols que han acabat els seus estudis en els últims tres anys estan treballant, segons Eurostat. Tot i això, la mitjana de la Unió Europea és del 80%. Un dels factors més importats que fa que aquests números siguin tan elevats, és l'impacte de la COVID-19. Un factor que ha afectat, sobretot, a les empreses petites i mitjanes, i entre elles als sectors d'hostaleria, oci i serveis i transports. Això ha fet que hi hagi menys oportunitats laborals i ha fet que els estudiants continuïn amb els seus estudis.

Independentment d'aquest factor que ha afectat greument, és evident que els universitaris espanyols tenen més dificultat per accedir a un lloc de treball acord amb el seu nivell de qualificació a diferència d'altres països de la Unió Europea. Es pot observar com aquells països amb menors taxes de sobrequalificació són també, en general, aquells que tenen les dades més baixes de desocupació.

3.1.2.1 Enquesta INE

El 2019, l'INE va realitzar una enquesta sobre la Inserció Laboral dels Titulats Universitaris el 2014 (i en actiu 2019), i es pot veure com hi ha diferències entre els graduats segons els àmbits d'estudis, l'experiència laboral posterior al primer treball, la grandària de l'empresa i inclús el salari percebut.

- **Ajust educatiu i experiència laboral**

L'experiència laboral adquirida en el primer treball pot contribuir a la millora en la qualitat del treball. Transcorreguts cinc anys des de la graduació, no tots els titulats tenen la mateixa experiència laboral.

En la Figura 3.2 s'observa el percentatge de titulats que estan sobrequalificats segons l'experiència laboral. Es pot veure com el percentatge és del 36% quan l'experiència laboral és menor de sis mesos, i d'un 20% quan és de més de dos anys, això significa que el percentatge de sobrequalificació s'ha disminuït un **44%**.

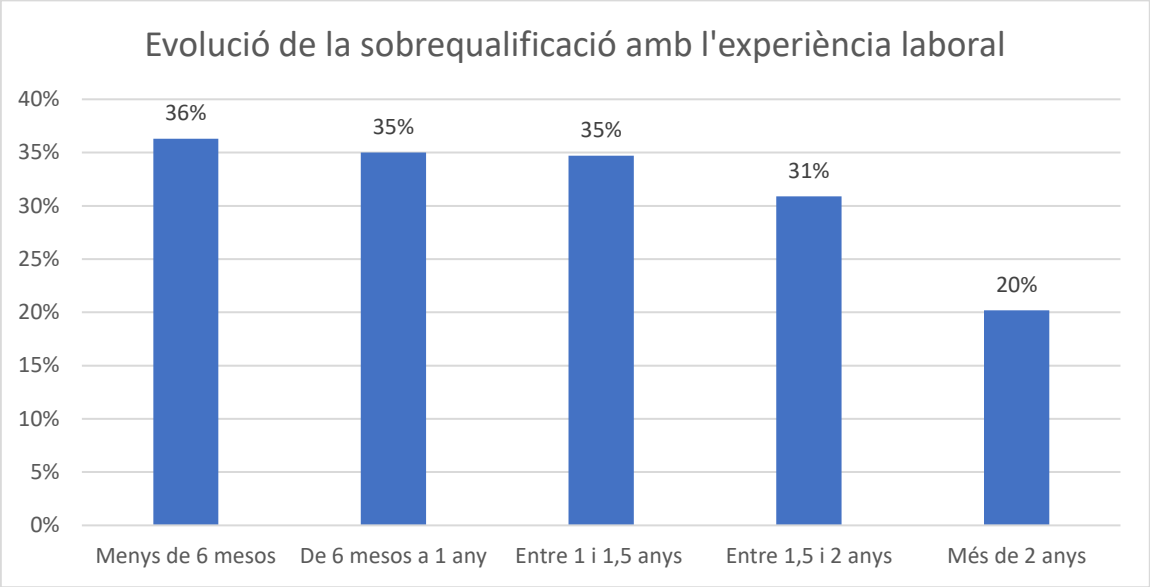


Figura 3.2: Evolució de la sobrequalificació amb l'experiència laboral. Elaboració pròpia. Font: INE.

- **Ajust educatiu i branca de titulació**

La sobrequalificació no afecta per igual a tots els àmbits educatius.

El Figura 3.3 es pot observar el percentatge de titulats, que segons l'àmbit d'estudis que han realitzat, declaren fer un treball pel qual és necessari els estudis universitaris. En primer lloc, hi ha Ciències de la Salut que té l'índex més gran d'ajust entre el que han estudiat a la universitat i el que apliquen en el seu lloc de treball, amb un 93% d'ocupacions ajustades als seus estudis, el que es tradueix en només 7% de sobrequalificació. D'altra banda, en Arts i Humanitats es concentra una taxa més elevada de sobrequalificació, ja que el 33% dels treballadors ocupen llocs que no requereixen el seu nivell de qualificació.

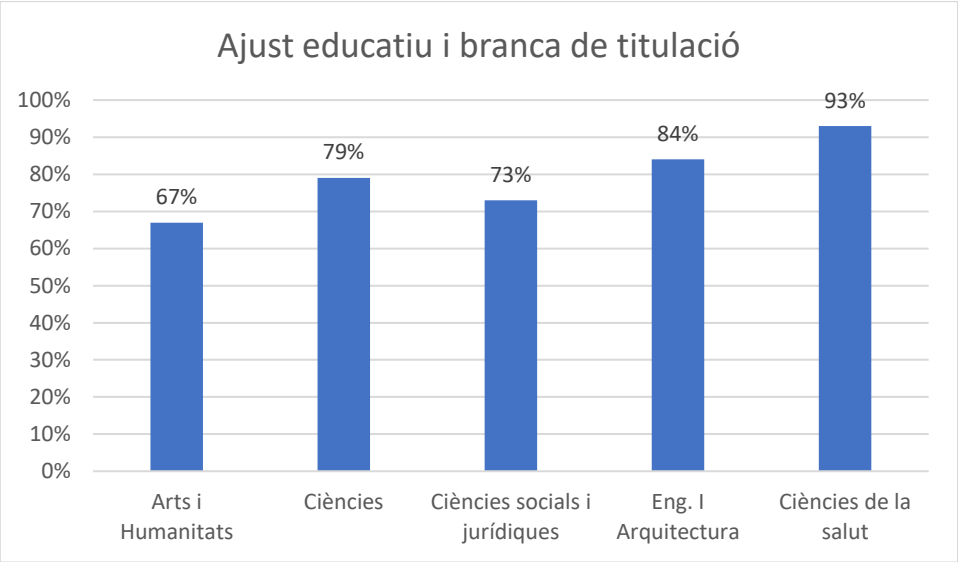


Figura 3.3: Ajust educatiu i branca de titulació. Elaboració pròpia. Font: INE.

- **Ajust educatiu i grandària de l'empresa**

Per altra part, la grandària de l'empresa també influeix. Les empreses petites i mitjanes són les que més treballadors sobrequalificats tenen.

En la Figura 3.4 es veu com les empreses de 250 o més treballadors, presenten un millor ajust educatiu, és a dir, els treballadors tenen llocs de treball més ajustats al seu nivell de qualificació. Es diu que només hi ha un 18% de sobrequalificació en empreses grans. Per altra banda, les empreses de menys de 10 treballadors, presenten un 32% de sobrequalificació. Aquest factor, podria estar relacionat en què en empreses grans, hi ha major productivitat i major salari.

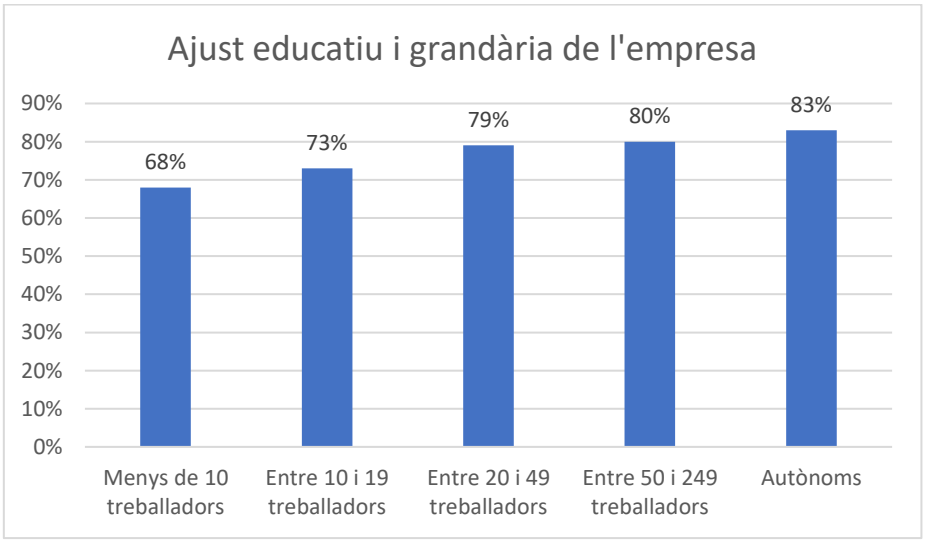


Figura 3.4: Ajust educatiu i grandària de l'empresa. Elaboració pròpia. Font: INE.

- **Ajust educatiu i salari mensual**

Com s'ha dit en el punt 3.1.1, la *Teoria del Capital Humà* prediu una relació positiva entre el capital humà i els salaris percebuts pels treballadors.

La Figura 3.5 compara els salaris de dos grups de treballadors, els de qualificació ajustada i els de sobrequalificació. S'observa com el 70% dels sobrequalificats cobren salaris inferiors als 1500 euros, mentre que el percentatge disminueix en el 30% per part dels treballadors que tenen una qualificació ajustada. Amb aquestes dades, es pot dir que hi ha una associació positiva entre la productivitat del treballador i l'adequació de la qualificació en el lloc de treball.

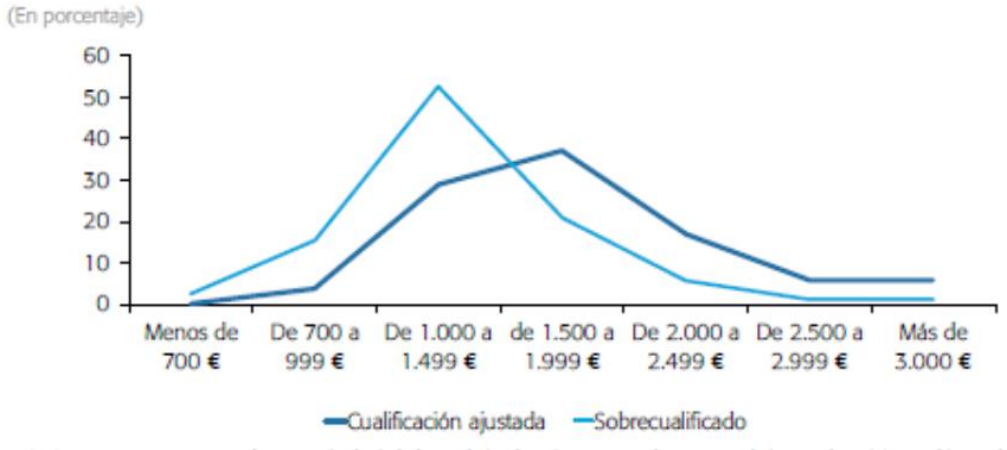


Figura 3.5: Distribució del salari brut mensual dels titulats amb sobrequalificació i sense. Font: INE.

En resum, a partir de l'enquesta realitzada per l'INE, es pot dir que:

- El 36% dels titulats amb una experiència laboral inferior a sis mesos estan sobrequalificats. Aquest percentatge es redueix al 20% entre els que tenen més de dos anys d'experiència laboral.
- Les titulacions amb menor sobrequalificació són les relacionades amb l'àmbit de Ciències de la Salut i les titulacions amb més sobrequalificació les incloses en la branca d'Arts i Humanitats.
- La sobrequalificació és més present en empreses petites. Trobem que mentre un de cada tres titulats a les petites empreses (menys de 10 treballadors) està sobrequalificat, aquest percentatge disminueix fins a un 20% a les empreses de 250 o més treballadors.
- Un 70% dels titulats sobrequalificats cobren salaris inferiors a 1.500 euros, mentre que aquest percentatge baixa al 30% entre els titulats amb una qualificació ajustada al lloc de treball.

3.1.3 Impacte de la sobrequalificació a Catalunya

En el cas de Catalunya, és important analitzar la taxa de sobrequalificació per a entendre el nivell de discrepància entre l'oferta i la demanda laboral en termes de formació i habilitats.

Catalunya és una de les regions més poblades i desenvolupades d'Espanya, amb una economia diversificada i un nivell educatiu alt en comparació amb altres regions del país. No obstant això, malgrat comptar amb un alt nombre de treballadors amb formació universitària i tècnica, Catalunya també presenta una taxa de sobrequalificació preocupant en alguns sectors laborals.

3.1.3.1 Enquesta Estudi de la Inserció Laboral de la població titulada de les universitats catalanes

Es realitza una enquesta sobre la inserció laboral de la població titulada el 2014 de les universitats catalanes. S'observarà si hi ha diferències entre els graduats segons l'experiència laboral posterior al primer treball, els àmbits d'estudis, la grandària de l'empresa i inclús el salari percebut.

Per calcular la taxa de sobrequalificació, es divideix el nombre de titulats sobrequalificats pel nombre total de titulats a Catalunya.

$$Taxa\ de\ sobrequalificació = \frac{Nombre\ de\ titulats\ sobrequalificats}{Nombre\ total\ de\ titulats}$$

Equació 3.1: Taxa de sobrequalificació.

En la Taula 3.1 s'afirma que el 29% dels joves catalans pateixen sobrequalificació en el seu lloc de treball. Això vol dir que 3 de cada 10 titulats tenen un nivell de formació i qualificació superior al necessari per al treball que exerceixen.

Com ja s'ha comentat, pot ser a causa de la falta d'oportunitats laborals adequades per als graduats universitaris o el fet que els estudiants triïn carreres amb una taxa d'ocupació més baixa que la demanda del mercat laboral.

Si es compara amb Espanya, com s'ha vist anteriorment, la taxa de sobrequalificació és del 38.8%, mentre que a Catalunya és del 29%. Això significa que el problema de la sobrequalificació és més agut en l'àmbit nacional en comparació amb Catalunya. És a dir, en mitjana, a Espanya més de tres de cada deu titulats tenen una formació superior a la

necessària per al treball que exerceixen o que estarà disponible en el mercat laboral, mentre que a Catalunya és menys de tres de cada deu titulats.

		Taxa de sobrequalificació
Nombre de titulats sobrequalificats	4503	**29.2%**
Nombre total de titulats	15421	

Taula 3.1: Classificació de titulats sobrequalificats i la taxa de sobrequalificació. Font: Elaboració pròpia.

És important tenir en compte que la sobrequalificació varia segons la regió i la indústria. Algunes regions i sectors poden tenir taxes de sobrequalificació més altes o més baixes que la mitjana nacional o regional. Per tant, és rellevant analitzar la sobrequalificació detalladament en cada context per a identificar les causes específiques i desenvolupar polítiques adequades per a abordar el problema.

- **Ajust educatiu i experiència laboral**

La sobrequalificació dels recentment graduats a Catalunya varia en funció de la seva experiència laboral, segons les dades disponibles.

La Figura 3.6 mostra l'evolució de la sobrequalificació dels estudiants de Catalunya. Es pot veure com el 31% dels titulats amb menys d'un any d'experiència laboral tenen sobrequalificació, i el percentatge es redueix a mesura que adquireixen més experiència laboral. Així i tot, la taxa de sobrequalificació és considerablement alta entre els titulats amb menys d'un any d'experiència laboral i aquells amb més de tres anys d'experiència laboral. Això pot ser degut al fet que molts titulats estan treballant en llocs de treball que no estan directament relacionats amb els seus estudis, la qual cosa els fa estar sobrequalificats per al seu lloc de treball. I pel que fa als titulats amb més de 3 anys d'experiència, pot ser degut al fet que, després d'un cert punt d'experiència, alguns treballadors poden trobar-se en posicions laborals que no estan alineades amb el seu nivell de qualificació.

En general, aquestes dades indiquen que la sobrequalificació continua sent un problema important a Catalunya, especialment per als graduats universitaris amb menys experiència laboral.

La taxa de sobrequalificació a Espanya és més alta que a Catalunya en tots els nivells d'experiència laboral. No obstant això, tots dos països presenten una tendència similar en la

qual la sobrequalificació disminueix a mesura que els titulats adquireixen més experiència laboral.

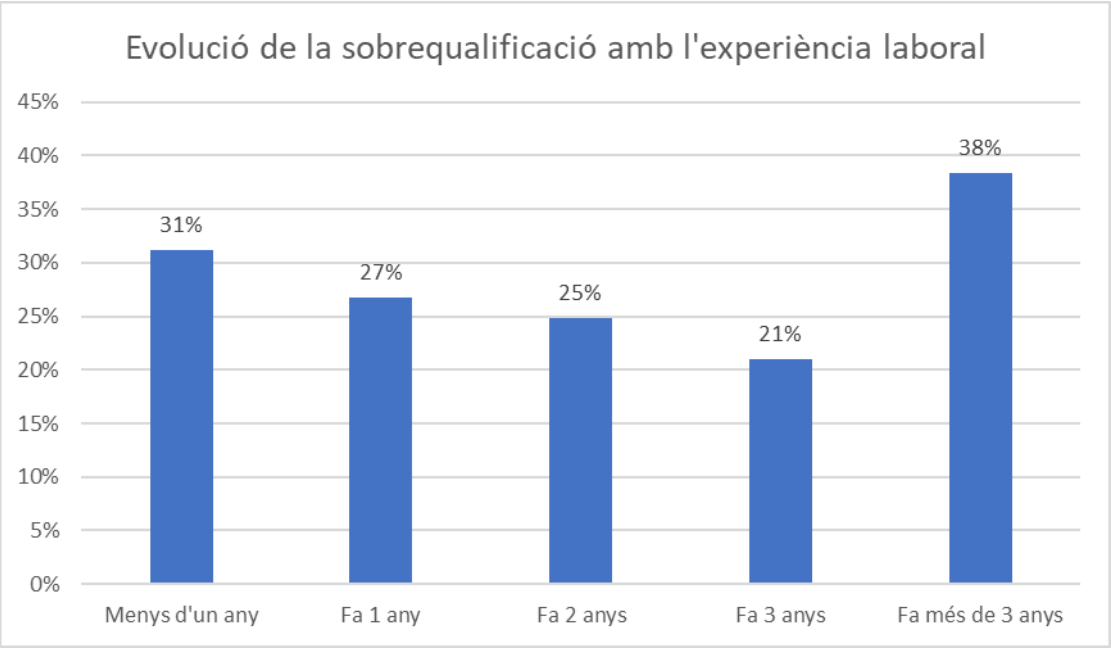


Figura 3.6: Evolució de la sobrequalificació amb l'experiència laboral. Font: Elaboració pròpia.

- **Ajust educatiu i branca de titulació**

Si es mira com afecta la sobrequalificació en els diferents àmbits, s'obté que en la Figura 3.7 l'ajust educatiu dels estudiants de Catalunya varia significativament segons la branca de titulació. Per exemple, en la branca d'Arts i Humanitats, el 50% dels titulats estan sobrequalificats per al seu treball, mentre que en la branca de Ciències aquest percentatge disminueix fins al 37%. En la branca de Ciències Socials i Jurídiques i Enginyeria i Arquitectura el percentatge de titulats sobrequalificats és del 28% i 27% respectivament. Finalment, en la branca de Ciències de la Salut, el percentatge de sobrequalificats és del 19%, la xifra més baixa entre les branques de titulació. Per tant, tenim que hi ha una major proporció de sobrequalificats en l'àrea d'Arts i Humanitats i una menor proporció en l'àrea de Ciències de la salut. Això pot ser degut a diversos factors, com ara la demanda del mercat laboral per a diferents àrees de coneixement i la disponibilitat de llocs de treball en aquests àmbits.

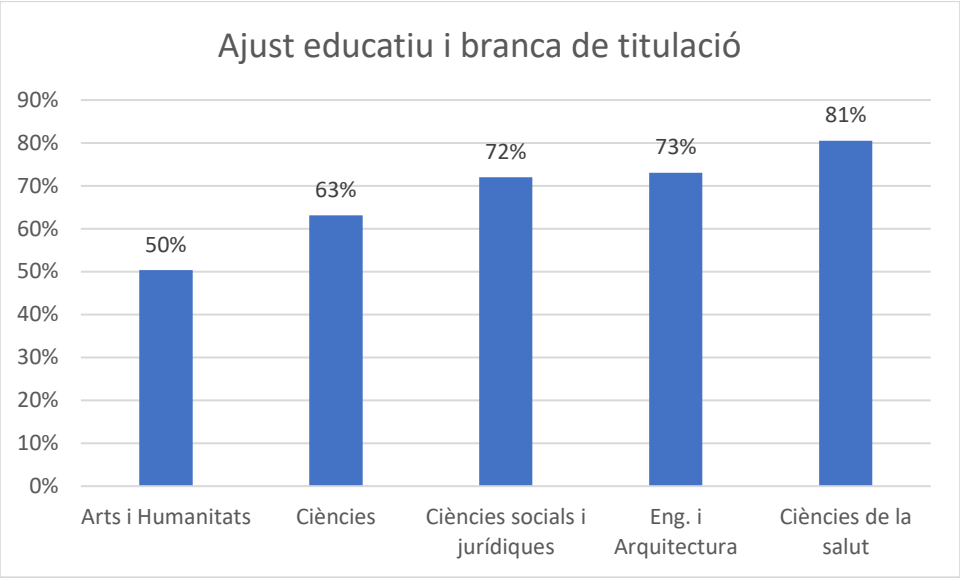


Figura 3.7: Ajust educatiu i branca de titulació. Font: Elaboració pròpia.

Com s'ha vist, en el cas d'Espanya, la branca amb més sobrequalificació és la de Ciències Socials i Jurídiques amb un 27%, seguida de la branca d'Arts i Humanitats amb un 33%, la branca de Ciències amb un 21%, la branca d'Enginyeria i Arquitectura amb un 16% i finalment la branca de Ciències de la Salut amb un 7%.

Per tant, s'observa que en general, els estudiants de Catalunya tenen una major sobrequalificació que els estudiants d'Espanya en totes les branques excepte en la branca de Ciències de la Salut, on els estudiants d'Espanya presenten una menor sobrequalificació.

- **Ajust educatiu i grandària de l'empresa**

Per altra banda, la grandària de l'empresa també influeix.

Segons les dades recollides, s'observa en la Figura 3.8 que l'ajust educatiu dels titulats de Catalunya varia segons la grandària de l'empresa on treballen. Així, els titulats que treballen en empreses amb 10 o menys treballadors presenten un ajust educatiu del 83%, mentre que els que treballen en empreses amb entre 11 i 50 treballadors presenten un ajust educatiu del 80%. A mesura que la grandària de l'empresa augmenta, l'ajust educatiu continua sent elevat, amb xifres com el 91% en empreses amb entre 51 i 100 treballadors i el 92% en empreses amb entre 101 i 250 treballadors. Cal destacar que en empreses de mida gran, amb més de 500 treballadors, i autònoms, l'ajust educatiu és més baix amb un 76% i 77% respectivament.

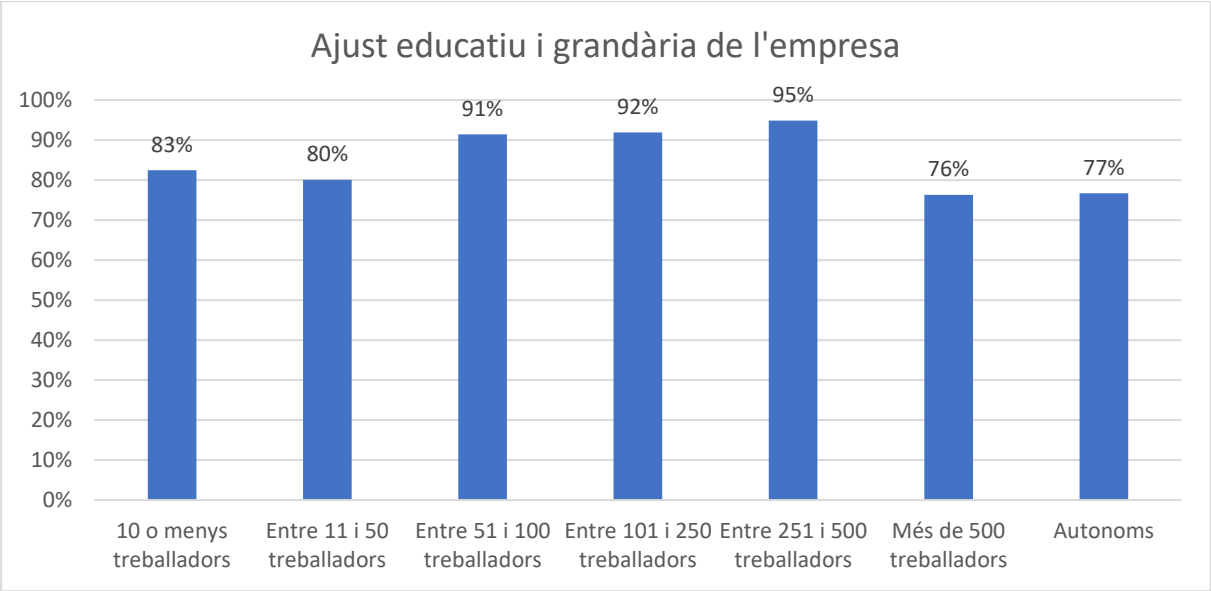


Figura 3.8: Ajust educatiu i grandària de l'empresa. Font: Elaboració pròpia.

En general, es pot observar que en la majoria dels casos Catalunya té un ajust educatiu més alt que la mitjana d'Espanya en totes les categories de grandària d'empresa, excepte en el cas d'empreses amb més de 500 treballadors, on l'ajust educatiu és més baix. Això pot indicar una major prioritat de l'educació i la formació a Catalunya en relació amb la grandària de l'empresa.

- **Ajust educatiu i salari mensual**

Finalment, si es mira els salaris percebuts pels treballadors amb sobrequalificació i sense, en la Figura 3.9 s'observa que hi ha una diferència significativa en la distribució del salari entre els dos grups. Els sobrequalificats tenen una major proporció de salaris baixos, amb un 61% dels titulats que guanyen menys de 1500 € al mes, en comparació amb el 43% dels titulats amb qualificació ajustada. A més a més, els titulats amb qualificació ajustada tenen una major proporció de salaris alts, amb un 18% dels treballadors que guanyen més de 2500 € al mes, en comparació amb l'11% dels sobrequalificats.

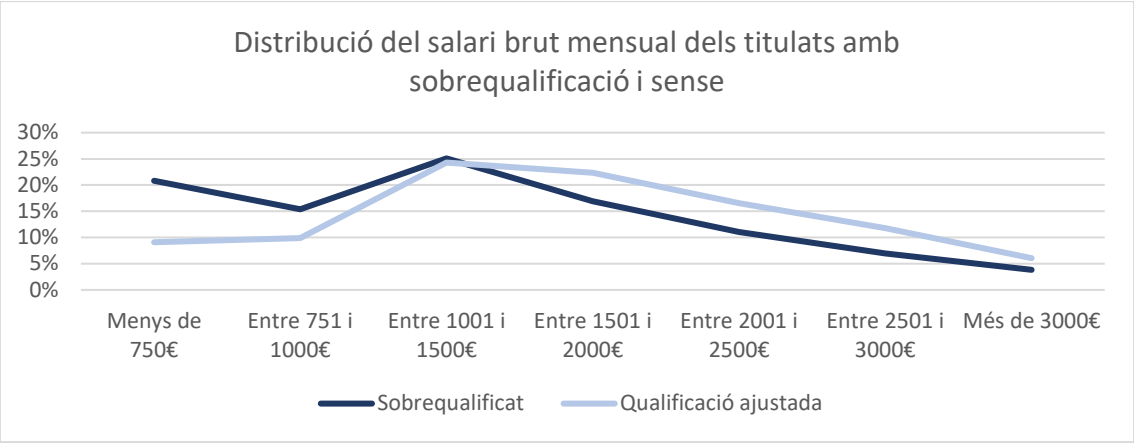


Figura 3.9: Distribució del salari brut mensual dels titulats amb sobrequalificació i sense. Font: Elaboració pròpia.

En Espanya també s'ha vist que els treballadors sobrequalificats tenen salaris més baixos en comparació els treballadors amb qualificació ajustada. Tot i això, el percentatge de titulats a Catalunya amb sobrequalificació que cobren menys de 1500 € al mes és inferior a la dels titulats a Espanya. A diferència dels titulats amb qualificació ajustada que cobren menys de 1500 € al mes, és més elevada a Espanya. Podem dir que els titulats a Catalunya amb sobrequalificació tenen una distribució de salari més elevada que els titulats amb sobrequalificació a Espanya, especialment en les categories de salari més altes (més de 1500 €).

En resum, de les dades extretes a partir de l'enquesta, afirmen que:

- El 31% dels titulats amb una experiència laboral inferior a un any estan sobrequalificats. Es redueix al 21% dels titulats aquells que l'experiència laboral és de tres anys.
- Les titulacions de les branques d'Arts i Humanitats tenen una sobrequalificació més elevada, a diferència de Ciències de la Salut que és la branca que menys sobrequalificació pateixen els estudiants.
- La sobrequalificació és més present en empreses de més de 500 treballadors, en canvi, els titulats que treballen en empreses mitjanes presenten l'ajust educatiu més gran.
- Un 61% dels titulats sobrequalificats cobren salaris inferiors als 1500 euros, en comparació amb el 43% dels titulats amb qualificació ajustada al lloc de treball.

3.2 Machine Learning

El *Machine Learning* – aprenentatge automàtic – és una branca de la intel·ligència artificial i la informàtica que permet que les màquines aprenguin sense ser expressament programades per fer-ho. Una habilitat indispensable per fer sistemes capaços d'identificar patrons entre les dades per fer prediccions. Aquesta tecnologia és present en infinites aplicacions com les recomanacions de Netflix o Spotify, les respostes intel·ligents de Gmail o la parla de Siri i Alexa (BBVA, 2019).

El terme es va utilitzar per primera vegada el 1959. Tot i això, ha guanyat rellevància en els últims anys a causa de l'augment de la capacitat de computació i del boom de les dades. De fet, les tècniques d'aprenentatge automàtic són una part fonamental del Big Data (IBERDROLA).

La intel·ligència artificial, per la seva banda, és un camp més ampli que busca desenvolupar sistemes que puguin executar tasques que normalment requereixen intel·ligència humana, com ara el raonament, l'aprenentatge i la resolució de problemes. Dins de la intel·ligència artificial, el *Machine Learning* és una de les tècniques més usades per assolir aquests objectius.

En resum, la relació entre *Machine Learning* i intel·ligència artificial és que el primer és una tècnica específica emprada en el camp més ampli de la intel·ligència artificial per ensenyar les màquines a aprendre i millorar el seu exercici en tasques específiques.

3.2.1 Àmbits d'aplicació del *Machine Learning*

Moltes activitats ja s'estan aprofitant actualment del *Machine Learning*. Sectors com el de les compres *Online*, per exemple com es decideix instantàniament els productes recomanats per a cada client al final d'un procés de compra, l'*Online advertising*, on posar un anunci perquè tingui més visibilitat en funció de l'usuari que visita la web, i molts més altres.

El camp d'aplicació pràctica depèn de la imaginació i de les dades que estiguin disponibles a l'empresa. Aquests són alguns exemples (¿Qué es Machine Learning y qué aplicaciones tiene?, 2018):

- Detectar frau en transaccions.
- Predir de fallades en equips tecnològics.
- Preveure quins empleats seran més rendibles l'any que ve (el sector dels Recursos Humans està apostant seriosament pel *Machine Learning*).

- Seleccionar clients potencials basant-se en comportaments a les xarxes socials, interaccions a la web...
- Predir el trànsit urbà.
- Saber quin és el millor moment per publicar piulades, actualitzacions de *Facebook* o enviar les *newsletter*.
- Fer prediagnòstics mèdics basats en símptomes del pacient.
- Canviar el comportament d'una app mòbil per adaptar-se als costums i les necessitats de cada usuari.
- Detectar intrusions a una xarxa de comunicacions de dades.
- Decidir quina és la millor hora per trucar a un client.

3.2.2 Tipus de *Machine Learning*

Depenent de les dades disponibles i la tasca que es vulgui abordar, es pot triar entre diferents tipus d'aprenentatge.

3.2.2.1 Aprenentatge supervisat

L'aprenentatge supervisat és un tipus d'entrenament en el qual el model rep dades etiquetades, és a dir, informació que indica què ha de predir. Per exemple, si es té una gelateria que ha estat registrant dades sobre la temperatura, el clima i el nombre de gelats venuts diàriament, es vol entrenar un model que pugui predir el nombre de gelats venuts en un dia determinat en funció de les dades climàtiques. Un altre exemple seria l'ús de fotografies etiquetades per entrenar un model que pugui identificar objectes similars a altres bases de dades (BBVA, 2019).

3.2.2.2 Aprenentatge no supervisat

Per altra banda, l'aprenentatge no supervisat treballa sense informació que indica que s'ha de predir. Aquests algoritmes s'utilitzen principalment en feines on és necessari analitzar les dades per extreure nous coneixements o agrupar entitats per afinitat. És a dir, no estan programats per detectar un tipus específic de dades, sinó que busquen exemples que s'assemblin i es puguin agrupar (Sanchez, 2020).

3.2.2.3 Aprenentatge semisupervisat

Com el seu nom indica, aquest mètode combina l'aprenentatge supervisat i el no supervisat. Aquesta tècnica es basa en l'ús d'una petita quantitat de dades etiquetades (ja han estat processades i identificades amb una categoria especifica que les defineix) i d'una gran sense

etiquetar, és a dir, sense identificar. L'avantatge d'aquest mètode és que no necessita grans quantitats de dades etiquetades. Es fa ús quan es treballa amb dades com a documents llargs que els humans trigarien molt a llegir i etiquetar (Sanchez, 2020).

3.2.2.4 Aprenentatge per esforç

Per acabar, l'aprenentatge per reforç és una tècnica d'aprenentatge automàtic que consisteix a recompensar els comportaments desitjats i penalitzar els no desitjats, a través de prova i error per aconseguir una solució òptima i una recompensa general màxima. Aquesta tècnica s'aplica en diversos àmbits, com ara els jocs, la robòtica, l'optimització de recursos o els sistemes de control. Aquesta tècnica permet als sistemes aprendre a partir de la seva pròpia experiència i a modificar la seva conducta per a prendre les millors decisions davant diferents situacions (BBVA, 2019).

3.2.3 Classificació en *Machine Learning*

La classificació en l'aprenentatge automàtic (*Machine Learning*) és una tècnica que consisteix a entrenar un algorisme perquè pugui classificar noves dades en una o diverses categories o classes prèviament definides. En altres paraules, la classificació és un procés d'etiquetatge automàtic de dades.

La classificació és una tasca important en el *Machine Learning*, ja que s'utilitza en una àmplia varietat d'aplicacions, com l'anàlisi de sentiments en xarxes socials, la detecció de *spam* en el correu electrònic, la identificació d'objectes en imatges, la segmentació de clients en el màrqueting, entre altres. A més, la classificació és una tasca fonamental en la intel·ligència artificial en general, ja que permet als sistemes automatitzats prendre decisions basades en les dades que reben.

Hi ha diversos tipus d'algoritmes de classificació que es fan servir per a predir la classe o categoria a la qual pertany un objecte. En els següents apartats es veurà diferents algoritmes aplicats en el cas de la sobrequalificació.

4. BASE DE DADES I EINES PER A L'ANÀLISI

Una vegada establerts els fonaments teòrics sobre el tema de sobrequalificació i aprenentatge automàtic, i els objectius del projecte, es va procedir a descriure la base de dades, els algoritmes que s'implementaran, les mètriques que s'utilitzaran per a aconseguir aquests objectius. En aquesta secció es fa esment a la base de dades emprada, els algoritmes i les eines computacionals emprades per a l'anàlisi de dades.

Per a obtenir classificacions i prediccions, tant de categories discretes com de valors continus, es va treballar amb un conjunt de dades. En primer lloc, es va dur a terme una cerca per a trobar una base adequada per acomplir el projecte. Posteriorment, es va realitzar un procés de tractament de les dades, seguit de l'aplicació d'algorismes amb els seus respectius mètodes i anàlisis.

4.1 Base de dades

L'obtenció i estructura de dades és un pas fonamental en qualsevol projecte de mineria de dades. Abans de poder aplicar algorismes d'aprenentatge automàtic per a extreure informació útil, és necessari tenir una bona comprensió de les dades que s'estan usant i com estan estructurats. En aquesta etapa, s'han de considerar aspectes com la qualitat de les dades, la quantitat, el format i la font d'aquests. A més, és important tenir en compte que l'obtenció i estructuració adequada de les dades pot marcar la diferència entre l'èxit i el fracàs d'un projecte de mineria de dades. En aquesta fase, es poden emprar diverses eines i tècniques per a recol·lectar, preprocessar i netejar les dades per a la seva posterior anàlisi.

Aquest projecte es basarà en la sobrequalificació dels estudiants recentment graduats de Catalunya en l'any 2017. La base de dades tracta sobre la inserció laboral dels estudiants graduats l'any 2014 a Catalunya. Conté dades sobre situació laboral actual, el nivell d'estudis requerit, les funcions, el tipus de contracte, els guanys anuals, la satisfacció, la formació adquirida, entre d'altres. La base de dades és extreta de l'enquesta realitzada sobre la inserció laboral de la població titulada el 2014 de les universitats catalanes. Té una mida mostral de 15563 registres i 63 variables.

A continuació, es presenta les variables més destacades i que s'utilitzaran en l'anàlisi posterior:

NOM VARIABLE	DESCRIPCIÓ	UNITATS DE MESURA
satisf1_10	Satisfacció amb el contingut de la feina	Escala ordinal (1-10)
satisf2_10	Satisfacció amb les perspectives de millora	Escala ordinal (1-10)
satisf3_10	Satisfacció amb el nivell de retribució	Escala ordinal (1-10)
satisf4_10	Satisfacció amb la utilitat dels coneixements	Escala ordinal (1-10)
satisf5_10	Satisfacció general amb la feina on treballes	Escala ordinal (1-10)
nteor_10	Nivell de formació teòrica	Escala ordinal (1-10)
npra_10	Nivell de formació pràctica	Escala ordinal (1-10)
ncoral_10	Nivell d'expressió oral	Escala ordinal (1-10)
ncescr_10	Nivell d'expressió escrita	Escala ordinal (1-10)
nequip_10	Nivell de treball en equip	Escala ordinal (1-10)
nlider_10	Nivell de lideratge	Escala ordinal (1-10)
nsolprob_10	Nivell de solució de problemes	Escala ordinal (1-10)
npresdec_10	Nivell de presa de decisions	Escala ordinal (1-10)
ncreat_10	Nivell de creativitat	Escala ordinal (1-10)
nptecrit_10	Nivell de pensament crític	Escala ordinal (1-10)
ngestio_10	Nivell de gestió	Escala ordinal (1-10)
ninform_10	Nivell d'informàtica	Escala ordinal (1-10)
nidiom_10	Nivell d'idiomes	Escala ordinal (1-10)
ndoc_10	Nivell d'habilitats de documentació	Escala ordinal (1-10)
ateor_10	Utilitat de formació teòrica	Escala ordinal (1-10)
apra_10	Utilitat de formació pràctica	Escala ordinal (1-10)
acoral_10	Utilitat d'expressió oral	Escala ordinal (1-10)
acescr_10	Utilitat d'expressió escrita	Escala ordinal (1-10)
aequip_10	Utilitat de treball en equip	Escala ordinal (1-10)
alider_10	Utilitat de lideratge	Escala ordinal (1-10)
asolprob_10	Utilitat de solució de problemes	Escala ordinal (1-10)
apresdec_10	Utilitat de presa de decisions	Escala ordinal (1-10)
acreat_10	Utilitat de creativitat	Escala ordinal (1-10)
aptecrit_10	Utilitat de pensament crític	Escala ordinal (1-10)
agestio_10	Utilitat de gestió	Escala ordinal (1-10)
ainform_10	Utilitat d'informàtica	Escala ordinal (1-10)
aidiom_10	Utilitat d'idiomes	Escala ordinal (1-10)
adoc_10	Utilitat d'habilitats de documentació	Escala ordinal (1-10)

sitact	Situació laboral actual	1 = Treballo 2 = No treballo però he treballat després dels estudis 3 = No he treballat mai després dels estudis
anyini_c	Temps des de l'inici de la feina actual	1 = Fa més de 3 anys 2 = Fa 3 anys 3 = Fa 2 anys 4 = Fa 1 any 5 = Menys d'un any
autonom_c	Tipus de contracte: autònom	0 = No 1 = Sí
codi_a	Branca Titulació	1 = Arts i Humanitats 2 = Ciències socials i jurídiques 3 = Ciències 4 = Ciències de la salut 5 = Eng. i Arquitectura
guanys	Guanys anuals bruts	-1 = Ns/Nc 1 = Menys de 9.000 € 2 = Entre 9.000 i 12.000 € 3 = Entre 12.001 i 15.000 € 4 = Entre 15.001 i 18.000 € 5 = Entre 18.001 i 24.000 € 6 = Entre 24.001 i 30.000 7 = Entre 30.001 i 40.000 € 8 = Entre 40.001 i 50.000 € 9 = Més de 50.000 €
numtreb	Nombre de treballadors	-1 = Ns/Nc 1 = 10 o menys 2 = Entre 11 i 50 3 = Entre 51 i 100 4 = Entre 101 i 250 5 = Entre 251 i 500 6 = Més de 500
sobreq	L'enquestat està sobrequalificat o no	0 = Normal 1 = Sobrequalificat

Taula 4.1: Descripció variables. Font: Elaboració pròpia.

Amb aquesta base de dades s'ha dut a terme un estudi de classificació per a determinar si els recentment graduats estan sobrequalificats o no. De manera que, s'ha buscat assignar a cada observació una categoria discreta, 1 si està sobrequalificat i 0 si no està sobrequalificat.

4.2 Algorismes d'aprenentatge automàtic per a la classificació i la predicció

L'anàlisi de dades i la classificació de patrons són aspectes clau en el camp de la ciència de les dades i l'aprenentatge automàtic. Els algoritmes de classificació són eines potents que permeten categoritzar i etiquetar les dades en diferents classes o categories en funció de les seves característiques.

L'objectiu principal dels algoritmes de classificació és aprendre un model a partir de dades d'entrenament per a posteriorment poder predir o classificar noves dades desconegudes. Això es fa mitjançant la identificació de patrons, tendències o relacions en les dades d'entrenament i la seva aplicació per realitzar prediccions en noves dades.

Els algoritmes de classificació utilitzen diferents tècniques i models, com ara arbres de decisió, màquines de vectors de suport (SVM), models bayesians, regressió logística, entre d'altres. Cada algoritme té els seus avantatges i desavantatges i pot ser més adequat per a certes situacions o tipus de dades.

En aquest treball, es farà una anàlisi i aplicació dels algoritmes de classificació per a categoritzar les dades i predir la pertinença a una determinada classe o categoria. A través de l'aprenentatge automàtic supervisat, es construiran models fent ús de dades d'entrenament i de prova, i s'avaluarà el seu rendiment mitjançant mètriques com l'exactitud, la precisió, la sensibilitat (Recall), l'especificitat i la corba ROC.

Mitjançant aquests algoritmes, es busca obtenir una comprensió més profunda de les dades i proporcionar eines per a la presa de decisions basada en la classificació. A més a més, a partir del model de classificació seleccionat, s'aconseguirà un perfil de treballador que sigui més proper a ser sobrequalificat.

4.2.1 Arbres de decisió

Els arbres de decisió o classificació són algorismes utilitzats en l'àmbit de l'aprenentatge automàtic i la intel·ligència artificial per a resoldre problemes de classificació. A diferència dels models estadístics tradicionals, els arbres de decisió no es basen en l'estimació de paràmetres, sinó en la creació de diagrames lògics que representen les decisions preses pel model.

Aquests algorismes són especialment útils quan es treballa amb grans volums de dades, ja que permeten identificar patrons i estructures complexes que no són detectables mitjançant mètodes estadístics convencionals.

Els arbres de decisió són de caràcter descriptiu, és a dir, proporcionen una explicació clara i comprensible de les decisions preses pel model. Això facilita la comprensió i la interpretació de les regles que defineixen les classificacions de dades.

El procés de construcció d'un arbre de decisió implica la partició successiva de les dades en grups homogenis mitjançant combinacions de variables independents. Aquesta tècnica, anomenada segmentació jeràrquica, es basa en iteracions de dalt a baix que formen grups específicament definits segons la variable dependent.

Durant la construcció de l'arbre de decisió s'utilitza una mostra d'entrenament que conté informació sobre el grup al qual pertany cada cas. L'objectiu és establir un criteri de classificació a partir d'aquesta mostra.

El procés comença amb un node inicial on s'escull una variable independent per realitzar la primera partició. Es busca una divisió de la variable per generar dos conjunts de dades que siguin tan homogenis com sigui possible respecte a la variable dependent. Això s'aconsegueix escollint un tall de punt a la variable. Per exemple, si la variable escollida és "x1", es determina un punt de tall "c" per dividir les dades en dos grups: "x1 ≤ c" i "x1 > c".

A partir del node inicial, es generen dos nous nodes, un al qual arriben les observacions de "x1 ≤ c" i l'altre al qual arriben les observacions de "x1 > c". El procés de selecció de variables i punts de tall es repeteix a cadascun d'aquests nodes generats. Pretén dividir els conjunts de dades en cada node de manera que es minimitzi la impuresa del grup (Parra, 2019).

Aquest procés de selecció i divisió de variables es repeteix contínuament a cada node fins que totes les observacions s'han classificat correctament en els seus grups respectius. En cada etapa, se seleccionen la variable i el punt de tall que millor contribueixen a la separació i classificació precisa de les dades.

Per implementar l'algoritme a R, hi ha diverses maneres d'obtenir arbres de decisió. Els algoritmes més freqüents són: CHAID, CHAID Exhaustiu, CART i QUEST. En aquest projecte, s'utilitzarà la que es coneix com a CART: *Classification And Regression Trees*. Els altres algorismes esmentats, com CHAID, CHAID Exhaustiu i QUEST, solen ser implementats en eines com SPSS.

CART és una tècnica d'aprenentatge supervisat de la que es pot aconseguir arbres de classificació, quan la variable resposta és discreta, i arbres de regressió, quan la variable objectiu és contínua.

Com que la variable resposta és la sobrequalificació, i és una variable discreta, en aquest cas es realitzarà classificació. La implementació particular que s'usarà de CART és coneguda com a *Recursive Partitioning and Regression Trees* o RPART. Utilitzant aquesta funció, existeixen dos algoritmes per dividir els nodes d'un arbre, *Information Gain* i *Gini Index*. Tots dos funcionen de manera diferent, però el propòsit és el mateix, trobar la millor manera de dividir les dades.

L'algoritme *Information Gain* funciona amb la fórmula següent:

$$-\big(p \cdot log2(p)\big) - (q \cdot log2(q))$$

Equació 4.1: Fórmula algoritme *Information Gain*.

On p és la probabilitat d'èxit o de casos positius. És a dir els casos de sobrequalificació. I q és la probabilitat de fracàs o de casos negatius, en aquest cas, els casos de normal. Si el resultat és 1, vol dir que les classes estan dividides (50%, 50%) i si el resultat fos 0, això significa que només existeixen dades d'una sola classe, el que l'arbre farà és dividir les dades on el resultat de la fórmula sempre sigui el més pròxim a 0 possible.

L'algoritme *Gini Index* utilitza la següent fórmula:

$$p^2 + q^2$$

Equació 4.2: Fórmula algoritme *Gini Index*.

En aquest cas, també es busca un número el més a prop possible a zero, ja que indica que les dades es divideixen de manera més homogènia.

4.2.2 Naive Bayes

El classificador *Naive Bayes* és un algorisme d'aprenentatge automàtic supervisat utilitzat per a classificar dades en categories. Es basa en el *Teorema de Bayes*, que calcula la probabilitat d'una hipòtesi donada l'evidència i el coneixement previ. Aquest classificador assumeix la independència entre totes les variables predictores, la qual cosa pot no ser sempre cert en el món real. Malgrat aquesta simplificació, el classificador NB és àmpliament usat a causa de la seva simplicitat, eficiència i bon rendiment en moltes aplicacions del món real (Ray, 2017).

Utilitza la teoria de la probabilitat condicional per a construir models que prediuen la probabilitat de possibles resultats, la qual cosa ho fa especialment útil en problemes de classificació i predicció (Gonzalez, Naive Bayes – Teoría, 2019):

$$P(A \,/\, B) = \frac{P(B \,/\, A)\, P(A)}{P(B)}$$

Equació 4.3: Probabilitat condicionada.

$P(A \,/\, B)$ és la probabilitat posterior de la *classe* (A, objectiu) donat el *predictor* (B, atributs).
$P(A)$ és la probabilitat prèvia de *classe*.
$P(B \,/\, A)$ és la probabilitat del *predictor* donat la *classe*.
$P(B)$ és la probabilitat prèvia del *predictor*.

A partir del teorema de *Bayes*, es prendrà aquella classe de la qual la probabilitat a posteriori és major, és a dir, aquella classe A de la qual $P(A|B)$ sigui màxima (Departamento de Matemática Aplicada, 2021):

$$classe(B) = \arg \max_{A \in \{1,\dots,m\}} P(A|B) = \arg \max_{A \in \{1,\dots,m\}} \frac{P(B|A)P(A)}{P(B)}$$

I com que $P(B)$ no depèn de A serà suficient amb prendre el valor màxim de:

$$classe(B) = \arg \max_{A \in \{1,\dots,m\}} P(B|A)\, P(A)$$

Assumint que els atributs observats en qualsevol instància són independents una vegada sabem que la instància pertany a la classe A. La probabilitat de $P(B|A)$ és:

$$P(B|A) = P(b_1|A)P(b_2|A) \dots P(b_n|A) = \prod_{i=1}^{n} P(b_i|A)$$

4.2.3 Support Vector Machines

L'algorisme de *Support Vector Machine* (SVM) és àmpliament utilitzat en l'aprenentatge automàtic per a problemes de classificació i, en menor mesura, per a problemes de regressió.

El seu objectiu principal és trobar un hiperplà en un espai d'alta dimensió que permeti classificar de manera clara els punts de dades. La dimensió de l'hiperplà depèn del nombre de característiques presents en les dades (Support Vector Machine Algorithm, 2021).

El model SVM considera les dades d'entrada com a vectors en un espai N-dimensional, on cada vector representa un punt de dades. Donat un conjunt de punts de dades etiquetades en dues possibles categories, l'algorisme SVM construeix un model capaç de predir a quina categoria pertany un nou punt de dades desconegut.

La SVM cerca trobar un hiperplà òptim que maximitzi la separació entre les diferents classes de punts de dades. Aquest enfocament es coneix com a "marge màxim". El model SVM cerca trobar un hiperplà que tingui la distància més gran possible (marge) amb els punts que estan més pròxims a ell. Per aquesta raó, sovint es denomina als SVM com a classificadors de marge màxim. Els punts de dades etiquetades amb una categoria es troben a un costat de l'hiperplà, mentre que els punts de l'altra categoria es troben a l'altre costat (Support Vector Machine (SVM) Algorithm - Javatpoint).

Per exemple, la Figura 4.1 mostra els possibles hiperplans que maximitzen la separació entre les diferents classes de punts de dades.

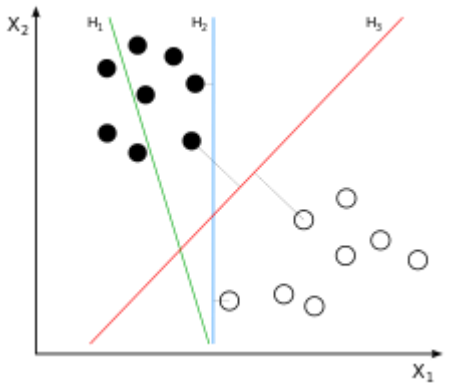


Figura 4.1: Hiperplans per classificar les classes. Font: Wikipedia.

L'hiperplà H_1 no separa les classes. L'hiperplà H_2 si separa les classes però amb un marge petit. En canvi, l'hiperplà H_3 separa les classes amb el màxim marge.

4.2.4 Regressió Logística

La Regressió Logística és un algorisme de classificació que s'utilitza per a predir la probabilitat d'una variable dependent categòrica. En aquest mètode, la variable dependent és binària i es

codifica com 1-0, sí-no, obert-tancat, etc. (Gonzalez, Curvas ROC y Área bajo la curva (AUC), 2019).

La Regressió Logística és un dels algorismes de *Machine Learning* més simples i comuns per a la classificació de dues classes. És fàcil d'implementar i es pot fer servir com a referència inicial per a qualsevol problema de classificació binària. Descriu i estima la relació entre una variable binària dependent i les variables independents.

Aquest model logístic binari s'usa per a estimar la probabilitat d'una resposta binària basada en una o més variables predictores o independents. Permet determinar que la presència d'un factor de risc augmenta la probabilitat d'un resultat específic en un percentatge determinat (Gonzalez, Curvas ROC y Área bajo la curva (AUC), 2019).

En general, aquest algorisme es pot emprar en diversos problemes de classificació, com la detecció de *spam*, la predicció de la diabetis, determinar si un client comprarà un producte en particular o si s'anirà a la competència, entre molts altres exemples on es pot aplicar aquest algorisme.

El model Logit es defineix a partir de la següent funció de distribució:

$$P(Y=1) = \frac{1}{1+e^{-Z_i}} = \frac{e^{Z_i}}{1+e^{Z_i}}$$

Equació 4.4: Funció de distribució.

On $Z_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$, $X_1, X_2, \ldots, X_k$ conjunt de variables independents o explicatives i $\beta_1, \beta_2, \ldots, \beta_k$ paràmetres del model a estimar. S'estima un model on l'estimació s'interpreta com la probabilitat de que succeeixi el grup codificat com 1, és a dir P(Y=1).

La linearització de la funció de distribució (Equació 4.4), es realitza mitjançant la definició de la Logit que denotem per L_i, prenent el logaritme natural de la raó entre les probabilitats P_i i la probabilitats complementàries $(1 - P_i)$.

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Equació 4.5: Logaritme natural de la raó.

En aquest cas, es farà servir aquest algoritme per classificar els estudiants en funció de si tenen sobrequalificació $(Y = 1)$ o no.

4.3 Eines per l'avaluació de la classificació

A més de la base de dades i els algorismes implementats, també s'han emprat altres eines per a l'anàlisi de classificació, que han estat útils per a estudiar els algoritmes i treure conclusions sobre elles.

Quan es fa servir un algorisme de classificació sobre un conjunt de dades, el que fem és mesurar com és de precís, mirant si la classe assignada a cada element classificat coincideix amb la real. Es pot mesurar la precisió de manera senzilla mirant, per exemple, el percentatge d'elements ben classificats respecte al total.

Més formalment, si es té un conjunt de N dades amb K>1 classes diferents, i que es té N1, …, NK elements de cada classe, i es construeix un classificador, aquest ens retornarà una classe per cada element, que podrà coincidir o no amb l'original. Això es pot representar amb el que es coneix com a matriu de confusió.

La matriu de confusió és una taula que mostra l'acompliment del classificador en comparar les classes predites amb les classes reals. En la matriu de confusió, es defineixen quatre valors:

- Veritables positius (TP): el nombre de casos en els quals el classificador va predir correctament una classe com a positiva.
- Veritables negatius (TN): el nombre de casos en els quals el classificador va predir correctament una classe com a negativa.
- Falsos positius (FP): el nombre de casos en els quals el classificador va predir incorrectament una classe com a positiva.
- Falsos negatius (FN): el nombre de casos en els quals el classificador va predir incorrectament una classe com a negativa.

Figura 4.2: Matriu de confusió. Font: *(Barrios Arce, 2019)*

A partir d'aquestes 4 opcions sorgeixen les mètriques de la matriu de confusió, d'una banda, l'exactitud i la precisió i per una altra la sensibilitat i l'especificitat.

L'exactitud o *accuracy*, mesura la proporció de prediccions correctes realitzades pel model respecte al nombre total de prediccions. És a dir, és un indicador de com de bé el model classifica correctament les instàncies en el conjunt de dades d'entrenament.

$$\frac{VP + VN}{VP + FP + FN + VN}$$

Equació 4.6: Fórmula Exactitud.

La precisió es representa per la proporció de casos classificats correctament en relació amb tots els casos positius identificats.

$$\frac{VP}{VP + FP}$$

Equació 4.7: Fórmula Precisió.

La sensibilitat (o recall), representa la proporció de casos positius reals que es classifiquen correctament com a positius.

$$\frac{VP}{VP + FN}$$

Equació 4.8: Fórmula Sensibilitat

En el cas de l'especificitat tracta dels casos negatius que l'algorisme ha classificat correctament. Expressa quan de bé pot el model detectar aquesta classe.

$$\frac{VN}{VN + FP}$$

Equació 4.9: Fórmula Especificitat.

Per altra banda, una altra mètrica utilitzada per avaluar els models de classificació és l'àrea sota la corba de ROC, mesura el rendiment del model de classificació binària. La corba ROC representa la relació entre la taxa de verdaders positius (TPR, True Positive Rate) i la taxa de falsos positius (FPR, False Positive Rate) a diferents punts de tall.

- True Positive Rate $TPR = Sensibilitat = \frac{VP}{VP+FN}$
- False Positive Rate $FPR = 1 - TPR = 1 - \frac{VP}{VP+FN}$

5. TRACTAMENT PREVI DE LES DADES

En els problemes de *Machine Learning* és molt important tractar prèviament les dades i realitzar totes aquelles transformacions necessàries, per a aconseguir una estructura del conjunt de dades que estigui adaptada a l'enfocament del problema i als algorismes implementats. A continuació, s'exposen els diferents processos realitzats al conjunt de dades, previs a l'aplicació dels algorismes.

5.1 Càlcul de la sobrequalificació

Primer de tot a partir de la base original s'ha calculat la variable a estudiar, és a dir la variable de sobrequalificació. La variable sobrequalificació es calcula per a avaluar si un individu està sobrequalificat pel lloc de treball en funció de les funcions requerides i el nivell de titulació. Per a fer aquest càlcul, s'ha utilitzat les variables "funprop1" i "funcions_c".

La variable "funprop1" indica si les funcions requerides són pròpies del nivell de titulació exigit. Els valors possibles són els següents:

- "-2": No aplica.
- "0": Les funcions requerides no són pròpies del nivell de titulació requerit.
- "1": Les funcions requerides són pròpies del nivell de titulació requerit.

D'altra banda, la variable "funcions_c" representa les funcions del lloc de treball i es divideix en les categories següents:

- "-3": No es recull informació.
- "-2": No aplica.
- "-1": No sap/No contesta.
- "1": Funcions específiques de la titulació.
- "2": Funcions universitàries.
- "3": Funcions no universitàries.

En base aquests valors, el càlcul de la variable s'ha realitzat de la següent manera:
Si el valor de "funcions_c" és menor o igual a 3 i "funprop1" és igual a 1, s'estableix la variable "sobrequalificació" en 0, la qual cosa indica que l'individu no està sobrequalificat per al lloc.

Si el valor de "funcions_c" és menor o igual a 3 i "funprop1" és igual a 0, s'assigna el valor 1 a la variable "sobrequalificació", la qual cosa indica que l'individu està sobrequalificat per al lloc.

És important tenir en compte que els valors "-2" i "-3" en totes dues variables no es consideren en el càlcul de la sobrequalificació, ja que no són aplicables o no es recullen dades en aquestes situacions.

```
DATASET ACTIVATE ConjuntoDatos.
IF  (funcions_c <= 3 & funprop1=1)
Sobreq=0.
IF  (funcions_c <= 3 & funprop1=0)
Sobreq=1.
EXECUTE.
```

Equació 5.1: Query càlcul de la sobrequalificació

5.2 Depuració de la base de dades

Per dur a terme el projecte, es farà una depuració inicial de la base de dades, on es seleccionarà tots els registres que és possible calcular el factor de la sobrequalificació. A més a més, es seleccionarà les variables d'interès.

Per a seleccionar les variables d'interès, en primer lloc, es van realitzar transformacions de les variables categòriques a factors amb els seus respectius nivells. A continuació, es van filtrar els registres d'aquells individus que actualment estan treballant, és a dir, aquells que compleixen la condició de tenir el valor "1" en la variable "sitact", la qual indica que estan ocupats en l'actualitat. Una vegada obtingudes les variables numèriques i els factors amb els seus corresponents nivells, es van eliminar les columnes amb major quantitat de valors que manca, és a dir, aquelles columnes que presentaven més de 12.000 valors absents. També es van eliminar els registres que contenien valors que manca (NA).

Després de totes aquestes modificacions i transformacions en les variables categòriques, es van seleccionar les variables numèriques d'interès per a l'anàlisi. Aquestes variables es van agrupar en tres conjunts: "dades_satisfaccio", "dades_nivell" i "dades_utilitat". A continuació, es va fer una anàlisi factorial exploratòria utilitzant el conjunt de dades combinat ("x") i es va calcular l'estadístic KMO. A partir d'aquesta anàlisis factorial, s'ha seleccionat els valors (scores) de les variables PA1, PA2 i PA3 que representen les dimensions ocultes del conjunt

de variables estudiades. Aquests valors són estimacions de la posició relativa de cada persona en aquestes dimensions.

Els valors es calculen usant uns coeficients que s'han obtingut a partir de l'anàlisi. Aquests coeficients permeten estimar com és de present cada dimensió en cada persona, tenint en compte les respostes que han donat a les preguntes.

Amb aquests valors, es pot veure i comparar els patrons individuals amb relació a les dimensions que s'han identificat. Si una persona té un valor alt en una dimensió, significa que té una presència important d'aquesta característica. En canvi, si té un punt baix, significa que té una presència més baixa en aquesta característica.

Els valors són útils per analitzar les dades de manera més senzilla i comprendre millor els resultats. En lloc de tractar cada pregunta per separat, es pot agrupar-les en aquestes dimensions i veure com es relacionen les persones entre elles. Això ens facilita fer altres anàlisis, com ara buscar correlacions o fer prediccions, utilitzant aquestes estimacions en lloc de les respostes originals a les preguntes.

D'altra banda, en el cas de les variables categòriques, es van seleccionar els factors estudiats en els apartats 3.1.2 i 3.1.3, ja que són els que teòricament influeixen en la sobrequalificació. Per tant, es van seleccionar 6 variables qualitatives, juntament amb la variable resposta de la sobrequalificació.

Després de depurar la matriu de dades, aquesta compte amb 12.182 registres i 10 columnes. D'aquestes columnes, 3 variables són numèriques, 7 són qualitatives de les quals una correspon a la variable resposta de la sobrequalificació.

5.3 Anàlisi descriptiu de les dades

Es comença per definir cadascuna de les variables que conformen la base de dades. Per a cadascuna de les variables numèriques, es va elaborar una anàlisi descriptiva, es van generar histogrames i diagrames de caixa (boxplots) amb la finalitat d'observar la distribució de les dades. D'altra banda, per a les variables qualitatives es van crear taules de freqüència i diagrames de sectors.

5.3.1 Variables numèriques

A partir de l'anàlisi factorial, s'han obtingut 3 variables transformades. Aquestes variables representen combinacions lineals de les variables originals i capturen la informació més rellevant de les dades.

La primera variable obtinguda, anomenada "PA1", té càrregues factorials altes i positives en les següents variables originals: “satisf1_10”, “satisf2_10”, “satisf3_10”, “satisf5_10”, “ateor_10”, “apra_10”, “acoral_10”, “acescr_10”, “aequip_10”, “alider_10”, “asolprob_10”, “apresdec_10”, “acreat_10”, “aptecrit_10”, “ainfom_10”, “aidiom_10”, “adoc_10”. Aquest factor es pot interpretar com la "Satisfacció laboral i la utilitat percebuda de la formació i les habilitats relacionades amb el treball".

En la Taula 5.1 es presenten els valors mínim, primer quartil, mitjana, mediana, tercer quartil i màxim, la qual cosa permet conèixer la distribució de les dades. El rang de la variable s'estén des del valor mínim de -4.11 fins al valor màxim de 2.78, amb una mitjana de 0.

A partir de l'histograma i el boxplot (Figura 5.1 i Figura 5.2), s'observa que els valors segueixen una distribució semblant a una campana, amb una asimetria cap a la dreta. També es poden veure diversos valors outliers, ja que les dades són combinacions lineals d'altres variables.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PA1	-4.11	-0.54	0.12	0	0.65	2.78

Taula 5.1: Anàlisi descriptiu variable *PA1. Font: Elaboració pròpia.*

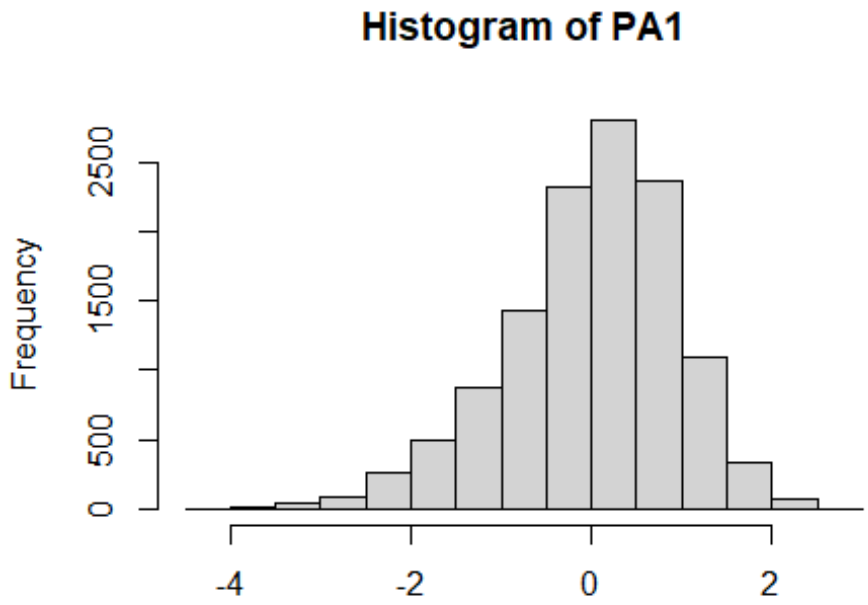


Figura 5.1: Histograma variable *PA1.* Font: Elaboració pròpia.

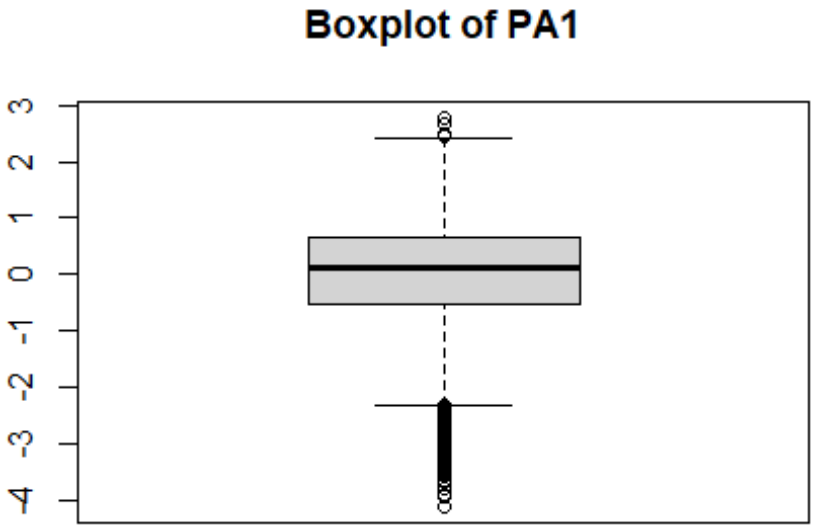


Figura 5.2: Boxplot variable *PA1*. Font: Elaboració pròpia.

La segona variable obtinguda, anomenada "PA2", té càrregues factorials altes i positives en les següents variables originals: “ncoral_10”, “ncescr_10”, “nequip_10”, “nlider_10”, “nsolprob_10”, “npresdec_10”, “ncreat_10”, “nptecrit_10”, “ngestio_10”, “ninform_10”, “nidiom_10”, “ndoc_10”, “satisf4_10”, “ateor_10”, “apra_10”, “aequip_10”, “alider_10”, “acreat_10”, “aptecrit_10”, “agestio_10”, “ainform_10”, “aidiom_10”, “adoc_10”. Aquest

factor es pot interpretar com les “Competències i habilitats tècniques i pràctiques relacionades amb el treball”.

S’aconsegueix una anàlisi descriptiu (Taula 5.2) similar a l’anterior variable, amb un rang que s'estén des del valor mínim de -3.68 fins al valor màxim de 2.99, amb una mitjana de 0.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PA2	-3.68	-0.58	0.11	0	0.68	2.99

Taula 5.2: Anàlisi descriptiu variable *PA2. Font: Elaboració pròpia.*

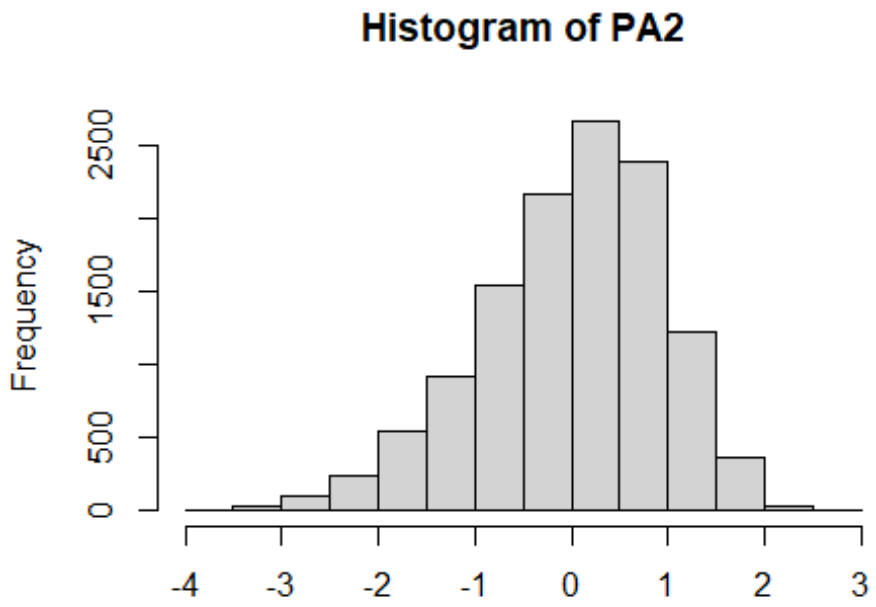


Figura 5.3: Histograma variable *PA2.* Font: Elaboració pròpia.

En últim lloc, obtenim la variable anomenada "PA3", té càrregues factorials altes i positives en les següents variables originals: “satisf4_10”, “ateor_10”, “apra_10”, “acoral_10”, “acescr_10”, “aequip_10”, “alider_10”, “acreat_10”, “aptecrit_10”, “agestio_10”, “adoc_10”. Aquest factor es pot interpretar com “Competències i habilitats de gestió i lideratge relacionades amb el treball”.

A partir de l'anàlisi i la Taula 5.3, s’assoleix una variable amb un rang de -4.49 a 2.11, amb una mitjana de 0.

En la Figura 5.4 s’observa una distribució en forma de campana amb simetria cap a la dreta.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PA3	-4.49	-0.46	0.15	0	0.61	2.11

Taula 5.3: Anàlisi descriptiu variable *PA3. Font: Elaboració pròpia.*

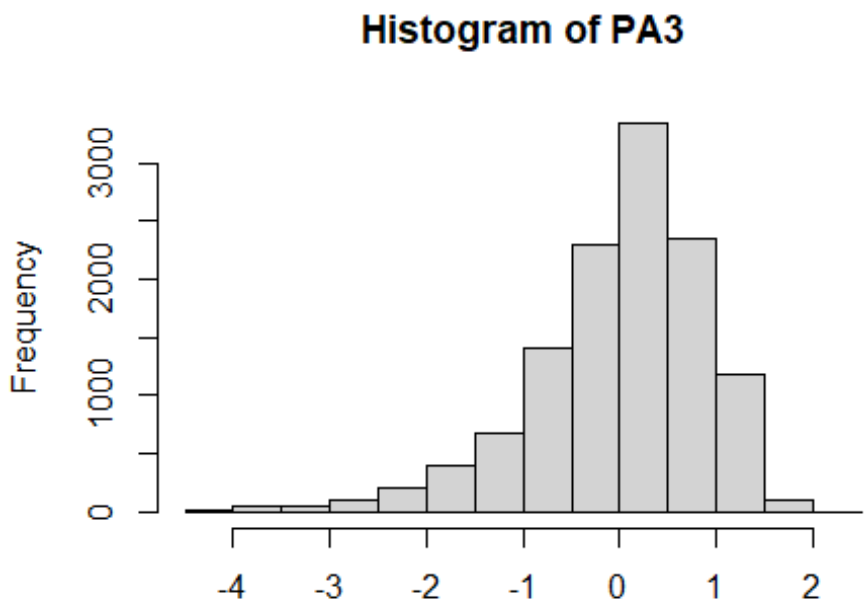


Figura 5.4: Histograma variable *PA3.* Font: Elaboració pròpia.

S'elabora una anàlisi bivariant d'aquestes variables mitjançant una matriu de correlacions per determinar el grau de relació entre elles. En la matriu de correlacions (Figura 5.5) es pot observar que no hi ha cap parell de variables amb un coeficient de correlació elevat.

Això indica que no hi ha una relació lineal fortament significativa entre les variables analitzades. En altres paraules, els valors de les variables no es correlacionen estretament entre si. Aquest resultat suggereix que les variables es comporten de manera independent i que no hi ha una dependència lineal evident entre elles.

Cal tenir en compte que aquesta anàlisi només considera la correlació lineal i no necessàriament reflecteix altres tipus de relacions no lineals o de dependències complexes que podrien existir entre les variables.

En resum, segons la matriu de correlacions, no es detecta cap relació lineal forta entre les variables analitzades.

	PA1	PA2	PA3
PA1	1.00000000	0.06318561	0.05013150
PA2	0.06318561	1.00000000	0.02094325
PA3	0.05013150	0.02094325	1.00000000

Figura 5.5: Matriu de correlacions variables numèriques. Font: Elaboració pròpia.

5.3.2 Variables categòriques

La variable "sitact", que fa referència a la situació laboral actual. En aquesta base de dades, s'ha filtrat anteriorment per seleccionar només els registres amb el valor "Treballo" en aquesta variable. Això significa que tots els registres de la variable "sitact" són "Treballo".

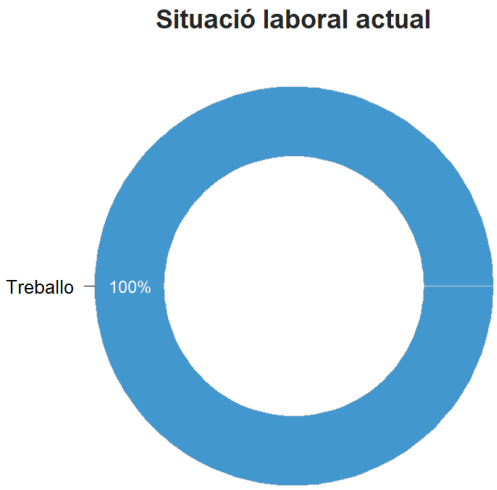


Figura 5.6: Diagrama de sectors Situació laboral actual. Font: Elaboració pròpia.

La variable "anyini_c" fa referència al temps transcorregut des de l'inici de la feina actual, és a dir, a l'experiència laboral. En la taula de freqüències, Taula 5.4, es pot observar que el grup amb més freqüència és "Menys d'un any" amb un 32.59%, seguit per "Fa 1 any" amb un 20.51%. D'altra banda, els grups amb menor freqüència són "Fa 3 anys" amb un 12.10% i "Fa més de 3 anys" amb un 17.15%. Aquesta informació proporciona una visió de la distribució dels empleats en funció de l'experiència laboral.

	Frequency	Percent
Fa 2 anys	2151	17.65720
Menys d'un any	3970	32.58907
Fa més de 3 anys	2089	17.14825
Fa 3 anys	1474	12.09982
Fa 1 any	2498	20.50566
Total	12182	100.00000

Taula 5.4: Taula de freqüències Experiència laboral. Font: Elaboració pròpia.

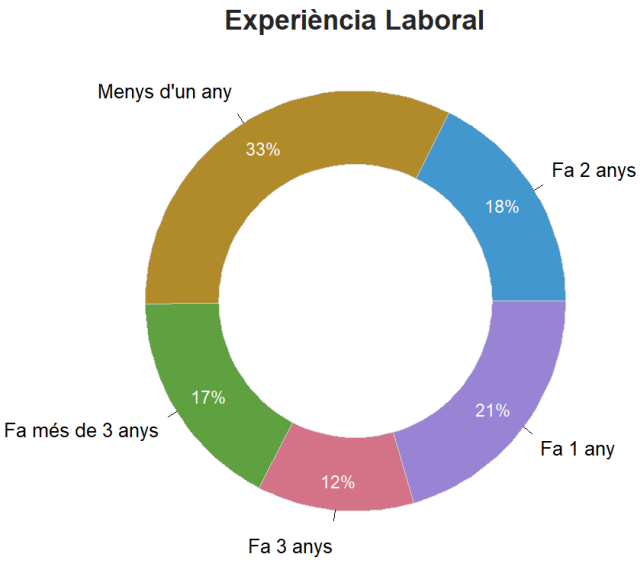


Figura 5.7: Diagrama de sectors Experiència laboral. Font: Elaboració pròpia.

La variable "autonom_c" indica si el tipus de contracte és autònom o no. En la taula de freqüències, Taula 5.5, es pot observar que només un 11% dels registres tenen un contracte autònom, el qual correspon a un total de 1366 registres.

	Frequency	Percent
No	10816	88.78673
Sí	1366	11.21327
Total	12182	100.00000

Taula 5.5: Taula de freqüències Tipus de contracte: autònom. Font: Elaboració pròpia.

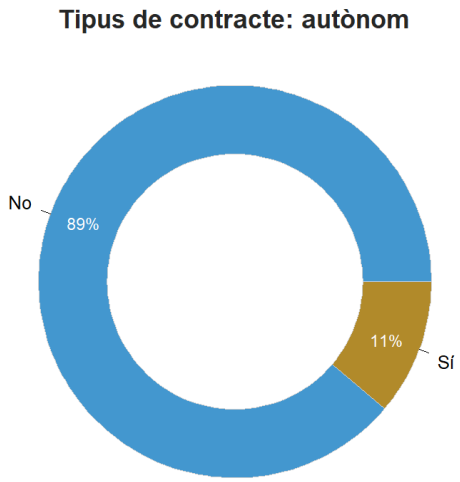


Figura 5.8: Diagrama de sectors Tipus de contracte: autònom. Font: Elaboració pròpia.

La variable “codi_a”, que correspon a la branca de titulació de l’enquestat. S'observa (Taula 5.6 i Figura 5.9) que la branca de titulació amb més freqüència és "Ciències socials i jurídiques" amb un 45%, que correspon a 5477 enquestats. A continuació, la branca d’"Enginyeria i Arquitectura" té una freqüència de 21% amb 2617 enquestats. La branca de "Ciències de la salut" representa un 18% amb 2196 enquestats.
Les branques d’"Arts i Humanitats" i "Ciències" tenen freqüències més baixes, amb 9% (1134 enquestats) i 6% (758 enquestats) respectivament.

	Frequency	Percent
Eng. i Arquitectura	2617	21.482515
Arts i Humanitats	1134	9.308816
Ciències de la salut	2196	18.026597
Ciències socials i jurídiques	5477	44.959777
Ciències	758	6.222295
Total	12182	100.000000

Taula 5.6: Taula de freqüències Branca Titulació. Font: Elaboració pròpia.

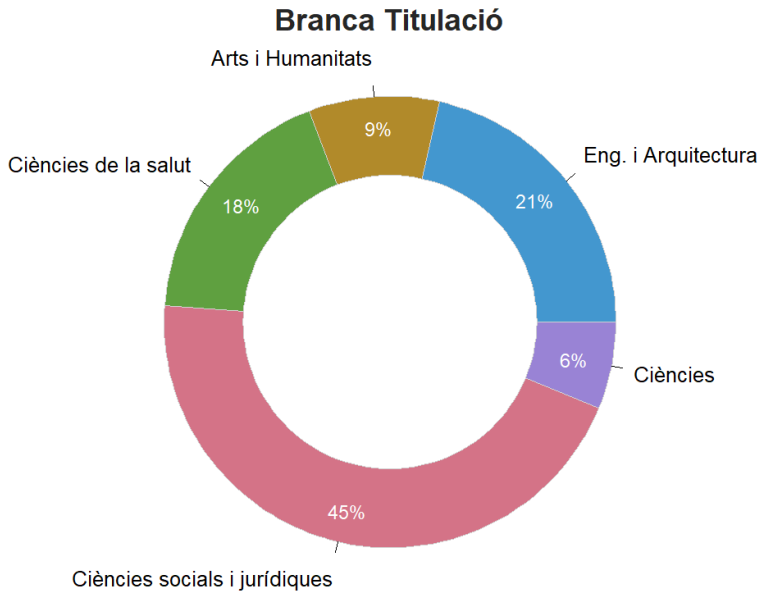


Figura 5.9: Diagrama de sectors Branca Titulació. Font: Elaboració pròpia.

La variable "guanys" fa referència als guanys anuals bruts dels enquestats. En la taula de freqüència (Taula 5.7) i el gràfic de sectors (Figura 5.10) es pot observar que el rang d'ingressos amb més freqüència és "Entre 24.001 i 30.000 €", amb un 14% de la mostra, el que correspon a 1740 enquestats. Seguidament, el rang "Entre 18.001 i 24.000 €" té una freqüència del 19%, amb 2373 enquestats.

Els rangs d'ingressos "Entre 40.000 i 50.000 €" i "Més de 50.000 €" tenen freqüències relativament molt baixes, amb un 3% i un 2%, respectivament.

Finalment, hi ha un grup que no ha especificat el rang d'ingressos ("NS/NC"), que té una freqüència d'un 12%.

La taula i el gràfic permet visualitzar la distribució dels enquestats segons els seus guanys anuals bruts i destaca la presència més gran d'enquestats en els rangs "Entre 18.001 i 24.000 €" i "Entre 24.001 i 30.000 €".

	Frequency	Percent
Entre 24.001 i 30.000	1740	14.283369
Entre 9.000 i 12.000 €	1139	9.349860
Menys de 9.000 €	1006	8.258086
Ns/Nc	1506	12.362502
Entre 18.001 i 24.000 €	2373	19.479560
Entre 15.001 i 18.000 €	1176	9.653587
Entre 40.000 i 50.000 €)	367	3.012642
Entre 12.001 i 15.000 €	1371	11.254310
Entre 30.001 i 40.000 €	1246	10.228205
Més de 50.000 €	258	2.117879
Total	12182	100.000000

Taula 5.7: Taula de freqüències Guanys anuals bruts. Font: Elaboració pròpia.

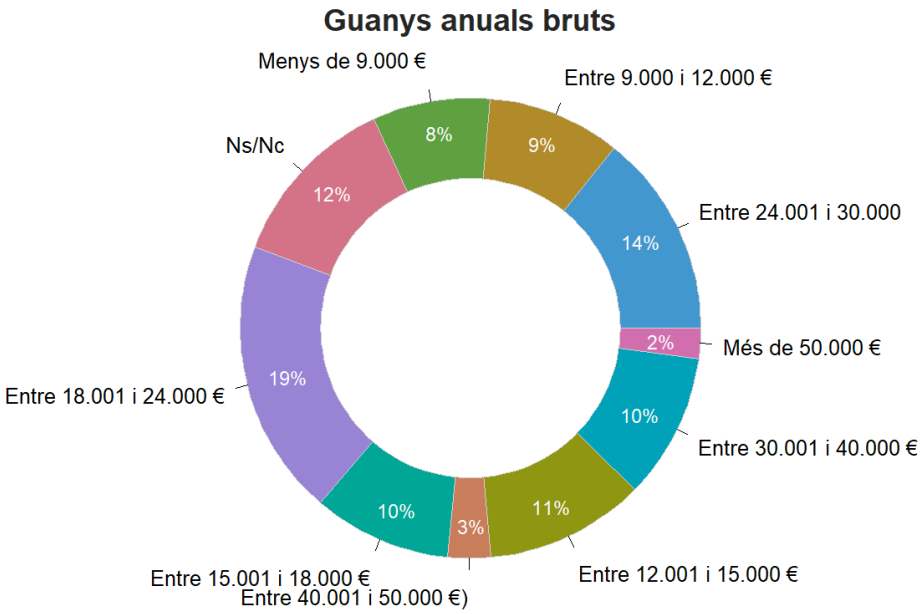


Figura 5.10: Diagrama de sectors Guanys anuals bruts. Font: Elaboració pròpia.

Per acabar, la variable “numtreb” proporciona informació sobre el nombre de treballadors de cada empresa on treballa l’estudiant. Analitzant la taula de freqüència, Taula 5.8, i el gràfic de sectors, Figura 5.11, es pot observar la distribució del nombre de treballadors en diferents rangs.

El rang amb la freqüència més alta és “Més de 500” treballadors, el qual representa un 26%. Això suggereix que hi ha un nombre considerable d'empreses amb una plantilla molt gran.

D’altra banda, “Entre 251 i 500” treballadors i "Ns/Nc" (No sap/No contesta), tenen la freqüència més baixa, amb un 6% cada una categoria. Això indica que hi ha un nombre moderat d'empreses amb una plantilla mitjana i un petit nombre d’enquestats que no van proporcionar informació sobre el nombre de treballadors de les seves respectives empreses.

A través d'aquesta variable, es pot obtenir una visió general de la distribució del nombre de treballadors, destacant tant les grans empreses com aquelles amb plantilles més reduïdes.

	Frequency	Percent
Entre 11 i 50	2749	22.566081
Entre 51 i 100	1206	9.899852
10 o menys	2516	20.653423
Més de 500	3133	25.718273
Ns/Nc	776	6.370054
Entre 251 i 500	737	6.049910
Entre 101 i 250	1065	8.742407
Total	12182	100.000000

Taula 5.8: Taula de freqüències Nombre de treballadors. Font: Elaboració pròpia.

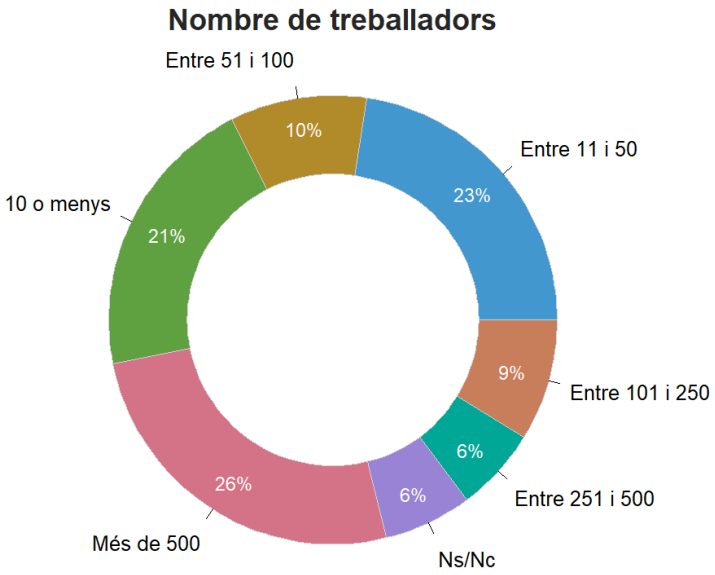


Figura 5.11: Diagrama de sectors Nombre de treballadors. Font: Elaboració pròpia.

6. APLICACIÓ DELS MODELS DE CLASSIFICACIÓ

A continuació, es realitzarà l'estimació de models de classificació: arbres de decisió, *Naive Bayes*, SVM (*Support Vector Machines*) i regressió logística. S'avaluarà el rendiment d'aquests models utilitzant mètriques com l'exactitud, la precisió, la sensibilitat (Recall), l'especificitat i l'àrea sota la corba ROC.

En primer lloc, s'ha fet la partició de les dades en un conjunt d'entrenament, representat per 'training', fent servir una mostra aleatòria d'aproximadament el 67% de les files de la base de dades. Aquesta mostra serà emprada per tots els models de classificació. Aquest conjunt de dades té 8121 observacions. Pel que fa a què el conjunt de dades de prova en té 4061.

6.1 Arbres de decisió

Es du a terme el model a partir de la funció 'rpart()' amb la variable resposta "Sobreq" i nou variables predictores: "PA1", "PA2", "PA3", "sitact", "anyini_c", "autonom_c", "codi_a", "guanys" i "numtreb". La mètrica d'impuresa usada per a dividir els nodes de l'arbre ha estat la "information gain" (guany d'informació).

A partir del resum del model (Figura 6.1), s'obté que s'han estimat quatre valors de paràmetres de complexitat (CP): 0.04115226, 0.04023777, 0.02103338 i 0.01000000. Aquests valors representen diferents nivells de complexitat de l'arbre de decisió. La variable més important en la construcció de l'arbre és "PA3", seguida de "PA1", "codi_a" i "PA2". Aquestes variables són les variables que millor defineixen les categories de la variable resposta i determinen els nodes fills.

```
Call:
rpart(formula = Sobreq ~ PA1 + PA2 + PA3 + sitact + anyini_c +
    autonom_c + codi_a + guanys + numtreb, data = dataTrain,
    method = "class", parms = list(split = "information"))
  n= 8121

          CP nsplit rel error    xerror       xstd
1 0.04115226      0 1.0000000 1.0000000 0.01827867
2 0.04023777      2 0.9176955 0.9780521 0.01814993
3 0.02103338      3 0.8774577 0.8948331 0.01762273
4 0.01000000      4 0.8564243 0.8779150 0.01750765

Variable importance
   PA3    PA1 codi_a    PA2
    80     14      6      1

Node number 1: 8121 observations,    complexity param=0.04115226
  predicted class=Normal           expected loss=0.2693018  P(node) =1
    class counts:  5934  2187
   probabilities: 0.731 0.269
  left son=2 (6277 obs) right son=3 (1844 obs)
  Primary splits:
      PA3      < -0.5228335 to the right, improve=302.68790, (0 missing)
      PA1      < -0.8417729 to the right, improve=103.04130, (0 missing)
      codi_a   splits as  LRLLR, improve= 82.43011, (0 missing)
      guanys   splits as  LRRLLLLRLL, improve= 78.17689, (0 missing)
      anyini_c splits as  LLRLL, improve= 42.30468, (0 missing)
  Surrogate splits:
      PA2 < 1.951311   to the left,  agree=0.775, adj=0.011, (0 split)
      PA1 < -2.716939  to the right, agree=0.774, adj=0.005, (0 split)

Node number 2: 6277 observations
  predicted class=Normal           expected loss=0.2010515  P(node) =0.7729344
    class counts:  5015  1262
   probabilities: 0.799 0.201
```

Figura 6.1: Model Arbre de decisió. Font: Elaboració pròpia.

L'arbre de decisió resultant (Figura 6.2) del model (Figura 6.1) té un total de 4 nodes, on el node arrel (número 1) representa a totes les observacions de la mostra. Al final de les branques de l'arbre, es troben els nodes terminals que representen les classes predites.

El node 1, és a dir, el node arrel, conte totes les observacions d'entrenament (8121 observacions en total). Es divideix en dos nodes fills, el node 2 i el node 3.

El node 2, és el node fill més gran i conté 6277 observacions d'entrenament. La classe majoritària en aquest node és "Normal", amb una proporció d'aproximadament el 79.9%. Aquesta divisió es basa en característiques com "PA3", "PA1", "codi_a", "guanys" i "anyini_c".

El node 3 conté 1844 observacions. Mostra una classificació més equilibrada entre les classes "Normal" i "Sobrequalificat". Aquesta divisió es basa en característiques com "PA1", "PA3", "codi_a", "guanys" i "PA2".

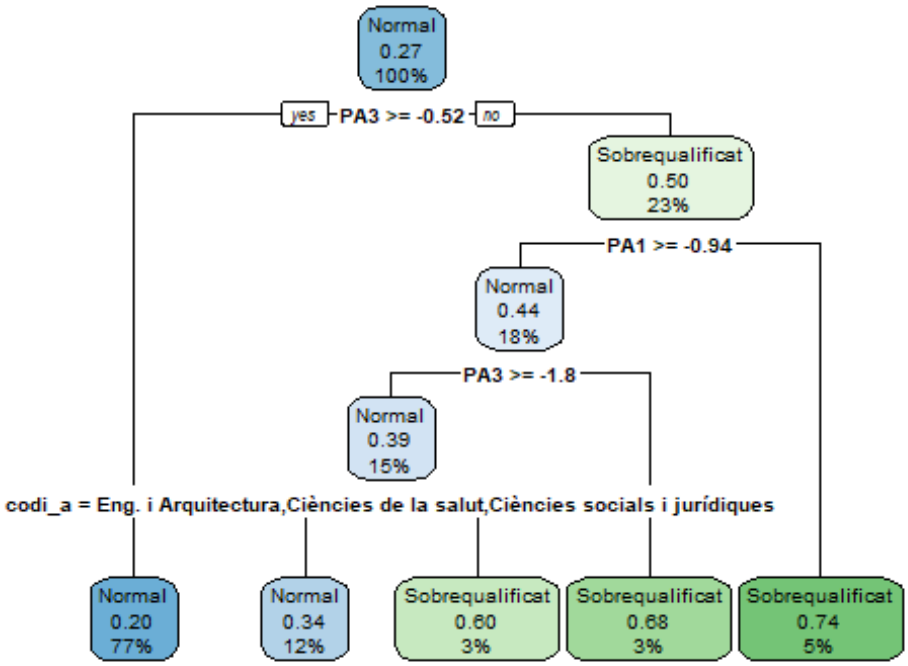


Figura 6.2: Arbre de decisió. Font: Elaboració pròpia.

Per a avaluar el model de classificació, en primer lloc, s'ha realitzat una matriu de confusió per tal de tenir una idea de com és de bo el model, és a dir, com seran de bones les prediccions. En aquest cas, la matriu de confusió, Figura 6.3, mostra que per la classe "Normal", el model va predir correctament 5665 casos com a "Normal" (veritables positius) i 1604 casos com "Sobrequalificat" (falsos positius). Per la classe "Sobrequalificat", el model va predir correctament 583 casos com "Sobrequalificat" (veritables negatius) i 269 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction        Normal Sobrequalificat
  Normal            5665             1604
  Sobrequalificat    269              583
```

Figura 6.3: Matriu de confusió dadesTrain. Font: Elaboració pròpia.

A partir d'aquesta matriu s'ha calculat les mètriques derivades. Es calcularà específicament les estadístiques de la classe "Normal" per tal d'obtenir una idea de com de bé el model està identificant correctament els casos de "Normal".

Primer, s'ha calculat l'exactitud (*accuracy*), que representa la proporció d'instàncies classificades correctament en el conjunt d'entrenament. En aquest cas, el valor aconseguit de 0.7694 indica que el model té una precisió del 76.94% en la classificació de les instàncies del conjunt d'entrenament.

Segon, s'ha calculat la precisió i s'ha aconseguit un valor de 0.7793, la qual cosa indica que s'han classificat correctament 77.93% en la classificació dels casos positius en el conjunt d'entrenament.

Seguidament, s'ha calculat la sensibilitat, que és un indicador de la proporció de casos positius que el model ha identificat correctament. En aquest cas, el valor assolit de 0.9547 indica que el model té un recall del 95.47% en la classificació dels casos positius en el conjunt d'entrenament.

Finalment, s'ha calculat l'especificitat, que és un indicador de la proporció de casos negatius que el model ha identificat correctament. El valor obtingut 0.2666 indica que el model té una especificitat del 26.66% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, s'ha calculat la corba ROC i l'àrea sota la corba (AUC). S'ha assolit un valor de 0.6460, la qual cosa indica la capacitat del model per a distingir entre les classes objectiu en les dades d'entrenament, és a dir, si l'enquestat està sobrequalificat o no.

En la Figura 6.4 es pot veure representada la corba de ROC on en l'eix X es representa la taxa de falsos positius, que és la proporció de casos negatius incorrectament classificats com a positius. I en l'eix Y, es troba la taxa de veritables positius o sensibilitat, que és la proporció de casos positius correctament classificats com a positius. S'observa que la corba es posiciona relativament per sobre de la línia clara d'aleatorietat el que gràficament indica una bona predicció.

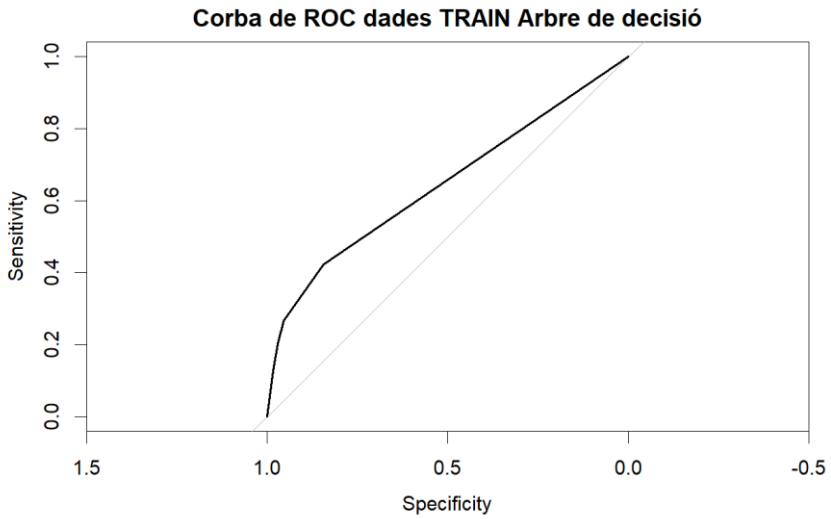


Figura 6.4: Corba de ROC model Arbre de decisió. Font: Elaboració pròpia.

És important tenir en compte que aquestes mètriques són utilitzades per a avaluar el rendiment del model en les dades d'entrenament. Per a obtenir una avaluació més completa, es recomana també avaluar el model en dades de prova.

Per tant, la matriu de confusió per les dades de prova, Figura 6.5, mostra que per la classe "Normal", el model va predir correctament 2905 casos com a "Normal" (veritables positius) i 849 casos com "Sobrequalificat" (falsos positius). Per la classe "Sobrequalificat", el model va predir correctament 232 casos com "Sobrequalificat" (veritables negatius) i 75 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction       Normal Sobrequalificat
  Normal           2905             849
  Sobrequalificat    75             232
```

Figura 6.5: Matriu de confusió dadesTest. Arbres de decisió. Font: Elaboració pròpia.

L'exactitud (*accuracy*) del model per a les dades de prova és de 0.7725. Aquest valor indica que el model té una precisió del 77.25% en la classificació de les instàncies del conjunt de prova.

La precisió és de 0.7738, la qual cosa indica que s'han classificat correctament 77.38% en la classificació dels casos positius en el conjunt d'entrenament.

El valor del recall per a les dades de prova és de 0.9748. Això significa que el model ha identificat correctament el 97.48% dels casos positius en el conjunt de prova.

Finalment, el valor obtingut de l'especificitat 0.2146 indica que el model té una especificitat del 21.46% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, l'àrea sota la corba ROC per a les dades de prova és de 0.6478. Això indica que el model té una capacitat moderada per a classificar correctament les instàncies en les dues classes, "Normal" i "Sobrequalificat". La representació de la corba és similar a les dades d'entrenament i es troba en l'annex (**ANNEX**).

	Dades entrenament	**Dades prova**
Exactitud	0.7694	0.7725
Precisió	0.7793	0.7738
Sensibilitat	0.9547	0.9748
Especificitat	0.2666	0.2146
Àrea Corba ROC	0.6460	0.6478

Taula 6.1: Mètriques d'avaluació del model de classificació Arbres de decisió. Font: Elaboració pròpia.

6.2 Naive Bayes

En primer lloc, s'ha aplicat l'algoritme mitjançant la funció 'naive_bayes()' amb totes les variables numèriques i categòriques amb la variable sobrequalificació com a variable resposta. Això significa que s'està classificant la variable resposta en funció de totes les altres variables predictores presents en el conjunt de dades d'entrenament.

A partir del model (Figura 6.6), s'obté que el paràmetre *Laplace* s'estableix en 0. La correcció de *Laplace* s'utilitza per evitar que es produeixin probabilitats de zeros en el cas que una categoria no aparegui en el conjunt d'entrenament. En aquest cas, en ser 0, indica que no s'ha aplicat cap suavitzat de *Lapalce*.

La distribució condicional de les característiques es divideix en tres: *Bernoulli*, Categòrica i Gaussiana. Això indica que l'algoritme *Naive Bayes* ha fet servir diferents distribucions per modelar les característiques depenen del seu tipus.

S'aconsegueix que la probabilitat a priori de la classe "Normal" és 0.7307 i la probabilitat a priori de la classe "Sobrequalificat" és 0.2693. Aquestes probabilitats representen la proporció de mostres en el conjunt d'entrenament que pertanyen a cada classe.

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

- Call: naive_bayes.formula(formula = Sobreq ~ ., data = dataTrain)
- Laplace: 0
- Classes: 2
- Samples: 8121
- Features: 9
- Conditional distributions:
    - Bernoulli: 1
    - Categorical: 5
    - Gaussian: 3
- Prior probabilities:
    - Normal: 0.7307
    - Sobrequalificat: 0.2693
```

Figura 6.6: Model *Naive Bayes*. Font: Elaboració pròpia.

S'han representat les variables en funció de la variable resposta. Les representacions gràfiques de les distribucions condicionals ajuden a visualitzar com el model *Naive Bayes* està separant les classes en funció de les característiques del conjunt de dades.

La Figura 6.7 s'observa una diferència en les categories de l'experiència laboral. En el gràfic, es destaca que la categoria d'experiència laboral de "Menys d'un any" mostra una probabilitat més alta tant en la classe "Sobrequalificat" com en la classe "Normal", en comparació amb les altres categories. Això suggereix que l'experiència laboral de menys d'un any pot influir en la classificació, independentment de si es considera com "Sobrequalificat" o "Normal".

D'altra banda, es veu que la categoria de "Fa 3 anys" té una probabilitat més baixa en totes dues classes de la variable resposta. Això implica que, en general, les mostres amb una experiència laboral de fa 3 anys tendeixen a tenir una menor probabilitat de pertànyer a les classes "Sobrequalificat" i "Normal" en comparació amb les altres categories.

Figura 6.7: Representació gràfica variable resposta i anyini_c. Font: Elaboració pròpia.

En relació amb la variable "Branca de titulació", es pot observar diferències en les probabilitats entre les categories i les classes de la variable resposta (Figura 6.8).

Específicament, la categoria "Ciències socials i jurídiques" mostra una probabilitat més elevada tant en la classe "Normal" com en la classe "Sobrequalificat". Això indica que les mostres pertanyents a aquesta categoria tenen una major probabilitat de classificar-se com a "Normal" o "Sobrequalificat" en comparació amb les altres categories de la variable "Branca de titulació".

D'altra banda, es nota que la categoria "Art i Humanitats" té una probabilitat més alta de pertànyer a la classe "Sobrequalificat" en comparació amb la classe "Normal". Això suggereix que les mostres relacionades amb el camp d'estudi d'"Art i Humanitats" tenen una major probabilitat de ser classificades com "Sobrequalificat" en lloc de "Normal".

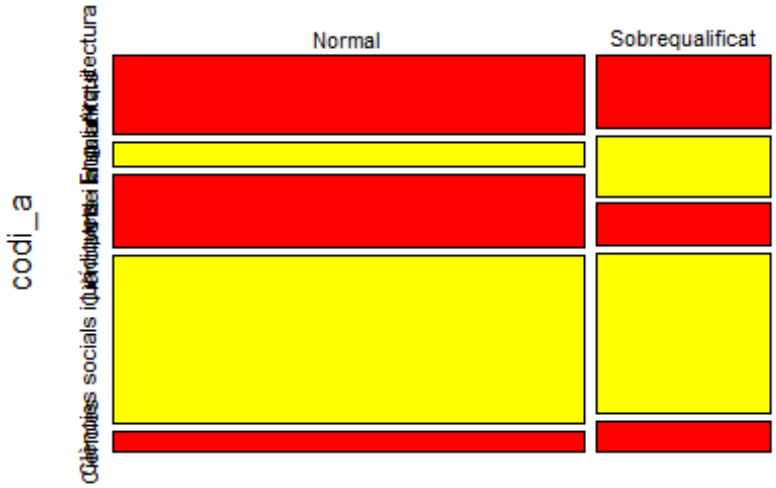


Figura 6.8: Representació gràfica variable resposta i codi_a. Font: Elaboració pròpia.

Per acabar, la Figura 6.9 que fa referència a la variable numèrica PA1, es pot observar una clara diferència en les probabilitats entre les classes "Sobrequalificat" i "Normal".

Específicament, si el valor de la variable "PA1" es troba al voltant de 0, és més probable que la mostra sigui classificada com a "Normal". No obstant això, quan la variable "PA1" està en el rang entre -2 i -2.5, les dues classes ("Sobrequalificat" i "Normal") tenen probabilitats aproximades. Això significa que les mostres amb valors de "PA1" en aquest rang tenen una probabilitat similar de pertànyer a qualsevol de les dues classes.

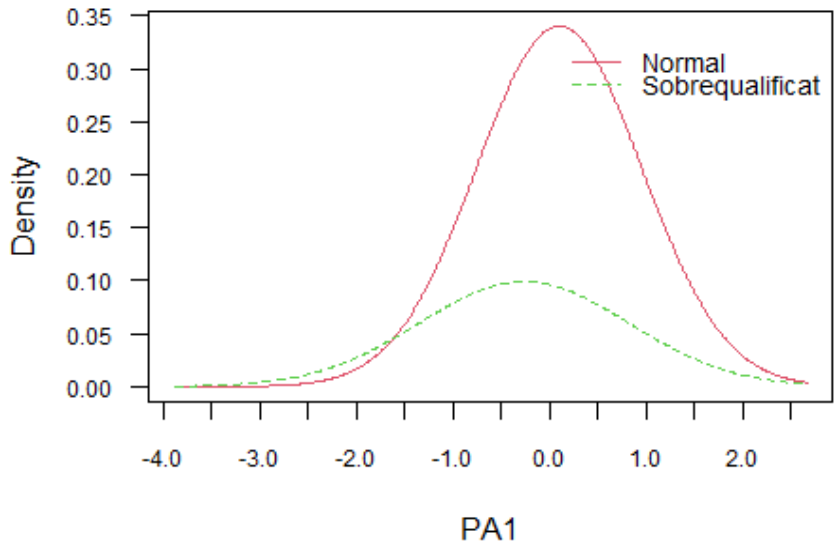


Figura 6.9: Gràfica densitat variable resposta i PA1. Font: Elaboració pròpia.

Per a avaluar el model de classificació, s'ha realitzat una matriu de confusió per tal de tenir una idea de com és de bo el model, és a dir, com seran de bones les prediccions. En aquest cas, la matriu de confusió, Figura 6.10, mostra que per la classe "Normal", el model va predir correctament 5458 casos com a "Normal" (veritables positius) i 1412 casos com "Sobrequalificat" (falsos positius). Per la classe "Sobrequalificat", el model va predir correctament 775 casos com "Sobrequalificat" (veritables negatius) i 476 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction       Normal Sobrequalificat
  Normal           5458            1412
  Sobrequalificat   476             775
```

Figura 6.10: Matriu de confusió dadesTrain. *Naive Bayes*. Font: Elaboració pròpia.

A partir d'aquesta matriu s'ha calculat les mètriques derivades, que ajudaran a la selecció del millor model de classificació.

Primer, s'ha calculat l'exactitud (*accuracy*), que representa la proporció d'instàncies classificades correctament en el conjunt d'entrenament. En aquest cas, el valor aconseguit de 0.7675 indica que el model té una precisió del 76.75% en la classificació de les instàncies del conjunt d'entrenament.

Segon, s'ha calculat la precisió i s'ha obtingut un valor de 0.7945, la qual cosa indica que s'han classificat correctament 79.45% en la classificació dels casos positius en el conjunt d'entrenament.

Després, s'ha calculat el *recall*, que és un indicador de la proporció de casos positius que el model ha identificat correctament. En aquest cas, el valor assolit de 0.9198 indica que el model té una sensibilitat del 91.98% en la classificació dels casos positius en el conjunt d'entrenament.

En últim lloc, s'ha calculat l'especificitat, que és un indicador de la proporció de casos negatius que el model ha identificat correctament. El valor obtingut 0.3544 indica que el model té una especificitat del 35.44% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, s'ha calculat la corba ROC i l'àrea sota la corba (AUC). S'ha aconseguit un valor de 0.7304, la qual cosa indica la capacitat del model per a distingir entre les classes objectiu en les dades d'entrenament, és a dir, si l'enquestat està sobrequalificat o no.

En la Figura 6.11 es pot veure representada la corba de ROC on en l'eix X es representa la taxa de falsos positius, que és la proporció de casos negatius incorrectament classificats com a positius. I en l'eix Y, es troba la taxa de veritables positius o sensibilitat, que és la proporció de casos positius correctament classificats com a positius. S'observa que la corba es posiciona per sobre de la línia clara d'aleatorietat el que gràficament indica una bona predicció.

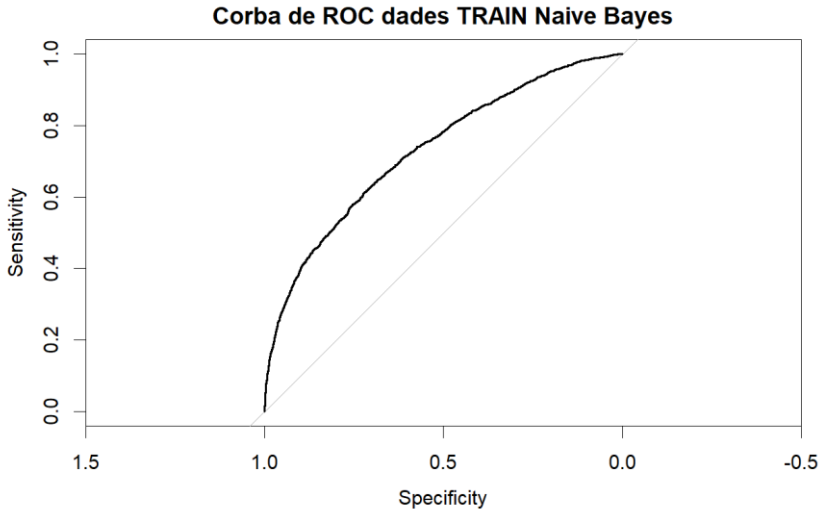


Figura 6.11: Corba de ROC model *Naive Bayes*. Font: Elaboració pròpia.

És important tenir en compte que aquestes mètriques són utilitzades per a avaluar el rendiment del model en les dades d'entrenament. Per a obtenir una avaluació més completa, es recomana també avaluar el model en dades de prova.

Per tant, la matriu de confusió per les dades de prova, Figura 6.12, mostra que per la classe "Normal", el model va predir correctament 2738 casos com a "Normal" (veritables positius) i 706 casos com "Sobrequalificat" (falsos positius). Per la classe "Sobrequalificat", el model va predir correctament 375 casos com "Sobrequalificat" (veritables negatius) i 242 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction       Normal Sobrequalificat
  Normal           2738             706
  Sobrequalificat   242             375
```

Figura 6.12: Matriu de confusió dadesTest. Naive Bayes. Font: Elaboració pròpia.

L'exactitud (*accuracy*) del model per a les dades de prova és de 0.7666. Aquest valor indica que el model té una precisió del 76.66% en la classificació de les instàncies del conjunt de prova.

La precisió és de 0.7950, la qual cosa indica que s'han classificat correctament 79.50% en la classificació dels casos positius en el conjunt d'entrenament.

El valor del recall per a les dades de prova és de 0.9188. Això significa que el model ha identificat correctament el 91.88% dels casos positius en el conjunt de prova.

Finalment, el valor obtingut de l'especificitat 0.3469 indica que el model te una especificitat del 34.69% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, l'àrea sota la corba ROC per a les dades de prova és de 0.7345. Això indica que el model té una capacitat moderada per a classificar correctament les instàncies en les dues classes, "Normal" i "Sobrequalificat". La representació de la corba és similar a les dades d'entrenament i es troba en l'annex (**ANNEX**).

	Dades entrenament	**Dades prova**
Exactitud	0.7675	0.7666
Precisió	0.7945	0.7950
Sensibilitat	0.9198	0.9188
Especificitat	0.3544	0.3469
Àrea Corba ROC	0.7304	0.7345

Taula 6.2: Mètriques d'avaluació del model de classificació *Naive Bayes*. Font: Elaboració pròpia.

6.3 Support Vector Machines

En aquest cas, primer s'ha eliminat la variable "sitact" pel fet que només tenia una categoria. A continuació s'ha aplicat l'algoritme mitjançant la funció 'svm()' amb totes les variables numèriques i categòriques i la variable sobrequalificació com a variable resposta. Això significa que s'està classificant la variable resposta en funció de totes les altres variables predictores presents en el conjunt de dades d'entrenament.

A partir del model, Figura 6.13, s'obté que es tracta d'un model de SVM per classificació. S'ha utilitzat *kernel* lineal, la qual cosa és adequat per a capturar relacions lineals entre les variables predictores i la variable resposta. El valor del paràmetre de cost usat és 1, el qual controla

l'equilibri entre la classificació correcta dels punts de dades i la suavitat de la frontera de decisió. Un valor més alt de cost pot conduir a una major precisió en les dades d'entrenament, però pot resultar en un model més complex i propens al sobreajust. A més, el model té un total de 4294 vectors de suport, que són punts de dades pròximes a la frontera de decisió i juguen un paper crucial en la classificació. Aquests vectors de suport representen els casos més difícils de classificar i són fonamentals per a l'eficàcia del model.

```
Call:
svm(formula = Sobreq ~ ., data = dataTrain2, kernel = "linear", scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  4294
```

Figura 6.13: Model SVM. Font: Elaboració pròpia.

La Figura 6.14, extret a partir del model, cada punt representa una observació del conjunt de dades i està situat en funció dels seus valors en les variables PA1 (eix x) i PA2 (eix y). Els punts estan acolorits segons la seva classe de sobrequalificació ("Normal" o "Sobrequalificat").

A més, s'han superposat els vectors de suport del model SVM en color vermell. Aquests vectors representen els punts de dades més pròximes a la frontera de decisió, és a dir, aquells que influeixen en la separació entre les classes. Els vectors de suport són punts clau per a determinar la ubicació i l'orientació de l'hiperplà que s'utilitza per a classificar noves observacions.

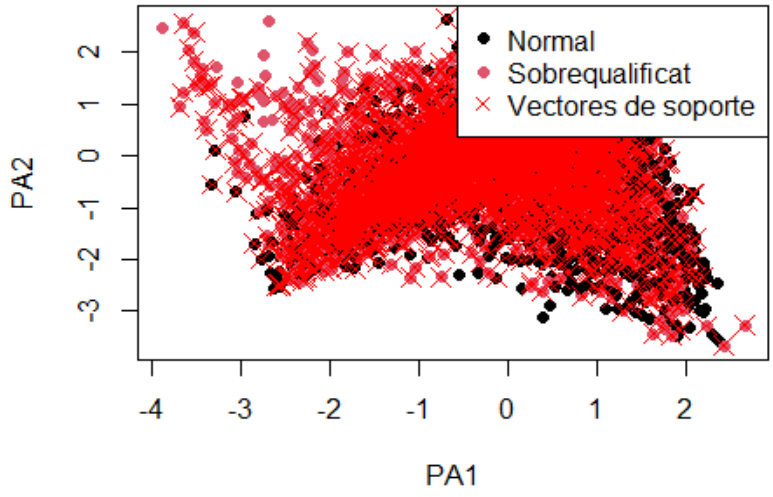


Figura 6.14: Representació gràfica variable resposta i PA1 i PA2. Font: Elaboració pròpia.

S’ha realitzat una matriu de confusió per tal de tenir una idea de com és de bo el model, és a dir, com seran de bones les prediccions. En aquest cas, la matriu de confusió, Figura 6.15, mostra que per la classe “Normal”, el model va predir correctament 5825 casos com a "Normal" (veritables positius) i 1774 casos com "Sobrequalificat" (falsos positius). Per la classe “Sobrequalificat”, el model va predir correctament 413 casos com "Sobrequalificat" (veritables negatius) i 109 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction       Normal Sobrequalificat
  Normal           5825            1774
  Sobrequalificat   109             413
```

Figura 6.15: Matriu de confusió dadesTrain2. SVM. Font: Elaboració pròpia.

En primer lloc, s’ha calculat l'exactitud (*accuracy*), que representa la proporció d'instàncies classificades correctament en el conjunt d'entrenament. En aquest cas, el valor de 0.7681 indica que el model té una precisió del 76.81% en la classificació de les instàncies del conjunt d'entrenament.

En segon lloc, s’ha calculat la precisió i s’ha obtingut un valor de 0.7665, la qual cosa indica que s’han classificat correctament 76.65% en la classificació dels casos positius en el conjunt d'entrenament.

Seguidament, s’ha calculat la sensibilitat (*recall*), que és un indicador de la proporció de casos positius que el model ha identificat correctament. En aquest cas, el valor de 0.9816 indica que el model té una sensibilitat del 98.16% en la classificació dels casos positius en el conjunt d'entrenament.

Finalment, s’ha calculat l’especificitat, que és un indicador de la proporció de casos negatius que el model ha identificat correctament. El valor aconseguit 0.1888 indica que el model té una especificitat del 18.88% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, s’ha calculat la corba ROC i l'àrea sota la corba (AUC). El valor assolit de 0.5852 indica la capacitat del model per a distingir entre les classes objectiu en les dades d'entrenament, és a dir, si l'enquestat està sobrequalificat o no.

En la Figura 6.16 es pot veure representada la corba de ROC on en l'eix X es representa la taxa de falsos positius, que és la proporció de casos negatius incorrectament classificats com a positius. I en l'eix Y, es troba la taxa de veritables positius o sensibilitat, que és la proporció de casos positius correctament classificats com a positius. S'observa que la corba es troba prop de la línia clara d'aleatorietat el que gràficament indica una predicció insignificant, ja que, és similar a una selecció aleatòria.

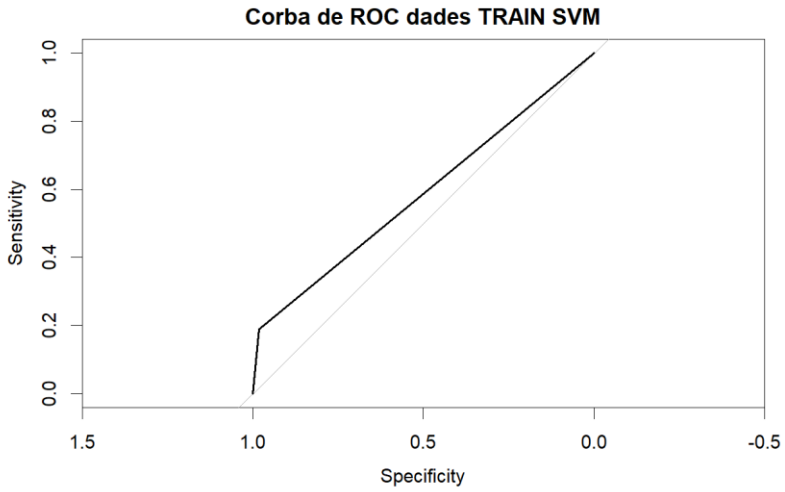


Figura 6.16: Corba de ROC model SVM. Font: Elaboració pròpia.

És important tenir en compte que aquestes mètriques són utilitzades per a avaluar el rendiment del model en les dades d'entrenament. Per a obtenir una avaluació més completa, es recomana també avaluar el model en dades de prova.

Per tant, la matriu de confusió per les dades de prova, Figura 6.17, mostra que per la classe "Normal", el model va predir correctament 2924 casos com a "Normal" (veritables positius) i 885 casos com "Sobrequalificat" (falsos positius). Per la classe "Sobrequalificat", el model va predir correctament 196 casos com "Sobrequalificat" (veritables negatius) i 56 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction       Normal Sobrequalificat
  Normal           2924              885
  Sobrequalificat    56              196
```

Figura 6.17: Matriu de confusió dadesTest2. SVM. Font: Elaboració pròpia.

L'exactitud (*accuracy*) del model per a les dades de prova és de 0.7683. Aquest valor indica que el model té una precisió del 76.83% en la classificació de les instàncies del conjunt de

prova. És a dir, aproximadament el 76.83% de les instàncies es van classificar correctament pel model.

La precisió és de 0.7677, la qual cosa indica que s'han classificat correctament 76.77% en la classificació dels casos positius en el conjunt d'entrenament.

El valor del *recall* per a les dades de prova és 0.9812. Això indica que el model ha identificat correctament el 98.12% dels casos positius en el conjunt de prova.

El valor obtingut de l'especificitat 0.1813 indica que el model té una especificitat del 18.13% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, l'àrea sota la corba ROC per a les dades de prova és de 0.5813. Això indica que el model té una capacitat limitada per a classificar correctament les instàncies en les dues classes, "Normal" i "Sobrequalificat". Aquest valor és més proper a 0.5, que indica que el model té una capacitat predictiva molt semblant a una predicció aleatòria. La representació de la corba és similar a les dades d'entrenament i es troba en l'annex (**ANNEX**).

	Dades entrenament	**Dades prova**
Exactitud	0.7681	0.7683
Precisió	0.7665	0.7677
Sensibilitat	0.9816	0.9812
Especificitat	0.1888	0.1813
Àrea Corba ROC	0.5852	0.5813

Taula 6.3: Mètriques d'avaluació del model de classificació SVM. Font: Elaboració pròpia.

6.4 Regressió Logística

En primer lloc, s'ha ajustat un model de classificació logística amb la funció 'glm()'. El model s'ha construït per predir la variable de sobrequalificació, 'Sobreq' en funció de les variables predictores "PA1", "PA2", "PA3", "anyini_c", "autonom_c", "codi_a", "guanys" i "numtreb". Excepte la variable "sitact", ja que només conté una categoria.

El model (Figura 6.18) presenta els coeficients estimats per cada una de les variables predictores. Si el coeficient d'una variable és positiu, com en el cas de "guanysEntre 9.000 i 12.000 €" amb un coeficient de 0.647411, significa que un increment en el valor d'aquesta variable està associat amb un augment en la probabilitat de pertànyer a la classe "Sobrequalificat". És a dir, a mesura que la variable "guanysEntre 9.000 i 12.000 €" augmenta,

la probabilitat que una instància sigui classificada com "Sobrequalificat" en lloc de "Normal" també augmenta. Si el coeficient d'una variable és negatiu, com en el cas de "PA3" amb un coeficient de -0.737409, indica que un increment en el valor d'aquesta variable s'associa amb una disminució en la probabilitat de pertànyer a la classe "Sobrequalificat". Un coeficient pròxim a zero, com en el cas de "numtrebEntre 251 i 500" amb un coeficient de 0.089029, suggereix que aquesta variable té una influència relativament petita en la probabilitat de pertànyer a la classe "Sobrequalificat" en comparació amb la classe de referència, que en aquest cas és "Normal". En altres paraules, els canvis en aquesta variable tenen menys impacte en la classificació de la classe "Sobrequalificat" en relació amb la classe "Normal".

El model inclou els nivells de les variables categòriques, com 'anyini_c' (amb categories com 'Menys d'un any', 'Fa més de 3 anys', etc.) i 'guanys' (amb categories com 'Menys de 9000 €', 'Entre 9000 i 12000 €', etc.). Per cada nivell, es mostra el coeficient estimat en comparació amb el nivell de referència. El nivell de referència serà aquella categoria que no apareix en la sortida del model.

A més a més, es proporciona informació sobre els graus de llibertat del model i les mètriques d'ajust, com la desviació nul·la ('Null Deviance'), la desviació residual ('Residual Deviance') i el criteri d'informació d'Akaike (AIC). Aquestes mètriques són útils per avaluar la qualitat de l'ajust del model.

```
Call:  glm(formula = Sobreq ~ PA1 + PA2 + PA3 + anyini_c + autonom_c +
    codi_a + guanys + numtreb, family = binomial(link = "logit"),
    data = dataTrain2)

Coefficients:
                          (Intercept)                                   PA1
                            -1.542824                             -0.356786
                                  PA2                                   PA3
                            -0.040543                             -0.737409
              anyini_cMenys d'un any              anyini_cFa més de 3 anys
                             0.128853                              0.674062
                   anyini_cFa 3 anys                     anyini_cFa 1 any
                            -0.161818                              0.016379
                         autonom_cSí                codi_aArts i Humanitats
                            -0.203316                              0.527348
        codi_aCiències de la salut  codi_aCiències socials i jurídiques
                            -0.614811                             -0.118371
                    codi_aCiències          guanysEntre 9.000 i 12.000 €
                             0.302833                              0.647411
             guanysMenys de 9.000 €                            guanysNs/Nc
                             0.873245                              0.311336
      guanysEntre 18.001 i 24.000 €         guanysEntre 15.001 i 18.000 €
                             0.098043                              0.270239
     guanysEntre 40.000 i 50.000 €)         guanysEntre 12.001 i 15.000 €
                             0.012401                              0.517493
      guanysEntre 30.001 i 40.000 €                 guanysMés de 50.000 €
                            -0.026161                              0.196699
               numtrebEntre 51 i 100                      numtreb10 o menys
                            -0.043246                              0.117421
                   numtrebMés de 500                             numtrebNs/Nc
                             0.277516                             -0.009747
             numtrebEntre 251 i 500                numtrebEntre 101 i 250
                             0.089029                             -0.054462

Degrees of Freedom: 8120 Total (i.e. Null);  8093 Residual
Null Deviance:      9462
Residual Deviance: 8124         AIC: 8180
```

Figura 6.18: Model Regressió Logística. Font: Elaboració pròpia.

En la Figura 6.19 es pot veure la representació visual del model de regressió logística i les dades d'entrenament utilitzades per ajustar el model. En l'eix X es representen la variable "PA1" i en l'eix Y es representa la variable "PA2". Els punts de color blau corresponen a les dades d'entrenament on la variable de resposta "Sobreq" és "Normal", mentre que els punts de color vermell corresponen a les dades on la variable de resposta és diferent de "Normal". A més dels punts de dades, també es representen els punts predits pel model fent ús de la funció 'predict()'. Aquests punts predits es mostren amb colors més clars: els punts de color clar blau corresponen a les prediccions on la probabilitat predita pel model és major que 0.5 (indicant la classe "Normal"), i els punts de color clar coral corresponen a les prediccions on la probabilitat predita és menor o igual a 0.5 (indicant la classe diferent de "Normal").

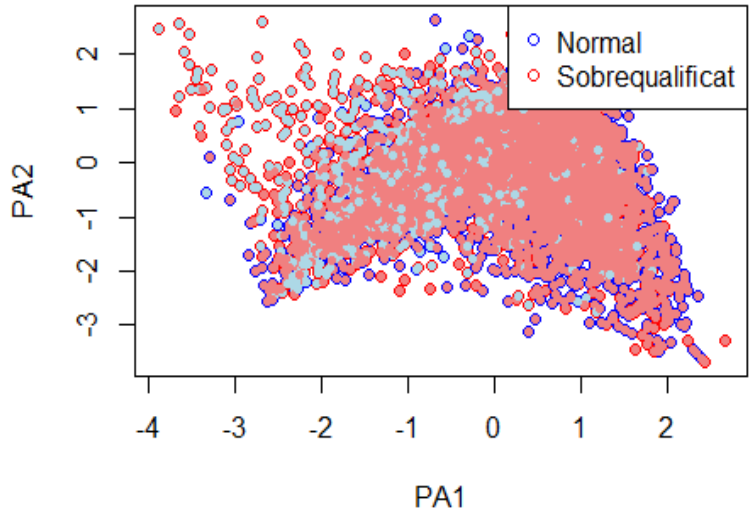


Figura 6.19: Representació gràfica variable resposta i PA1 i PA2. Font: Elaboració pròpia.

S'ha realitzat una matriu de confusió per tal de tenir una idea de com és de bo el model, és a dir, com seran de bones les prediccions. En aquest cas, la matriu de confusió, Figura 6.20, mostra que per la classe "Normal", el model va predir correctament 5635 casos com a "Normal" (veritables positius) i 1526 casos com "Sobrequalificat" (falsos positius). Per la classe "Sobrequalificat", el model va predir correctament 661 casos com "Sobrequalificat" (veritables negatius) i 299 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction        Normal Sobrequalificat
  Normal            5635            1526
  Sobrequalificat    299             661
```

Figura 6.20: Matriu de confusió dadesTrain2. Regressió Logística. Font: Elaboració pròpia.

A partir de la matriu de confusió, l'exactitud (*accuracy*) del model en el conjunt d'entrenament és del 0.7753, cosa que significa que aproximadament el 77.53% de les instàncies es classifiquen correctament segons el model. Això és un indicador de la precisió del model en la classificació de les dades d'entrenament.

El càlcul de la precisió ha obtingut un valor de 0.7869, la qual cosa indica que s'han classificat correctament 78.69% en la classificació dels casos positius en el conjunt d'entrenament.

El valor obtingut de la sensibilitat és del 94.96% en la classificació dels casos positius en les dades d'entrenament. Això significa que el model és capaç d'identificar la majoria dels casos sobrequalificats de manera precisa.

Finalment, s'ha calculat l'especificitat, on el valor aconseguit és 0.3022. Indica que el model té una especificitat del 30.22% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, la corba ROC i el valor de l'AUC indiquen la capacitat del model per a distingir entre les classes objectiu (sobrequalificat o no) en les dades d'entrenament. Un valor d'AUC de 0.7389 suggereix que el model té una capacitat moderada per a realitzar aquesta distinció.

En la Figura 6.21 es pot veure representada la corba de ROC on en l'eix X es representa la taxa de falsos positius, que és la proporció de casos negatius incorrectament classificats com a positius. I en l'eix Y, es troba la taxa de veritables positius o sensibilitat, que és la proporció de casos positius correctament classificats com a positius. S'observa que la corba es posiciona totalment per sobre de la línia clara d'aleatorietat, això indica gràficament una bona predicció.

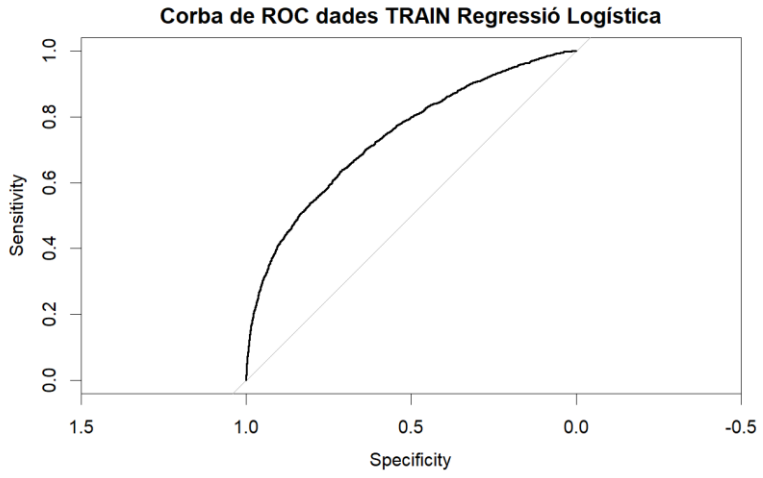


Figura 6.21: Corba de ROC model Regressió Logística. Font: Elaboració pròpia.

És important tenir en compte que aquestes mètriques són utilitzades per a avaluar el rendiment del model en les dades d'entrenament. Per a obtenir una avaluació més completa, es recomana també avaluar el model en dades de prova o fer ús de tècniques de validació creuada.

Per tant, la matriu de confusió per les dades de prova, Figura 6.22, mostra que per la classe “Normal”, el model va predir correctament 5635 casos com a "Normal" (veritables positius) i 1526 casos com "Sobrequalificat" (falsos positius). Per la classe “Sobrequalificat”, el model va predir correctament 661 casos com "Sobrequalificat" (veritables negatius) i 299 casos com a "Normal" (falsos negatius).

```
Confusion Matrix and Statistics

                Reference
Prediction       Normal Sobrequalificat
  Normal           2822             761
  Sobrequalificat   158             320
```

Figura 6.22: Matriu de confusió dadesTest2. Regressió Logística. Font: Elaboració pròpia.

L'exactitud (*accuracy*) del model en les dades de prova és de 0.7737, la qual cosa implica que el model té una precisió del 77.37% en la classificació de les instàncies del conjunt de prova. Això significa que aproximadament el 77.37% de les instàncies es van classificar correctament segons el model.

La precisió és de 0.7876, la qual cosa indica que s’han classificat correctament 78.76% en la classificació dels casos positius en el conjunt d'entrenament.

El valor de la sensibilitat per a les dades de prova és de 0.9470. Això indica que el model ha identificat correctament el 94.70% dels casos positius en el conjunt de prova. En altres paraules, el model ha aconseguit capturar la majoria dels casos de sobrequalificat de manera precisa.

El valor obtingut de l’especificitat 0.2960 indica que el model té una especificitat del 29.60% en la classificació dels casos negatius en el conjunt d'entrenament.

Per altra banda, es va assolir un valor de 0.7439 en l'àrea sota la corba ROC. Això indica que el model té capacitat moderada per a classificar correctament les instàncies en les dues classes,

"Normal" i "Sobrequalificat". La representació de la corba és similar a les dades d'entrenament i es troba en l'annex (**ANNEX**).

	Dades entrenament	Dades prova
Exactitud	0.7753	0.7737
Precisió	0.7869	0.7876
Sensibilitat	0.9496	0.9470
Especificitat	0.3022	0.2960
Àrea Corba ROC	0.7389	0.7439

Taula 6.4: Mètriques d'avaluació del model de classificació Regressió Logística. Font: Elaboració pròpia.

7. SELECCIÓ DEL MILLOR MODEL DE CLASSIFICACIÓ

En aquesta comparació de mètodes de classificació, s'analitzaran els quatre algorismes utilitzats anteriorment: Arbres de Decisió, Màquines de Vectors de Suport (SVM), *Naive Bayes* (NB) i Regressió Logística. L'objectiu és avaluar el seu rendiment en la tasca de classificació utilitzant les tres mètriques clau: l'exactitud (*accuracy*), la precisió, la sensibilitat, l'especificitat i l'àrea sota la corba ROC (AUC).

Cal recordar que, l'exactitud (*accuracy*) ens permet avaluar que tan bé els algorismes classifiquen correctament les instàncies en general. Com més alt sigui el valor d'exactitud, millor serà el rendiment de l'algorisme en la classificació global de les instàncies.

La precisió es refereix a la proporció d'instàncies classificades com a positives que són realment positives. És a dir, és una mesura de quantes de les instàncies classificades com a positives són realment positives. Una alta precisió indica que el model té una baixa taxa de falsos positius.

El *recall*, també conegut com a la sensibilitat, ens dóna una mesura de la capacitat dels algorismes per a identificar correctament els casos positius. Un alt valor de sensibilitat indica que l'algorisme pot detectar la majoria dels casos positius, la qual cosa és especialment rellevant en problemes en els quals la identificació de casos positius és crucial com en l'àmbit sanitari.

L'especificitat és la proporció d'instàncies negatives que el model ha identificat correctament. És una mesura de quantes de les instàncies realment negatives han estat correctament identificades pel model.

Finalment, l'AUC de la corba ROC proporciona una mesura de la capacitat discriminativa de cada algorisme. Si el valor de l'AUC és major a 0.5, significa que el model prediu de forma representativa a les observacions. Si és menor a 0.5, la predicció és insignificant, ja que, és similar a una selecció aleatòria. Un valor d'AUC pròxim a 1 indica un excel·lent rendiment en la classificació, cosa que significa que l'algorisme pot distingir amb precisió entre les classes objectives. Comparant els valors d'AUC, podrem determinar quin algorisme té un millor rendiment en aquest aspecte.

Mitjançant la comparació d'aquests valors per a cada algorisme, es podrà determinar quin d'ells té el millor rendiment en termes de capacitat discriminativa, classificació global i

detecció de casos positius. Això ajudarà a seleccionar l'algorisme més adequat per al problema de classificació en qüestió.

És important considerar tant les mètriques calculades en les dades d'entrenament com les mètriques obtingudes en les dades de prova.

Les mètriques calculades en les dades d'entrenament proporcionen una avaluació del rendiment del model en les dades utilitzades per al seu ajust. Això ens dóna una idea de com s'està ajustant el model a les dades d'entrenament i quin tan bé pot realitzar la classificació en aquest conjunt específic. No obstant això, el rendiment en les dades d'entrenament pot no ser un indicador precís del rendiment real del model en dades noves i no vists.

D'altra banda, les mètriques calculades en les dades de prova brinden una avaluació més de confiança del rendiment del model en dades no usades durant l'entrenament. Aquestes mètriques ens donen una idea de com es generalitza el model i com pot classificar noves dades. És important avaluar el rendiment en les dades de prova per a tenir una millor comprensió de la capacitat de generalització del model i la seva capacitat per a fer prediccions precises en situacions reals.

En aquest cas, s'han calculat les cinc mètriques amb la variable resposta "Sobreq", si hi ha sobrequalificació o no en l'estudiant. I amb els mateixos conjunts de dades d'entrenament i de prova en els quatre algoritmes aplicats.

En aquesta primera part, s'analitzen les mètriques en els diferents mètodes pel cas de les dades d'entrenament (*Train*).

Es pot observar en la Taula 7.1, termes de l'exactitud (a*ccuracy*), la Regressió Logística presenta el valor més alt (0.7753), el que indica que classifica correctament al voltant del 77.53% de les instàncies en el conjunt d'entrenament.

Pel que fa a la precisió, el model de Naive Bayes mostra el millor acompliment amb un valor de 0.7945. Això significa que de les instàncies classificades com a positives per aquest model, al voltant del 79.45% són realment positives.

Si es mira la sensibilitat, és a dir, la capacitat d'identificar correctament els casos positius, el model de SVM obté el millor resultat amb un valor de 0.9816. Això indica que el 98.16% dels casos positius ha estat identificat correctament per aquest model.

No obstant això, en analitzar l'especificitat, que mesura la capacitat d'identificar correctament els casos negatius, cap dels models mostra un rendiment destacat. El model de *Naive Bayes* té l'especificitat més alta amb un valor de 0.3544.
Finalment, en avaluar l'àrea sota la corba ROC (AUC), que avalua la capacitat general del model per a distingir entre classes, novament el model de Regressió Logística presenta el millor resultat amb un valor de 0.7389.

DADES TRAIN	Arbres de decisió	Naive Bayes	SVM	Regressió Logística
Exactitud	0.7694	0.7675	0.7681	0.7753
Precisió	0.7793	0.7945	0.7665	0.7869
Sensibilitat	0.9547	0.9198	0.9816	0.9496
Especificitat	0.2666	0.3544	0.1888	0.3022
Àrea Corba ROC	0.6460	0.7304	0.5852	0.7389

Taula 7.1: Mètriques d'avaluació del model de classificació dades TRAIN. *Font: Elaboració pròpia.*

En general, l'elecció del millor model depèn de la mètrica que es consideri més rellevant en el context del problema en particular. En base aquests resultats, es pot concloure que el model de Regressió Logística mostra un bon acompliment en termes d'exactitud i d'àrea sota la corba ROC en les dades d'entrenament. Tot i que no tingui el valor més alt de precisió entre els models avaluats, continua sent relativament bona, el que significa que té una proporció considerable de prediccions positives correctes. Encara que el model de Regressió Logística no té l'especificitat més alta, la seva especificitat (0.3022) és raonablement bona, la qual cosa implica que pot identificar correctament una proporció significativa de casos negatius reals. Per tant, es pot considerar com el millor model en aquest conjunt de dades específic.

No obstant això, és important tenir en compte que aquestes avaluacions es van realitzar en les dades d'entrenament. Per a obtenir una avaluació més completa i de confiança, es recomana avaluar els models en dades de prova. Això permetrà determinar el rendiment i la generalització dels models en situacions reals i amb dades no vistes prèviament.

Pel cas de les dades de prova, en la Taula 7.2 es pot observar que en termes de l'exactitud (*accuracy*), la Regressió Logística també té el valor més alt (0.7737), la qual cosa indica que classifica correctament al voltant del 77.37% de les instàncies en el conjunt de prova.

Respecte a la precisió, el model de *Naive Bayes* obté el millor acompliment amb un valor de 0.7950. Això significa que de les instàncies classificades com a positives per aquest model, aproximadament el 79.50% són realment positives.

En termes de la sensibilitat, el model SVM mostra el millor resultat amb un valor de 0.9812. Això indica que el 98.12% dels casos positius ha estat identificat correctament per aquest model en les dades de prova.

Si es mira l'especificitat, novament cap dels models mostra un rendiment destacat. El model de *Naive Bayes* té l'especificitat més alta amb un valor de 0.3469.

Finalment, en avaluar l'àrea sota la corba ROC (AUC), el model de Regressió Logística presenta el millor resultat amb un valor de 0.7439.

En base aquests resultats, es pot concloure que el model de Regressió Logística també mostra un bon acompliment en termes d'exactitud i d'àrea sota la corba ROC en les dades de prova.

DADES TEST	**Arbres de decisió**	**Naive Bayes**	**SVM**	**Regressió Logística**
Exactitud	0.7725	0.7666	0.7683	0.7737
Precisió	0.7738	0.7950	0.7677	0.7876
Sensibilitat	0.9748	0.9188	0.9812	0.9470
Especificitat	0.2146	0.3469	0.1813	0.2960
Àrea Corba ROC	0.6478	0.7345	0.5813	0.7439

Taula 7.2: Mètriques d'avaluació del model de classificació dades TEST. *Font: Elaboració pròpia.*

Comparant els resultats en les dades de prova amb els obtinguts en les dades d'entrenament, es pot veure que els valors de les mètriques en general són similars per a cada model. Això indica que els models tenen un rendiment similar en les dades de prova en comparació amb les dades d'entrenament.

En termes generals, la Regressió Logística sembla tenir un bon rendiment en totes les mètriques avaluades tant en les dades d'entrenament com en els de prova. Per tant, tenint en compte l'exactitud com a mètrica més rellevant i que les altres tinguin valors raonablement bons, es pot dir que la Regressió Logística és el model que millor classifica aquestes dades.

El segon millor model pot variar depenent de la mètrica específica que es consideri.

Seguint el mateix criteri anterior, on l'exactitud es considera la mètrica més rellevant i es busca que els altres valors siguin considerablement bons, es va determinar que el segon millor model, tant en el conjunt de dades d'entrenament com en el conjunt de dades de prova, és l'Arbre de Decisió, amb una precisió de 0.7694 i 0.7725 respectivament.

En resum, en termes d'exactitud, el model de Regressió Logística és el millor model seguit de l'Arbre de Decisió.

7.1 Perfil de treballador sobrequalificat.

Centrant-se en el millor model de classificació escollit, el model de Regressió Logística. A partir d'aquest model es pot obtenir un perfil de treballador sobrequalificat.

En primer lloc, s'ajusta el model de classificació logística amb la funció 'glm()'. L'objectiu del model és predir la variable de resposta "Sobreq" basant-se en les variables predictores "PA1", "PA2", "PA3", "anyini_c", "autonom_c", "codi_a", "guanys" i "numtreb". Excepte la variable 'sitact', ja que només conté una categoria. Com a base de dades s'ha agafat tot el conjunt de dades inicial. El model utilitza una funció d'enllaç logit i assumeix una distribució binomial per a la variable de resposta.

El resum del model (Figura 7.1) proporciona els coeficients estimats per cada una de les variables predictores. Cada coeficient indica la contribució relativa de la variable predictora al logaritme de la raó d'odds (probabilitats) de la variable resposta. El log-odds es defineix com el logaritme de la raó de les probabilitats entre l'esdeveniment d'interès (en aquest cas, tenir sobrequalificació) i l'esdeveniment complementari (no tenir sobrequalificació). Els coeficients positius indiquen un augment en el log-odds i, per tant, un major risc de pertànyer a la categoria "Sobrequalificat", mentre que els coeficients negatius indiquen una disminució en el log-odds i, doncs, un menor risc.

Els valors 'Std.Error' representen els errors estàndard corresponents als coeficients estimats. Com menor sigui el valor de l'error estàndard, major serà la precisió de l'estimació del coeficient.

Els valors 'z value' són la relació entre el coeficient estimat i el seu error estàndard. S'utilitzen per a avaluar la significança estadística de cada coeficient. Valors absoluts més alts indiquen una major significança estadística.

La columna 'Pr(>|z|):' mostra el valor 'p' associat a cada coeficient. Indica la probabilitat d'obtenir un valor del coeficient almenys tan extrem com l'observat si la veritable relació entre la variable predictora i la resposta fos nul·la. Valors petits (per sota d'un cert llindar de significança, normalment 0.05) indiquen una significança estadística, la qual cosa implica que la variable predictora té un efecte significatiu sobre la variable resposta.

Els valors de la 'Null deviance' i 'Residual Deviance', representen la desviació nul·la i la desviació residual del model, respectivament. S'usen per a avaluar l'ajust del model en comparació amb un model nul. Una disminució en la desviació residual indica un millor ajust del model.

Per acabar, l'AIC (Akaike Information Criterion) és un criteri de selecció de models que penalitza la complexitat del model. Com menor sigui el valor d'AIC, millor serà l'ajust del model.

```
Call:
glm(formula = Sobreq ~ PA1 + PA2 + PA3 + anyini_c + autonom_c +
    codi_a + guanys + numtreb, family = binomial, data = bbdd2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3082  -0.7444  -0.5498   0.6813   2.6211

Coefficients:
                                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)                     -1.512850   0.098929 -15.292 < 0.0000000000000002 ***
PA1                             -0.368719   0.024047 -15.333 < 0.0000000000000002 ***
PA2                             -0.062388   0.024930  -2.503             0.012331 *
PA3                             -0.711624   0.026091 -27.274 < 0.0000000000000002 ***
anyini_cMenys d'un any           0.131831   0.068148   1.934             0.053055 .
anyini_cFa més de 3 anys         0.600183   0.075410   7.959  0.00000000000000173 ***
anyini_cFa 3 anys               -0.164476   0.090531  -1.817             0.069250 .
anyini_cFa 1 any                -0.023180   0.074974  -0.309             0.757191
autonom_cSí                     -0.238444   0.084509  -2.822             0.004780 **
codi_aArts i Humanitats          0.565502   0.085825   6.589  0.00000000004427582 ***
codi_aCiències de la salut      -0.673704   0.081317  -8.285 < 0.0000000000000002 ***
codi_aCiències socials i jurídiques -0.116679   0.061463  -1.898             0.057648 .
codi_aCiències                   0.315703   0.097755   3.230             0.001240 **
guanysEntre 9.000 i 12.000 €     0.674730   0.097282   6.936  0.00000000000403888 ***
guanysMenys de 9.000 €           0.884650   0.100474   8.805 < 0.0000000000000002 ***
guanysNs/Nc                      0.301972   0.091612   3.296             0.000980 ***
guanysEntre 18.001 i 24.000 €    0.104309   0.082434   1.265             0.205740
guanysEntre 15.001 i 18.000 €    0.273666   0.096643   2.832             0.004630 **
guanysEntre 40.001 i 50.000 €)  -0.057410   0.151967  -0.378             0.705595
guanysEntre 12.001 i 15.000 €    0.437254   0.092820   4.711  0.00000246798194184 ***
guanysEntre 30.001 i 40.000 €   -0.061805   0.097848  -0.632             0.527623
guanysMés de 50.000 €            0.014618   0.180102   0.081             0.935311
numtrebEntre 51 i 100           -0.001666   0.086822  -0.019             0.984694
numtreb10 o menys                0.144052   0.073424   1.962             0.049772 *
numtrebMés de 500                0.258687   0.066768   3.874             0.000107 ***
numtrebNs/Nc                    -0.072764   0.103533  -0.703             0.482173
numtrebEntre 251 i 500           0.127221   0.105038   1.211             0.225824
numtrebEntre 101 i 250          -0.062810   0.091845  -0.684             0.494059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14168  on 12181  degrees of freedom
Residual deviance: 12177  on 12154  degrees of freedom
AIC: 12233

Number of Fisher Scoring iterations: 4
```

Figura 7.1: Sortida model Regressió Logística amb tot el conjunt de dades. Font: Elaboració pròpia.

A partir dels coeficients estimats del model, és a dir els log-odds, es poden convertir en odds utilitzant la funció exp() (Figura 7.2). En aquest cas, valors superiors a 1, indiquen una associació positiva entre la variable predictora i l'esdeveniment de sobrequalificació, mentre que valors inferiors a 1, implica una associació negativa entre la variable predictora i l'esdeveniment d'interès. A més a més, es pot calcular les probabilitats de què succeeixi

l'esdeveniment d'interès, és a dir, pertànyer a la categoria "Sobrequalificat", donat el valor de la variable predictora.

Per exemple, la variable “codi_aArts i Humanitats”, el coeficient log-odds és 0.565502, això significa que, en comparació amb la categoria de referència “codi_aEng. i Arquitectura”, pertànyer a la categoria “codi_aArts i Humanitats” s’associa a un increment esperat en el log-odds de ser classificat com “Sobrequalificat” en 0.565502 unitats.

Per altra banda, l’odd de la variable “codi_aArts i Humanitats” és 1.7603307. Això indica que la raó de probabilitats (odds) de sobrequalificació per a aquells amb una formació en "codi_aArts i Humanitats" és aproximadament 1.76 vegades la raó de probabilitats per a aquells amb una formació en "codi_aEng i Arquitectura" (categoria de referència). En altres paraules, els individus amb formació en "codi_aArts i Humanitats" tenen majors probabilitats d'experimentar sobrequalificació en comparació amb aquells amb formació en "codi_aEng i Arquitectura".

Per calcular l’odd ràtio de comparació entre les categories "codi_aArts i Humanitats" i "codi_aCiències de la salut" amb relació a la variable predictora, segons els coeficients del model de regressió logística:

Coeficient de "codi_aArts i Humanitats": 0.565502
Coeficient de "codi_aCiències de la salut": -0.673704

Diferència de coeficients: 0.565502 - (-0.673704) = 1.239206

Odd Ràtio de comparació entre "codi_aArts i Humanitats" i "codi_aCiències de la salut":
Odd Ràtio = exp(1.239206) = 3.456467

Això significa que, en comparació amb la categoria "codi_aCiències de la salut", la categoria "codi_aArts i Humanitats" té un odd ràtio de 3.456467. En termes d'interpretació, això implica que la probabilitat de pertànyer a la categoria “Sobrequalificat” és aproximadament 3.46 vegades major per a aquells que pertanyen a la categoria "codi_*aArts i Humanitats" en comparació amb aquells que pertanyen a la categoria "codi_*aCiències de la salut".

Utilitzant la transformació logística inversa, la probabilitat estimada de sobrequalificació per a aquells amb formació en "codi_aArts i Humanitats" és de 0.6377246.

	Coeficient	Odds	Probabilitat
(Intercept)	-1.512849614	0.2202814	0.1805169
PA1	-0.368718884	0.6916198	0.4088506
PA2	-0.062387700	0.9395186	0.4844081
PA3	-0.711624169	0.4908463	0.3292401
anyini_cMenys d'un any	0.131831155	1.1409157	0.5329101
anyini_cFa més de 3 anys	0.600183103	1.8224525	0.6456982
anyini_cFa 3 anys	-0.164476012	0.8483381	0.4589734
anyini_cFa 1 any	-0.023179847	0.9770867	0.4942053
autonom_cSí	-0.238444296	0.7878526	0.4406698
codi_aArts i Humanitats	0.565501680	1.7603307	0.6377246
codi_aCiències de la salut	-0.673704055	0.5098167	0.3376679
codi_aCiències socials i jurídiques	-0.116679426	0.8898704	0.4708632
codi_aCiències	0.315703180	1.3712232	0.5782767
guanysEntre 9.000 i 12.000 €	0.674729537	1.9635018	0.6625614
guanysMenys de 9.000 €	0.884650392	2.4221374	0.7077850
guanysNs/Nc	0.301971815	1.3525231	0.5749245
guanysEntre 18.001 i 24.000 €	0.104309310	1.1099437	0.5260537
guanysEntre 15.001 i 18.000 €	0.273665801	1.3147753	0.5679926
guanysEntre 40.001 i 50.000 €)	-0.057409662	0.9442072	0.4856515
guanysEntre 12.001 i 15.000 €	0.437254310	1.5484498	0.6076046
guanysEntre 30.001 i 40.000 €	-0.061804635	0.9400665	0.4845538
guanysMés de 50.000 €	0.014617878	1.0147252	0.5036544
numtrebEntre 51 i 100	-0.001665649	0.9983357	0.4995836
numtreb10 o menys	0.144051560	1.1549437	0.5359507
numtrebMés de 500	0.258686756	1.2952280	0.5643134
numtrebNs/Nc	-0.072764173	0.9298201	0.4818170
numtrebEntre 251 i 500	0.127220580	1.1356675	0.5317623
numtrebEntre 101 i 250	-0.062809930	0.9391220	0.4843027

Figura 7.2: Càlcul d'odds i probabilitats del model de Regressió Logística. Font: Elaboració pròpia.

Per determinar quin perfil té més probabilitat de ser classificat com a "Sobrequalificat" segons el model anterior de regressió logística, és necessari analitzar els coeficients estimats i la seva influència en la variable resposta. Fixant amb la Figura 7.1 es pot identificar les variables predictores que tenen una associació positiva significativa amb la probabilitat de ser classificat com "Sobrequalificat". Aquestes variables són:

- 'anyini_cFa més de 3 anys': Té un coeficient estimat de 0.600183, la qual cosa suggereix que pertànyer a la categoria "anyini_cFa més de 3 anys" s'associa amb un augment en la probabilitat de ser classificat com "Sobrequalificat".
- 'codi_aArts i Humanitats': Té un coeficient estimat de 0.565502, la qual cosa indica que pertànyer a la categoria "codi_aArts i Humanitats" es relaciona amb un augment en la probabilitat de ser classificat com "Sobrequalificat".
- 'guanysEntre 9.000 i 12.000 €': Té un coeficient estimat de 0.674730, la qual cosa indica que tenir un nivell d'ingressos "guanysEntre 9.000 i 12.000 €" es relaciona amb un augment en la probabilitat de ser classificat com "Sobrequalificat".

- ‘guanysMenys de 9.000 €’: Té un coeficient estimat de 0.884650, la qual cosa suggereix que tenir un nivell d'ingressos "guanysMenys de 9.000 €" s'associa amb un augment en la probabilitat de ser classificat com "Sobrequalificat".
- ‘numtrebMés de 500’: Té un coeficient estimat de 0.258687, la qual cosa indica que tenir un nombre de treballadors "numtrebMés de 500" es relaciona amb un augment en la probabilitat de ser classificat com "Sobrequalificat".

Per tant, en base aquests resultats obtinguts, es pot concloure que un perfil amb les següents característiques té més probabilitat de ser classificat com “Sobrequalificat” segons el model de regressió logística:

- Experiència laboral de més de 3 anys (categoria ‘anyini_cFa més de 3 anys’).
- Branca de titulació en el camp de les Arts i Humanitats (categoria ‘codi_aArts i Humanitats’).
- Nivell d’ingressos baix (categories ‘guanysMenys de 9.000 €’, i ‘guanysEntre 9.000 i 12.000 €’).
- Pertànyer a empreses amb un nombre de treballadors superior a 500 (categoria ‘numtrebMés de 500’).

En resum, un treballador amb aquestes categories té una major probabilitat de ser classificat com “Sobrequalificat” segons el model de regressió logística.

8. CONCLUSIONS

Amb aquest treball s'ha abordat i aprofundit en el problema de la sobrequalificació utilitzant algoritmes de classificació, amb l'objectiu de seleccionar el millor model per a aquestes dades.

En primer lloc, s'ha analitzat i entès el fenomen de la sobrequalificació, identificant els factors que influeixen en aquesta situació. La sobrequalificació es produeix quan un treballador posseeix un nivell d'educació i habilitats que superen els requisits d'un lloc de treball específic. Els factors que s'han identificat com a influents en la sobrequalificació són l'experiència laboral, la branca de titulació, els guanys anuals bruts i el nombre de treballadors en l'empresa.

Amb aquesta conclusió dels factors influents es pot afirmar la primera hipòtesi, de què un dels factors que influeix en la sobrequalificació és la branca de titulació dels treballadors.

A continuació, s'han aplicat i avaluat quatre tècniques de classificació per determinar quin model és el més efectiu en la classificació dels titulats sobrequalificats. Els resultats obtinguts han estat satisfactoris i han mostrat un rendiment similar entre els models. Tot i això, es pot afirmar que el model de regressió logística ha demostrat un bon rendiment, especialment en l'accuracy i l'àrea sota la corba de ROC, i totes les altres mètriques d'avaluació, tant en les dades d'entrenament com en les dades de prova. Per tant, es pot concloure que el model de regressió logística és una opció adequada per classificar aquestes dades.

Un dels objectius del treball era identificar el perfil de treballador sobrequalificat utilitzant el millor model de classificació escollit. Com s'ha mencionat anteriorment, el model de regressió logística ha estat considerat el millor per a aquestes dades. A partir de l'aplicació d'aquest model a tot el conjunt de dades, s'han obtingut els següents resultats: un treballador amb més de 3 anys d'experiència laboral, una titulació en el camp de les Arts i Humanitats, uns ingressos baixos o mitjans, i que treballa en empreses amb més de 500 treballadors té una major probabilitat de ser classificat com a "sobrequalificat" segons el model de regressió logística.

A partir d'aquestes conclusions, es pot rebutjar la segona hipòtesi i afirmar la tercera hipòtesi. En primer lloc, no hi ha diferències significatives entre els diferents models de classificació, tots presenten un bon rendiment en les mètriques avaluades. En segon lloc, existeix una relació significativa entre la branca de titulació dels treballadors i la sobrequalificació, sent la branca d'Arts i Humanitats la que té una major probabilitat de ser sobrequalificada.

En resum, aquest treball ha permès abordar la qüestió de la sobrequalificació des d'una perspectiva àmplia i profunda. S'ha identificat una sèrie de factors que influeixen en aquest fenomen, com ara l'experiència laboral, la branca de titulació, els guanys anuals bruts i el nombre de treballadors de les empreses. A través de l'aplicació de tècniques de classificació, s'ha determinat que el model de regressió logística és eficaç per classificar els treballadors sobrequalificats. Les conclusions d'aquest estudi proporcionen una comprensió més profunda de la sobrequalificació i poden ser útils per prendre mesures per abordar aquest problema en els àmbits educatiu i laboral.

REFERÈNCIES

¿Cuáles son los tipos de algoritmos del machine learning? (April / 2019). *APD España*. Recollit de https://www.apd.es/algoritmos-del-machine-learning/

¿Qué es Machine Learning y qué aplicaciones tiene? (2018). *¿Qué es Machine Learning y qué aplicaciones tiene?* Retrieved June 7, 2023, from https://www.ibertech.org/que-es-machine-learning-y-que-aplicaciones-tiene-en-nuestro-dia-a-dia-2/

Arias Montoya, R., Jairo Santa Chávez, J., & Veloza Mora, J. (2013, June). Aplicación del aprendizaje automático con árboles de decisión en el diagnóstico médico. *Cultura del cuidado*. doi:10.18041/1794-5232/cultrua.2013v10n1.2102

Barrios Arce, J. (2019, July). La matriz de confusión y sus métricas – Inteligencia Artificial –. *La matriz de confusión y sus métricas – Inteligencia Artificial –*. Retrieved June 11, 2023, from https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

BBVA. (2019, November). Te contamos qué es el 'machine learning' y cómo funciona. *Te contamos qué es el 'machine learning' y cómo funciona*. Retrieved June 7, 2023, from https://www.bbva.com/es/innovacion/machine-learning-que-es-y-como-funciona/

Corrales Galán, R. (2021, July). La educación como señal: el modelo de Spence y su respaldo empírico. Retrieved June 19, 2023, from https://gredos.usal.es/handle/10366/147188

Departamento de Matemática Aplicada. (2021). Introducción al Aprendizaje Automático. Recollit de https://dcain.etsin.upm.es/~carlos/bookAA/introAA.html

España suma 11 años seguidos como país de la UE con más tasa de trabajadores sobrecualificados. (2023, April). *España suma 11 años seguidos como país de la UE con más tasa de trabajadores sobrecualificados*. Retrieved June 7, 2023, from https://www.lavanguardia.com/economia/20230430/8933009/espana-suma-11-anos-seguidos-pais-ue-mas-tasa-trabajadores-sobrecualificados.html

García Montalvo, J. (2001). *Formación y empleo de los graduados de enseñanza superior en España y en Europa.* Retrieved June 10, 2023, from https://dialnet.unirioja.es/servlet/libro?codigo=72378

GARCÍA MONTALVO, J. (2009). LA INSERCIÓN LABORAL DE LOS UNIVERISTARIOS Y EL FENÓMENO DE LA SOBRECUALIFICACIÓN EN ESPAÑA. *Funcas*. Recollit de https://www.funcas.es/wp-content/uploads/Migracion/Articulos/FUNCAS_PEE/119art12.pdf

García Montalvo, J., & Peiró, J. M. (2009). *ANÁLISIS DE LA SOBRECUALIFICACIÓN Y LA FLEXIBILIDAD LABORAL.* (I. Instituto Valenciano de Investigaciones Económicas, Ed.) Fundación Bancaja. Recollit de http://jgmontalvo.com/wp-content/uploads/2018/12/OBSERVATORIO2008.pdf

García Montalvo, J., Peiró, J. M., & Soro Bonmatí, A. (2003). *Observatorio de la inserción laboral de los jóvenes: 1996-2002.* Instituto Valenciano de Investigaciones Económicas (IVIE). Retrieved June 10, 2023, from https://dialnet.unirioja.es/servlet/libro?codigo=757748

García Montalvo, J., Soro Bonmatí, A., & Peiró Silla, J. M. (2006). *Los jóvenes y el mercado del trabajo en la España urbana: resultados del Observatorio de Inserción Laboral 2005.* Retrieved June 10, 2023, from https://dialnet.unirioja.es/servlet/libro?codigo=268724

Gavilá, S. (2014, May). La sobrecualificación: implicaciones para los empleados y para las empresas | InfoJobs. *La sobrecualificación: implicaciones para los empleados y para las empresas | InfoJobs*. Retrieved June 7, 2023, from https://recursos-humanos.infojobs.net/empleados-sobrecualificados

González, A. (sense data). Conceptos básicos de Machine Learning – Cleverdata. Recollit de https://cleverdata.io/conceptos-basicos-machine-learning/

Gonzalez, L. (2019, May). Curvas ROC y Área bajo la curva (AUC). *Curvas ROC y Área bajo la curva (AUC)*. Retrieved June 19, 2023, from https://aprendeia.com/curvas-roc-y-area-bajo-la-curva-auc-machine-learning/

Gonzalez, L. (September / 2019). Naive Bayes – Teoría. *Aprende IA*. Recollit de https://aprendeia.com/algoritmo-naive-bayes-machine-learning/

González, X., & Miles, D. (2021). La transición de la universidad al trabajo y el fenómeno de la sobrecualificación en {España}. Recollit de https://www.funcas.es/articulos/la-transicion-de-la-universidad-al-trabajo-y-el-fenomeno-de-la-sobrecualificacion-en-espana/

IBERDROLA. (n.d.). Descubre los principales beneficios del Machine Learning. *Descubre los principales beneficios del Machine Learning*. Retrieved June 7, 2023, from https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

Julio de Manuel. (2022, July). Nueva generación JASP: España, país con mayor nivel de jóvenes sobrecualificados de la UE. *Nueva generación JASP: España, país con mayor nivel de jóvenes sobrecualificados de la UE*. Retrieved June 7, 2023, from https://www.epe.es/es/activos/20220717/espana-tasa-sobrecualificacion-union-europea-jovenes-empleo-formacion-14085617

Martínez Martín, R. (sense data). Aproximaciones teóricas a los procesos de inserción laboral. Recollit de https://vlex.es/vid/aproximaciones-teoricas-procesos-insercion-116419

Montes Pineda, O., Garrido Yuste, R., & Gallo, M. (April / 2019). Sobre-cualificación o Falta de Oportunidades Laborales: Un análisissectorial en España. *Economics of Education*. Recollit de https://2019.economicsofeducation.com/user/pdfsesiones/121.pdf

MORENO BECERRA, J. (sense data). La educación como determinante del salario: capital humano versus credencialismo. Recollit de https://repositorio.uam.es/bitstream/handle/10486/5669/35448_6.pdf

Naive Bayes Classifier in Machine Learning - Javatpoint. (sense data). Recollit de https://www.javatpoint.com/machine-learning-naive-bayes-classifier

Ollé, J. (2020, November). ¿Cómo interpretar una máquina de vectores de soporte? *¿Cómo interpretar una máquina de vectores de soporte?* Retrieved June 19, 2023, from https://conceptosclaros.com/que-es-maquina-vectores-soporte/

Parra, F. (2019). *Estadística y Machine Learning con R.* Recollit de https://bookdown.org/content/2274/metodos-de-clasificacion.html

Pereira Toledo, A., López Cabrera, J., & Quintero Domínguez , L. (2017). Estudio experimental para la comparación del desempeño de Naïve Bayes con otros clasificadores bayesianos. *Revista Cubana de Ciencias Informáticas*. Recollit de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992017000400006

Precision, Recall, F1, Accuracy en clasificación - IArtificial.net. (2019, November). *Precision, Recall, F1, Accuracy en clasificación - IArtificial.net*. Retrieved June 7, 2023, from https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

Ray, S. (September / 2017). Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier. Recollit de https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Recuero de los Santos, P. (2021, December). Tipos de aprendizaje en Machine Learning: supervisado y no supervisado. *Tipos de aprendizaje en Machine Learning: supervisado y no supervisado*. Retrieved June 7, 2023, from https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/

Rodríguez, V. (2018). Decision trees / Árboles de decisión para clasificar en python. Recollit de https://vincentblog.xyz/posts/decision-trees-arboles-de-decision-para-clasificar-en-python

Roman, V. (2019, April). Algoritmos Naive Bayes: Fudamentos e Implementación. Retrieved June 7, 2023, from https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f

Sanchez, J. A. (2020). ¿Cómo aprenden las máquinas? Machine Learning y sus diferentes tipos. Recollit de https://datos.gob.es/es/blog/como-aprenden-las-maquinas-machine-learning-y-sus-diferentes-tipos#:~:text=Estos son: aprendizaje supervisado, aprendizaje,supervisado y aprendizaje por refuerzo

Sandoval, L. J. (2018). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS. *REVISTA TECNOLÓGICA N° 11.* Recollit de http://redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Saturino Martínez García, J. (2017). Sobrequalificació dels titulats universitaris i mobilitat social. *Papers: revista de sociologia*. Retrieved June 7, 2023, from https://raco.cat/index.php/Papers/article/view/v102-n1-martinez

Saturnino Martínez García, J. (2013). Sobrecualificación de los titulados universitarios y movilidad social. Recollit de https://www.educacionyfp.gob.es/inee/dam/jcr:4fc86178-c1de-474e-a843-1dc0f0620ae4/saturninomartinezpiaac2013vol2.pdf

Sicherman, N., & Galor, O. (1990). A Theory of Career Mobility. *Journal of Political Economy, 98*, 169–192. Retrieved June 10, 2023, from https://ideas.repec.org//a/ucp/jpolec/v98y1990i1p169-92.html

Soler, P. (2021, August). España, bronce en universitarios más sobrecualificados de la Unión Europea. *España, bronce en universitarios más sobrecualificados de la Unión Europea*. Retrieved June 7, 2023, from https://www.elconfidencial.com/economia/2021-08-22/espana-bronce-universitarios-sobrecualificados_3238586/

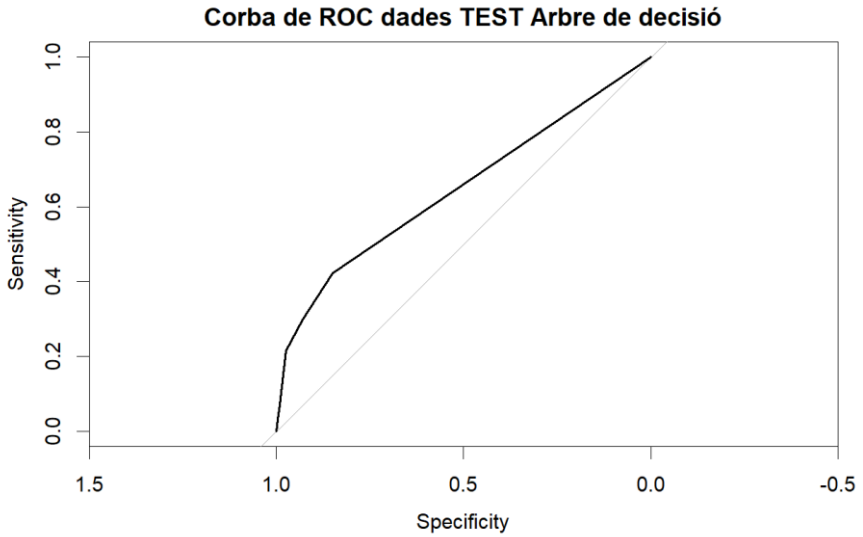
Support Vector Machine (SVM) Algorithm - Javatpoint. (sense data). Recollit de https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

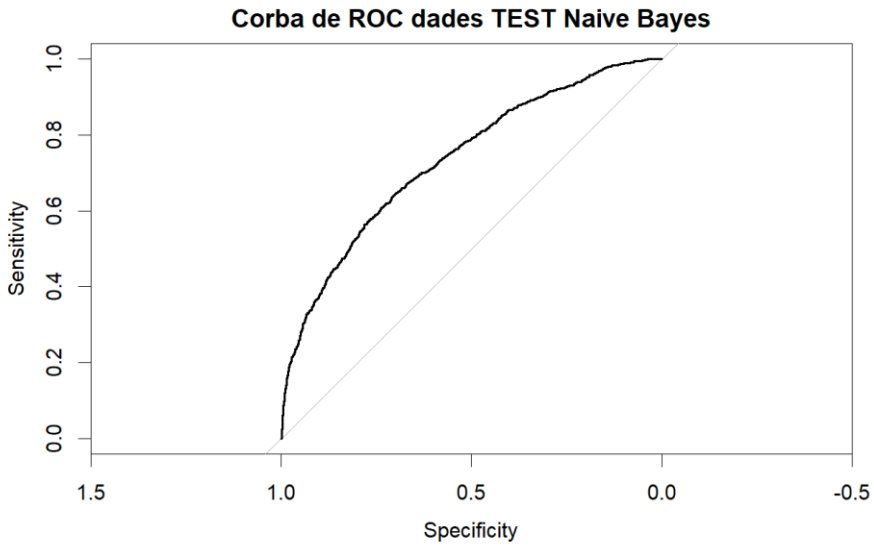
Support Vector Machine Algorithm. (2021). Recollit de https://www.geeksforgeeks.org/support-vector-machine-algorithm/

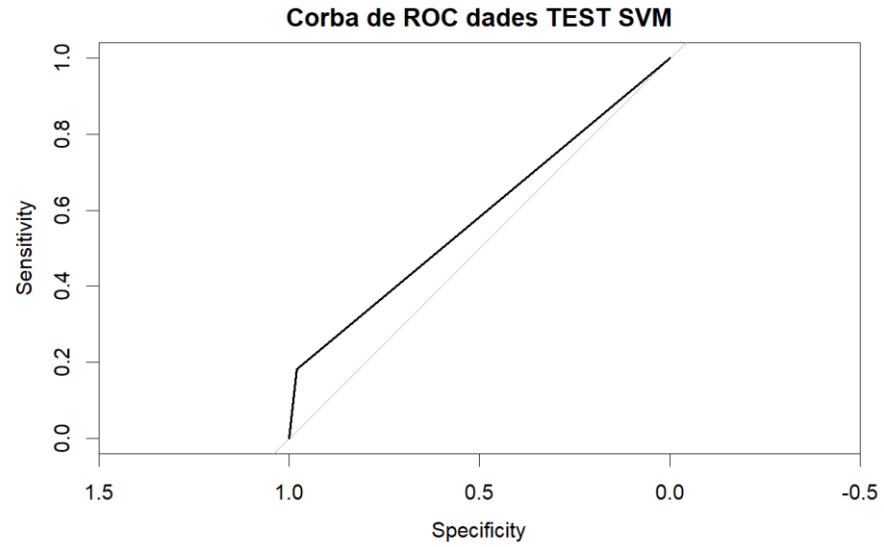
Turmo-Garuz, J., Bartual-Figueras, M.-T., & Sierra-Martinez, F.-J. (2019). Factors Associated with Overeducation Among Recent Graduates During Labour Market Integration: The Case of Catalonia (Spain). *Social Indicators Research*. doi:10.1007/s11205-019-02086-z

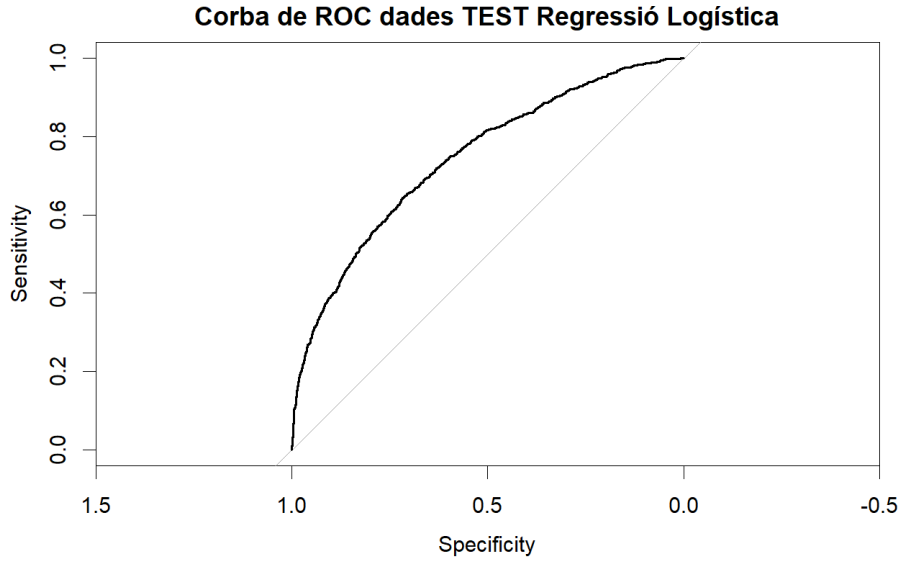
ANNEX

10.1 CORBES DE ROC DADES DE PROVA









10.2 CODI DE R

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
setwd("C:/Users/marin/Desktop/uni marina/4RT ANY/TFG/bbdd dep")
library(readr)
library(ggplot2)
library(descr)
library(kableExtra)
library(class)
library(lessR)
library(knitr)
#install.packages("corrplot")
library(corrplot)
library(Hmisc)
library(dplyr)
#install.packages("mice")
#install.packages("VIM")
library(mice)
library(VIM)
```
```

```
library(VIM)
library(cluster)
library(psych)
#library(psy)
library(rpart)
library(rpart.plot)
library(rattle)
library(caret)
library(MASS)
#install.packages("naivebayes")
library(naivebayes)
library(e1071)
library(rpart)
library(cluster)
library(pROC)
library(caret)
```

```{r}
bbdd <- read_delim("base_TFG.csv", delim = ";",
    escape_double = FALSE, locale = locale(decimal_mark = ","))
str(bbdd)
summary(bbdd)
```

```{r}
#---------Modificació de variables categòriques----------------
for (i in 16:57) {
  bbdd[[i]] <- gsub(",", ".", bbdd[[i]])
  bbdd[[i]] <- as.numeric(bbdd[[i]])
}

for(i in 1:15){
  levels <- as.character(unique(bbdd[[i]]))
  bbdd[[i]] <- factor(bbdd[[i]], levels = levels)
}

for(i in 58:60){
  levels <- as.character(unique(bbdd[[i]]))
  bbdd[[i]] <- factor(bbdd[[i]], levels = levels)
}

#Filtrar por los que trabajan ahora
bbdd <- subset(bbdd, sitact == "1")

#Eliminem les columnes que contenen masses NA's
bbdd <- subset(bbdd, select = -c(a_dific1_10, a_dific2_10, a_dific3_10, a_dific4_10,
a_dific5_10, a_dific6_10, a_dific7_10, a_dific8_10, a_dific9_10, pondera, filter_))

bbdd <- read_delim("base_TFG.csv", delim = ";",
    escape_double = FALSE, locale = locale(decimal_mark = ","))
str(bbdd)
summary(bbdd)

#---------Modificació de variables categòriques----------------
for (i in 16:57) {
  bbdd[[i]] <- gsub(",", ".", bbdd[[i]])
  bbdd[[i]] <- as.numeric(bbdd[[i]])
}

for(i in 1:15){
  levels <- as.character(unique(bbdd[[i]]))
  bbdd[[i]] <- factor(bbdd[[i]], levels = levels)
}
```

```r
for(i in 58:60){
  levels <- as.character(unique(bbdd[[i]]))
  bbdd[[i]] <- factor(bbdd[[i]], levels = levels)
}

#Filtrar por los que trabajan ahora
bbdd <- subset(bbdd, sitact == "1")

#Eliminem les columnes que contenen masses NA's
bbdd <- subset(bbdd, select = -c(a_dific1_10, a_dific2_10, a_dific3_10, a_dific4_10, a_dific5_10, a
_dific6_10, a_dific7_10, a_dific8_10, a_dific9_10, pondera, filter_))

#Eliminem els registres que contenen NA's
bbdd <- bbdd[complete.cases(bbdd), ]

bbdd$sitact <- factor(bbdd$sitact, labels = c("Treballo"))
bbdd$anyini_c <- factor(bbdd$anyini_c, labels = c("Fa 2 anys",
"Menys d'un any", "Fa més de 3 anys", "Fa 3 anys", "Fa 1 any"))
bbdd$autonom_c <- factor(bbdd$autonom_c, labels = c("No",
"Sí"))
bbdd$codi_a <- factor(bbdd$codi_a, labels = c("Eng. i Arquitectura",
"Arts i Humanitats", "Ciències de la salut", "Ciències socials i jurídiques", "Ciències"))
bbdd$cno_1digit <- factor(bbdd$cno_1digit, labels = c("Tècnics i professionals de suport", "Empleat
s comptables, administratius i empleats d'oficina", "Tècnics i professionals científics i intel·lec
tuals", "Ns/Nc", "Operadors d'instal·lacions i maquinària i muntadors", "Ocupacions elementals", "T
reballadors al servei de la restauració, personals, protecció i venedors", "Directors i gerents", "
Artesans i treballadors qualificats de les indústries manufactureres i la construcció", "Treballado
rs qualificats en activitats agrícoles, ramaderes, forestals i perqueres", "No aplica", "Ocupacions
militars"))
bbdd$requisit <- factor(bbdd$requisit, labels = c("La vostra titulació específica", "Només ser titu
lat universitari", "No calia titulació universitària"))
bbdd$funprop1 <- factor(bbdd$funprop1, labels = c("Sí", "No"))
bbdd$funprop2 <- factor(bbdd$funprop2, labels = c("No aplica", "No", "Sí"))
bbdd$funprop1_antic <- factor(bbdd$funprop1_antic, labels = c("Sí", "No aplica", "No"))
bbdd$tipcontr <- factor(bbdd$tipcontr, labels = c("Fix / Indefinit", "Temporal", "Autónom"))
bbdd$jorn_tc <- factor(bbdd$jorn_tc, labels = c("Sí", "No"))
bbdd$durcontr <- factor(bbdd$durcontr, labels = c("No aplica", "Entre sis mesos i un any", "Menys d
e sis mesos", "Ns/Nc", "Més d'un any"))
bbdd$ambit <- factor(bbdd$ambit, labels = c("Privat", "Públic", "Ns/Nc"))
bbdd$guanys <- factor(bbdd$guanys, labels = c("Entre 24.001 i 30.000", "Entre 9.000 i 12.000 €", "M
enys de 9.000 €", "Ns/Nc", "Entre 18.001 i 24.000 €", "Entre 15.001 i 18.000 €", "Entre 40.001 i 50
.000 €)", "Entre 12.001 i 15.000 €", "Entre 30.001 i 40.000 €", "Més de 50.000 €"))
bbdd$numtreb <- factor(bbdd$numtreb, labels = c("Entre 11 i 50", "Entre 51 i 100", "10 o menys", "M
és de 500", "Ns/Nc", "Entre 251 i 500", "Entre 101 i 250"))
bbdd$sexe <- factor(bbdd$sexe, labels = c("Dona", "Home"))
bbdd$adequacio_c <- factor(bbdd$adequacio_c, labels = c("Requisit titulació específica i funcions p
ròpies tit. esp.", "Requisit titulació universitària i funcions no pròpies", "Requisit titulació un
iversitària i funcions pròpies universitàries", "Cap requisit i funcions no universitàries", "Cap r
equisit però funcions universitàries", "Requisit titulació específica i funcions no pròpies tit. es
p."))
bbdd$funcions_c <- factor(bbdd$funcions_c, labels = c("Funcions específiques de la titulació", "Fun
cions no universitàries", "Funcions universitàries"))
bbdd$Sobreq <- as.numeric(bbdd$Sobreq)
bbdd$Sobreq <- factor(bbdd$Sobreq, labels = c("Normal", "Sobrequalificat"))

#Eliminem els registres que contenen NA's
bbdd <- bbdd[complete.cases(bbdd), ]

bbdd$sitact <- factor(bbdd$sitact, labels = c("Treballo"))

bbdd$anyini_c <- factor(bbdd$anyini_c, labels = c("Fa 2 anys",
"Menys d'un any", "Fa més de 3 anys", "Fa 3 anys", "Fa 1 any"))

bbdd$autonom_c <- factor(bbdd$autonom_c, labels = c("No",
"Sí"))

bbdd$codi_a <- factor(bbdd$codi_a, labels = c("Eng. i Arquitectura",
"Arts i Humanitats", "Ciències de la salut", "Ciències socials i jurídiques",
"Ciències"))
```

```
bbdd$cno_1digit <- factor(bbdd$cno_1digit, labels = c("Tècnics i professionals de
suport", "Empleats comptables, administratius i empleats d'oficina", "Tècnics i
professionals científics i intel·lectuals", "Ns/Nc", "Operadors d'instal·lacions i
maquinària i muntadors", "Ocupacions elementals", "Treballadors al servei de la
restauració, personals, protecció i venedors", "Directors i gerents", "Artesans i
treballadors qualificats de les indústries manufactureres i la construcció",
"Treballadors qualificats en activitats agrícoles, ramaderes, forestals i perqueres",
"No aplica", "Ocupacions militars"))

bbdd$requisit <- factor(bbdd$requisit, labels = c("La vostra titulació específica",
"Només ser titulat universitari", "No calia titulació universitària"))

bbdd$funprop1 <- factor(bbdd$funprop1, labels = c("Sí", "No"))

bbdd$funprop2 <- factor(bbdd$funprop2, labels = c("No aplica", "No", "Sí"))

bbdd$funprop1_antic <- factor(bbdd$funprop1_antic, labels = c("Sí", "No aplica", "No"))

bbdd$tipcontr <- factor(bbdd$tipcontr, labels = c("Fix / Indefinit", "Temporal",
"Autónom"))

bbdd$jorn_tc <- factor(bbdd$jorn_tc, labels = c("Sí", "No"))

bbdd$durcontr <- factor(bbdd$durcontr, labels = c("No aplica", "Entre sis mesos i un
any", "Menys de sis mesos", "Ns/Nc", "Més d'un any"))

bbdd$ambit <- factor(bbdd$ambit, labels = c("Privat", "Públic", "Ns/Nc"))

bbdd$guanys <- factor(bbdd$guanys, labels = c("Entre 24.001 i 30.000", "Entre 9.000 i
12.000 €", "Menys de 9.000 €", "Ns/Nc", "Entre 18.001 i 24.000 €", "Entre 15.001 i
18.000 €", "Entre 40.001 i 50.000 €)", "Entre 12.001 i 15.000 €", "Entre 30.001 i 40.000
€", "Més de 50.000 €"))

bbdd$numtreb <- factor(bbdd$numtreb, labels = c("Entre 11 i 50", "Entre 51 i 100", "10 o
menys", "Més de 500", "Ns/Nc", "Entre 251 i 500", "Entre 101 i 250"))

bbdd$sexe <- factor(bbdd$sexe, labels = c("Dona", "Home"))

bbdd$adequacio_c <- factor(bbdd$adequacio_c, labels = c("Requisit titulació específica i
funcions pròpies tit. esp.", "Requisit titulació universitària i funcions no pròpies",
"Requisit titulació universitària i funcions pròpies universitàries", "Cap requisit i
funcions no universitàries", "Cap requisit però funcions universitàries", "Requisit
titulació específica i funcions no pròpies tit. esp."))

bbdd$funcions_c <- factor(bbdd$funcions_c, labels = c("Funcions específiques de la
titulació", "Funcions no universitàries", "Funcions universitàries"))

bbdd$Sobreq <- as.numeric(bbdd$Sobreq)
bbdd$Sobreq <- factor(bbdd$Sobreq, labels = c("Normal", "Sobrequalificat"))
```
#SELECCIÓ DE VARIABLES
```{r}
#Seleccio de variables numeriques més rellevants
## Factorial exploratorio

# seleccionar variables satisfaccion
#17-21
datos_satisfaccio <- bbdd[ , 16:20]

# seleccionar variables nivel
#22-35
datos_nivel <- bbdd[ , 21:34]
```

```
# seleccionar variables utilidad
#36-49
datos_utilidad <- bbdd[ , 35:48]

x <- cbind(datos_nivel,datos_satisfaccio,datos_utilidad)
x <- na.omit(x)

#scree.plot(x,type = 'R',title = "Gráfico de Sedimentación")

r <- cor(x)
KMO(r) # Estadístico KMO
cortest.bartlett(r,n=nrow(datos_satisfaccio)) # Prueba de esfericidad de Barlett

# Modelo factorial exploratorio  con todas las variables 3 factores,
# rotación varimax, método de estimación ejes principales (fa=pa).
fa1 <- fa(x,nfactors = 3,fm="pa",rotate="varimax",scores="regression")
names(fa1)
print(fa1,sort = T)

# Diagrama
diagram(fa1, e.size = 0.1,cex=0.75)

#fa1$scores
```

```{r}
#VARIABLES SCORES + EXPLICATIVES ESTUDIADES
bbdd1 <- cbind(bbdd, fa1$scores)
bbdd1 <- subset(bbdd1, select = c(Sobreq, PA1, PA2, PA3, sitact, anyini_c, autonom_c,
codi_a, guanys, numtreb))
str(bbdd1)
```
#Anàlisi descriptiva univariant
```{r}
# Anàlisi descriptiva univariant
## Numèriques Histograma i boxplot per les variables numèriques
for (k in 1:ncol(bbdd1)) {
  if (is.numeric(bbdd1[[k]])) {
      hist(bbdd1[[k]], main=paste("Histogram of", names(bbdd1)[k]))
      boxplot(bbdd1[[k]], main=paste("Boxplot of", names(bbdd1)[k]))
  }
}

# Obtener columnas numéricas
cols_numericas <- which(sapply(bbdd1, is.numeric))

# Crear lista para guardar resultados
lista_tablas <- lapply(cols_numericas, function(col) {
  # Obtener nombre de la columna
  nombre_col <- names(bbdd1)[col]

  # Crear tabla de resumen
  tabla <- summary(bbdd1[[col]])
  tabla_matriz <- t(as.matrix(tabla))
  tabla_df <- as.data.frame(tabla_matriz)

  # Agregar nombre de la columna como variable a la tabla
  tabla_df$Variable <- nombre_col
  tabla_df <- tabla_df[, c("Variable", names(tabla))]

  # Devolver tabla
  print(tabla_df)
})
```

```{r}
#Sitact
vNC <- freq(bbdd1$sitact, plot = FALSE)
kable(vNC)%>%
kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>%
column_spec(1, border_right = T)
pcNC <- bbdd1$sitact
PieChart(pcNC, main="Situació laboral actual")
```

```{r}
#anyini_c
vNC <- freq(bbdd1$anyini_c, plot = FALSE)
kable(vNC)%>%
kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>%
column_spec(1, border_right = T)
pcNC <- bbdd1$anyini_c
PieChart(pcNC, main="Experiència Laboral")
```

```{r}
#autonom_c
vNC <- freq(bbdd1$autonom_c, plot = FALSE)
kable(vNC)%>%
kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>%
column_spec(1, border_right = T)
pcNC <- bbdd1$autonom_c
PieChart(pcNC, main="Tipus de contracte: autònom")
```

```{r}
#codi_a
vNC <- freq(bbdd1$codi_a, plot = FALSE)
kable(vNC)%>%
kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>%
column_spec(1, border_right = T)
pcNC <- bbdd1$codi_a
PieChart(pcNC, main="Branca Titulació")
```

```{r}
#guanys
vNC <- freq(bbdd1$guanys, plot = FALSE)
kable(vNC)%>%
kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>%
pcNC <- bbdd1$guanys
PieChart(pcNC, main="Guanys anuals bruts")
```

```{r}
#numtreb
vNC <- freq(bbdd1$numtreb, plot = FALSE)
kable(vNC)%>%
kable_styling(bootstrap_options = "bordered", stripe_color = TRUE,
position = "center", full_width = TRUE)%>%
column_spec(1, border_right = T)
pcNC <- bbdd1$numtreb
PieChart(pcNC, main="Nombre de treballadors")
```

```{r}
# ANÀLISI BIVARIANT
library(corrplot)
library(Hmisc)
# seleccionar solo las variables numéricas
datos_numericos <- bbdd1 %>% select_if(is.numeric)

# Eliminar las filas con valores faltantes en cualquier variable numérica
datos_numericos_completos <- datos_numericos[complete.cases(bbdd1),]
cor(datos_numericos)
```

MODELS DE CLASSIFICACIÓ
```{r}
set.seed(123)
training<-sample(1:nrow(bbdd1), round(2*nrow(bbdd1)/3))
# Crear un subconjunto de los datos de prueba para visualizar
dataTrain <- bbdd1[training, ]
dataTest <- bbdd1[-training, ]
```

```{r}
dataTrain2 <- subset(dataTrain, select = -c(sitact))
dataTest2 <- subset(dataTest, select = -c(sitact))
```

ARBRES DE DECISIÓ
```{r}
#DATOS TRAIN
model1 <- rpart(Sobreq~PA1 +PA2 + PA3 + sitact + anyini_c + autonom_c + codi_a + guanys
+ numtreb, data=dataTrain, method="class", parms = list(split="information"))
summary(model1)
rpart.plot::rpart.plot(model1)

#DATOS TEST
model2 <- rpart(Sobreq~PA1 +PA2 + PA3 + sitact + anyini_c + autonom_c + codi_a + guanys
+ numtreb, data=dataTest, method="class", parms = list(split="information"))
rpart.plot::rpart.plot(model2)
```
MATRIU CONFUSIÓ
```{r}
# Realitza les prediccions utilitzant les dades TRAIN
class_pred <- predict(model1, newdata = dataTrain, type = "class")

confusion_matrix_train <- confusionMatrix(class_pred, dataTrain[["Sobreq"]])
confusion_matrix_train

# Realitza les prediccions utilitzant les dades TEST
class_pred2 <- predict(model2, newdata = dataTest, type = "class")

confusion_matrix_test <- confusionMatrix(class_pred2, dataTest[["Sobreq"]])
confusion_matrix_test
```

METRIQUES DADES TRAIN
```{r}
VN <- confusion_matrix_train$table[2, 2]
VP <- confusion_matrix_train$table[1, 1]
FP <- confusion_matrix_train$table[1, 2]
FN <- confusion_matrix_train$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy
```

```
# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TRAIN
prob_pred <- predict(model1, newdata = dataTrain, type = "prob")

# Extraer las probabilidades de la clase positiva
prob_positive <- prob_pred[, "Sobrequalificat"]

# Calcular la curva ROC
roc_obj <- roc(dataTrain$Sobreq, prob_positive)

# Graficar la curva ROC TRAIN
plot(roc_obj, main = "Curva ROC")

# Calcular el área bajo la curva ROC
auc <- roc(dataTrain$Sobreq, prob_positive)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de entrenament:", auc))
```
METRIQUES DADES TEST
```{r}
VN <- confusion_matrix_test$table[2, 2]
VP <- confusion_matrix_test$table[1, 1]
FP <- confusion_matrix_test$table[1, 2]
FN <- confusion_matrix_test$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy

# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TEST
prob_pred <- predict(model2, newdata = dataTest, type = "prob")

# Extraer las probabilidades de la clase positiva
prob_positive <- prob_pred[, "Sobrequalificat"]

# Calcular la curva ROC
roc_obj <- roc(dataTest$Sobreq, prob_positive)

# Graficar la curva ROC
plot(roc_obj, main = "Curva ROC")
```

```
# Calcular el área bajo la curva ROC
auc2 <- roc(dataTest$Sobreq, prob_positive)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de proba:", auc2))
```


#NAIVE BAYES
```{r}
#DADES TRAIN
nb <- naive_bayes(Sobreq ~ ., data = dataTrain)
summary(nb)
plot(nb, legend=T)
plot(nb, which = "codi_a", legend = TRUE)
plot(nb, which = "anyini_c", legend = TRUE)

#DADES TEST
nb2 <- naive_bayes(Sobreq ~ ., data = dataTest)
summary(nb2)
plot(nb2, legend=T)
```
MATRIU CONFUSIÓ
```{r}
# Obtener las probabilidades predichas TRAIN
class_pred <- predict(nb, newdata = dataTrain, type = "class")

confusion_matrix_train <- confusionMatrix(class_pred, dataTrain[["Sobreq"]])
confusion_matrix_train

# Obtener las probabilidades predichas TEST
class_pred2 <- predict(nb2, newdata = dataTest, type = "class")

confusion_matrix_test <- confusionMatrix(class_pred2, dataTest[["Sobreq"]])
confusion_matrix_test
```

METRIQUES DADES TRAIN
```{r}
VN <- confusion_matrix_train$table[2, 2]
VP <- confusion_matrix_train$table[1, 1]
FP <- confusion_matrix_train$table[1, 2]
FN <- confusion_matrix_train$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy

# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TRAIN
prob_pred <- predict(nb, newdata = dataTrain, type = "prob")

# Extraer las probabilidades de la clase positiva
prob_positive <- prob_pred[, "Sobrequalificat"]
```

```
# Calcular la curva ROC
roc_obj <- roc(dataTrain$Sobreq, prob_positive)

# Graficar la curva ROC TRAIN
plot(roc_obj, main = "Curva ROC")

# Calcular el área bajo la curva ROC
auc <- roc(dataTrain$Sobreq, prob_positive)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de entrenament:", auc))
```

METRIQUES DADES TEST
```{r}
VN <- confusion_matrix_test$table[2, 2]
VP <- confusion_matrix_test$table[1, 1]
FP <- confusion_matrix_test$table[1, 2]
FN <- confusion_matrix_test$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy

# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TEST
prob_pred <- predict(nb2, newdata = dataTest, type = "prob")

# Extraer las probabilidades de la clase positiva
prob_positive <- prob_pred[, "Sobrequalificat"]

# Calcular la curva ROC
roc_obj <- roc(dataTest$Sobreq, prob_positive)

# Graficar la curva ROC
plot(roc_obj, main = "Curva ROC")

# Calcular el área bajo la curva ROC
auc2 <- roc(dataTest$Sobreq, prob_positive)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de proba:", auc2))
```


#SVM
```{r}
# DADES TRAIN
# Ajustar el modelo SVM
svm_modelfit <- svm(Sobreq ~ ., data = dataTrain2,kernel="linear",scale=FALSE)
print(svm_modelfit)

# Gráfico de dispersión de las variables PA1 y PA2
plot(dataTrain2$PA1, dataTrain2$PA2, col = as.numeric(dataTrain2$Sobreq), pch = 19, xlab
```

```
= "PA1", ylab = "PA2")
# Superponer los vectores de soporte
points(dataTrain2$PA1[svm_modelfit$index], dataTrain2$PA2[svm_modelfit$index], col =
"red", pch = 4, cex = 1.5)

# Agregar leyenda
legend("topright", legend = c("Normal", "Sobrequalificat", "Vectores de soporte"), col =
c(1, 2, "red"), pch = c(19, 19, 4), cex = 1)


#DADES TEST
# Ajustar el modelo SVM
svm_modelfit2 <- svm(Sobreq ~ ., data = dataTest2,kernel="linear",scale=FALSE)
print(svm_modelfit2)
```

MATRIU CONFUSIÓ
```{r}
# Obtener las probabilidades predichas TRAIN
class_pred_train <- predict(svm_modelfit, dataTrain2, type = "class")

confusion_matrix_train <- confusionMatrix(class_pred_train, dataTrain2[["Sobreq"]])
confusion_matrix_train

# Obtener las probabilidades predichas TEST
class_pred_test <- predict(svm_modelfit2, newdata = dataTest2, type = "class")

confusion_matrix_test <- confusionMatrix(class_pred_test, dataTest2[["Sobreq"]])
confusion_matrix_test
```

METRIQUES DADES TRAIN
```{r}
VN <- confusion_matrix_train$table[2, 2]
VP <- confusion_matrix_train$table[1, 1]
FP <- confusion_matrix_train$table[1, 2]
FN <- confusion_matrix_train$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy

# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TRAIN
prob_pred <- predict(svm_modelfit, newdata = dataTrain, type = "prob")

# Convert factor to numeric probabilities
prob_pred_numeric <- as.numeric(prob_pred == "Sobrequalificat")

# Calcular la curva ROC
roc_obj <- roc(dataTrain$Sobreq, prob_pred_numeric)

# Graficar la curva ROC TRAIN
plot(roc_obj, main = "Curva ROC")
```

```
# Calcular el área bajo la curva ROC
auc <- roc(dataTrain$Sobreq, prob_pred_numeric)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de entrenament:", auc))
```

METRIQUES DADES TEST
```{r}
VN <- confusion_matrix_test$table[2, 2]
VP <- confusion_matrix_test$table[1, 1]
FP <- confusion_matrix_test$table[1, 2]
FN <- confusion_matrix_test$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy

# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TEST
prob_pred <- predict(svm_modelfit2, newdata = dataTest, type = "prob")

# Convert factor to numeric probabilities
prob_pred_numeric2 <- as.numeric(prob_pred == "Sobrequalificat")

# Calcular la curva ROC
roc_obj <- roc(dataTest$Sobreq, prob_pred_numeric2)

# Graficar la curva ROC
plot(roc_obj, main = "Curva ROC")

# Calcular el área bajo la curva ROC
auc2 <- roc(dataTest$Sobreq, prob_pred_numeric2)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de proba:", auc2))
```


REGRESSIÓ LOGISTICA
```{r}
# Ajustar el modelo de clasificación logística con los datos de entrenamiento
model <- glm(Sobreq ~ PA1 + PA2 + PA3 + anyini_c + autonom_c + codi_a + guanys +
numtreb, data = dataTrain2, family = binomial)
print(model)

# Graficar el modelo y los datos de entrenamiento
plot(dataTrain2$PA1, dataTrain2$PA2, col = ifelse(dataTrain2$Sobreq == "Normal", "blue",
"red"),
     xlab = "PA1", ylab = "PA2")
points(dataTrain2$PA1, dataTrain2$PA2, col = ifelse(predict(model, type = "response") >
0.5, "lightblue", "lightcoral"), pch = 20)
legend("topright", legend = levels(dataTrain2$Sobreq), col = c("blue", "red"), pch = 1)
```

```
model2 <- glm(Sobreq ~ PA1 + PA2 + PA3 + anyini_c + autonom_c + codi_a + guanys +
numtreb, data = dataTest2, family = binomial)
print(model2)

# Graficar el modelo y los datos de entrenamiento
plot(dataTest2$PA1, dataTest2$PA2, col = ifelse(dataTest2$Sobreq == "Normal", "blue",
"red"),
     xlab = "PA1", ylab = "PA2")
points(dataTest2$PA1, dataTest2$PA2, col = ifelse(predict(model2, type = "response") >
0.5, "lightblue", "lightcoral"), pch = 20)
legend("topright", legend = levels(dataTest2$Sobreq), col = c("blue", "red"), pch = 1)
```

MATRIU CONFUSIÓ

```{r}
# Obtener las probabilidades predichas TRAIN
predicted <- predict(model, newdata = dataTrain2, type = "response")
predicted_class <- ifelse(predicted >= 0.5, "Sobrequalificat", "Normal")
predicted_class <- factor(predicted_class, levels = levels(dataTrain2$Sobreq))

confusion_matrix_train <- confusionMatrix(predicted_class, dataTrain2$Sobreq)
confusion_matrix_train

# Obtener las probabilidades predichas TEST
predicted2 <- predict(model2, newdata = dataTest2, type = "response")
predicted_class2 <- ifelse(predicted2 >= 0.5, "Sobrequalificat", "Normal")
predicted_class2 <- factor(predicted_class2, levels = levels(dataTest2$Sobreq))

confusion_matrix_test <- confusionMatrix(predicted_class2, dataTest2$Sobreq)
confusion_matrix_test
```

METRIQUES DADES TRAIN

```{r}
VN <- confusion_matrix_train$table[2, 2]
VP <- confusion_matrix_train$table[1, 1]
FP <- confusion_matrix_train$table[1, 2]
FN <- confusion_matrix_train$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy

# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TRAIN
prob_pred <- predict(model, newdata = dataTrain2, type = "response")

# Calcular la curva ROC
roc_obj <- roc(dataTrain2$Sobreq, prob_pred)

# Graficar la curva ROC TRAIN
plot(roc_obj, main = "Curva ROC")
```

```
# Calcular el área bajo la curva ROC
auc <- roc(dataTrain2$Sobreq, prob_pred)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de entrenament:", auc))
```

METRIQUES DADES TEST
```{r}
VN <- confusion_matrix_test$table[2, 2]
VP <- confusion_matrix_test$table[1, 1]
FP <- confusion_matrix_test$table[1, 2]
FN <- confusion_matrix_test$table[2, 1]

# Exactitud / Accuracy
accuracy <- (VP + VN)/(VP + FP + FN + VN)
accuracy

# Precisio
precision <- VP / (VP + FP)
precision

# Recall / Sensibilitat
recall <- VP / (VP + FN)
recall

# Especificitat
espe <- VN / (VN + FP)
espe

# Obtener las probabilidades predichas TEST
prob_pred <- predict(model2, newdata = dataTest2, type = "response")

# Calcular la curva ROC
roc_obj <- roc(dataTest2$Sobreq, prob_pred)

# Graficar la curva ROC
plot(roc_obj, main = "Curva ROC")

# Calcular el área bajo la curva ROC
auc2 <- roc(dataTest2$Sobreq, prob_pred)$auc

# Imprimir el valor del área bajo la curva ROC
print(paste("L'area sota la corba de ROC per a les dades de proba:", auc2))
```

PERFIL ESTUDIANT MODEL REGRESSIÓ LOGÍSTICA
```{r}
bbdd2 <- subset(bbdd1, select = -c(sitact))

model <- glm(Sobreq ~ PA1 + PA2 + PA3 + anyini_c + autonom_c + codi_a + guanys +
numtreb, data = bbdd2, family = binomial)
summary(model)
```

```{r}
coef(model) %>%
  as_tibble() %>%
  mutate(odds = exp(value))
```

```{r}
# Obtener los coeficientes del modelo
coeficientes <- coef(model)
```

```
# Calcular las odds ratios
odd <- exp(coeficientes)

# Calcular las probabilidades
probabilidades <- 1 / (1 + exp(-coeficientes))

# Crear una tabla con los coeficientes, las odds ratios y las probabilidades
tabla <- cbind(coeficientes, odd, probabilidades)

# Renombrar las columnas
colnames(tabla) <- c("Coeficients", "Odds", "Probabilitat")

# Imprimir la tabla
print(tabla)
```