Treball final de grau

GRAU D'ENGINYERIA INFORMÀTICA

Facultat de Matemàtiques
Universitat de Barcelona

IMPORTÀNCIA DE LES DADES EN LA PREDICCIÓ DE L'ABANDONAMENT UNIVERSITARI

Autor: Lluc Slatosch Vives

Directora: Dra. Laura Igual
Realitzat a: Departament de Matemàtiques i Informàtica
Data: 17 de Gener de 2024

Resum

Actualment l'abandonament universitari representa un problema a la majoria dels estudis de les universitats catalanes. Aquest treball es tracta d'una investigació sobre quina informació sobre un estudiant universitari és més important a l'hora de predir si l'estudiant abandonarà els seus estudis. Per a fer això s'ha fet una comparació entre dos conjunts de dades acadèmiques, un de la Universitat de Barcelona (UB) i un altre del Polytechnic Institute of Portalegre (PIP). El conjunt de dades de la UB disposa únicament de les qualificacions dels estudiants. El conjunt de dades del PIP disposa d'un nombre elevat de característiques sobre cada estudiant, les quals s'analitzen per definir la importància d'aquestes en la predicció de l'abandonament i fer una proposta a la UB sobre quines dades s'haurien de considerar, a més de les qualificacions. Aquest estudi s'ha realitzat mitjançant ciència de dades i models d'aprenentatge automàtic.

Resumen

Actualmente, el abandono universitario representa un problema en la mayoría de los estudios de las universidades catalanas. Este trabajo es una investigación sobre qué información sobre un estudiante universitario es más importante para predecir si el estudiante abandonará sus estudios. Para ello se ha realizado una comparación entre dos conjuntos de datos académicos, uno de la Universidad de Barcelona (UB) y otro del Polytechnic Institute of Portalegre (PIP). El conjunto de datos de la UB dispone únicamente de las calificaciones de los estudiantes. El conjunto de datos del PIP dispone de un número elevado de características sobre cada estudiante, las cuales se analizan para definir la importancia de éstas en la predicción del abandono y hacer una propuesta en la UB sobre qué datos se deberían considerar, además de las calificaciones. Este estudio se ha realizado mediante ciencia de datos y modelos de aprendizaje automático.

Abstract

Currently, university dropout is a problem in most studies at Catalan universities. This paper is an investigation of which information about a university student is most important in predicting whether the student will drop out. For this purpose, a comparison has been made between two academic datasets, one from the University of Barcelona (UB) and the other

from the Polytechnic Institute of Portalegre (PIP). The UB dataset has only students' grades. The PIP dataset has a large number of characteristics about each student, which are analysed to define the importance of these in predicting dropout and to make a proposal at UB about what data should be considered in addition to grades. This study has been carried out using data science and machine learning models.

Índex

1. Introducció

L'abandonament d'estudiants a la universitat és un problema recurrent. La indecisió dels joves per triar una carrera que els hi agradi, junt amb la manca de rendiment o dificultats econòmiques que poden tenir, fa que molts estudiants abandonin o canviin el seu grau abans de finalitzar-lo, suposant això una pèrdua de temps i diners per a ells mateixos i la seva família. Per què esperar a que un jove abandoni els estudis al tercer any de la carrera, si analitzant el seu primer curs podem predir si finalitzarà o no i realitzar alguna acció en conseqüència?

1. 1 Motivació

Un estudi realitzat per la Facultat d'Educació de la Universidad Complutense de Madrid [1], ha tret com a conclusió que un 11% dels estudiants de menys de 30 anys que es van inscriure a la universitat el curs 2015-2016 varen abandonar el grau. Més de la meitat ho varen fer al primer curs. Si ampliem la mostra a alumnes majors de 30 anys el percentatge augmenta fins a un 13%. De 1.340.000 estudiants, 174.200 van abandonar la universitat. Aquest estudi, entén com a abandonament aquells alumnes que s'han matriculat per primer cop en un grau i no ho han tornat a fer els dos següents cursos ni tampoc s'han titulat en un màxim de quatre cursos a partir del primer [1]. L'U-ranking 2019 [2], en canvi, va fer un informe on detalla que el 33% d'alumnes espanyols no finalitza el grau on es van matricular. D'aquest 33%, el 21% abandona completament la universitat, mentre que l'altre 12% es canvia de grau. Aquestes elevades taxes, reflexen un important desaprofitament de recursos privats i públics destinats a la formació universitària. Les pèrdues annuals s'estimen a prop dels mil milions d'euros. L'U-ranking 2019 analitzà 62 universitats espanyoles, 48 públiques i 14 privades.

Sigui com sigui, està clar que l'abandonament universitari és un problema a Espanya i té unes taxes bastant millorables. Per això, la Universitat de Barcelona disposa d'un *Pla d'Acció Tutorial,* el qual ofereix als estudiants tutors d'estudis. Aquests tutors han de guiar als alumnes durant els seus estudis, per ajudar-los en cas de estar bloquejats o de no saber com seguir en el grau. El problema d'això, és que cada tutor disposa d'un grup elevat d'estudiants, i per tant és molt difícil que conegui les situacions de tots i que prengui mesures per alumnes delicats. Per a tractar aquest problema són d'ajuda els models de predicció d'abandonament universitari. En comptes de construïr un model, la motivació és ajudar a futurs projectes dedicats a la construcció d'aquests models fent una proposta de millora de les dades disponibles a la Universitat de Barcelona.

És per els problemes mencionats anteriorment que la Dra. Laura Igual (directora d'aquest treball) i jo hem decidit fer aquest projecte.

1.2 Contexte del projecte

Aquest treball tracta d'ajudar a millorar els projectes emmarcats dins d'un projecte d'innovació docent del grup INDOMAIN, format per professors de la Facultat de Matemàtiques i Informàtica.

Per a aquest projecte ja s'han realitzat Treballs de Final de Grau anteriorment. Un dels treballs que construeix un model de predicció d'abandonament universitari és el realitzat per Sergi Rovira al 2017, *Data Science for a New Generation of Tutors: Building an Academic-Guidance System Based on Dropout and Grades Prediction.* És per ajudar a futurs projectes d'aquest tipus que s'ha realitzat aquest treball.

El projecte d'innovació docent es centra en un anàlisi de dades de diferents facultats de la UB per estudiar les qualificacions dels alumnes, entre altre informació, per predir comportaments com l'abandonament. En aquest treball previ s'ha demostrat que la intenció

d'abandonament es pot predir amb una precisió de fins al 85%. Aquesta informació es vol emprar per reconduir alumnes amb intenció d'abandonar, per tal d'ajudar-los a seguir amb la seva carrera. Es vol ampliar l'anàlisi per tal de poder millorar les dades que recolleix la UB, i així poder crear models de predicció millors dels que es tenen.

1.3 Problema tractat i objectius

La idea principal és estudiar estrategies de predicció de l'abandonament fent servir dos conjunts de dades. L'objectiu final és definir quina informació sobre els estudiants s'ha de recollir a la UB per tal de poder predir l'abandonament.

Volem saber si es pot predir l'abandonament de forma precisa i quines dades dels alumnes calen per fer-ho. Per a fer aquest estudi s'han emprat dues bases de dades diferents, la obtinguda de la Universitat de Barcelona i una altra obtinguda on-line del Polytechnic Institute of Portalegre. Aquest segon conjunt de dades disposa d'un nombre elevat de característiques i ens ajudarà a determinar la importància de casdascuna d'aquestes. Estudiarem i decidirem quines d'aquestes variables serien fàcils de recollir per la UB i ajudarien a aconseguir resultats més precisos a l'hora de definir models de predicció d'abandonament.

Per a dur això a terme s'hauràn de realitzar les següents tasques:

1. Formació personal sobre el tema
2. Plantejament d'objectius
3. Obtenció i processament de les dades
4. Estudi i visualització de les dades
5. Estudi i elecció de models predictius
6. Implementació del model predictiu
7. Avaluació de resultats i proposta de millora de dades

2. Obtenció de les dades

Per a començar a treballar en aquest projecte, el primer que es necessita són dos conjunts de dades, el de la UB, i un amb un nombre significativament més elevat de característiques per a fer un estudi de quines són més importants. En aquest capítol s'explicarà l'obtenció dels conjunts de dades amb els que s'ha treballat.

2.1 Conjunt de dades de la UB

Les dades de la UB han estat sol·licitades per la directora del TFG al propi centre, a través del Portal d'Estadístiques de l'Àmbit Acadèmic [18] on s'han obtingut dades acadèmiques des dels anys 2009 al 2023. Aquestes dades inclouen dades acadèmiques d'assignatures de 36 graus diferents.

2.2 Competició i conjunt de dades del PIP

Per a poder complir els nostres objectius, necessitarem estudiar en detall un dataset de dades universitàries amb un nombre elevat de variables. Per a això s'utilitza un dataset molt interessant disponible a la web Kaggle [4]. La competició té com a objectiu trobar les maneres més efectives per a calcular les probabilitats de que un estudiant abandoni la carrera o no. Per a fer això, els creadors de la competició ens proposen un dataset de dades acadèmiques de Portugal. En particular del *Polytechnic Institute of Portalegre*, que és el centre encarregat de realitzar l'estudi presentat a la competició. Els usuaris participants de la competició tenen l'objectiu de implementar un model de predicció sobre l'abandonament universitari dels alumnes. D'aquí en endavant, per a referir-nos a aquest dataset l'anomenarem com a "dades del PIP" (*Polytechnic Institute of Portalegre*). L'estudi es va fer amb resultats acadèmics dels anys 2008/09 fins a 2018/19. Inclou dades de 17 graus de

diferents camps d'aprenentatge [5]. El motiu per el qual s'ha triat aquest dataset és perquè disposa de 34 variables, per tant ens serà força útil per veure quines de totes aquestes són les més importants i poden ser recollides per la UB en el futur per poder millorar les prediccions d'abandonament. A la secció *3.1.1 Descripció dades PIP* explicarem com són aquestes dades més en detall.

3. Anàlisi exploratori de les dades

En aquesta secció es mostrarà amb precisió com s'han analitzat les dades. Primer s'exposarà la preparació de les dades. Després es farà un anàlisi d'aquestes, on es mostraran les característiques més rellevants a l'hora de resoldre el nostre problema. Finalment, mitjançant diferents gràfics es visualitzaran les dades que es considerin més importants.

3.1 Descripció de les dades

Abans de començar a processar les dades s'han de descriure els conjunts de dades amb els que es treballarà. Aquestes són les dades de la Universitat de Barcelona (UB), i les dades del *Polytechnic Institute of Portalegre* (PIP).

3.1.1 Descripció de les dades del PIP

El dataset del PIP té un total de 35 columnes, incloent la columna "Target". El contingut d'aquesta columna és de "Dropout", "Graduate" o "Enrolled". Aquesta característica ens indica si l'alumne, després dels anys que durin naturalment els seus estudis, ha acabat graduat, ha abandonat la carrera o si encara hi està inscrit. Les altres 34 columnes indiquen dades personals o acadèmiques de l'alumne després del 1r any de carrera.

Dins de la secció "Discussion" del kaggle, trobam una pàgina que explica amb més claretat les dades emprades per l'autor [5]. Allà es pot veure que les dades tenen 6 categories principals:

- "Demographic data"
- "Socioeconomic data"
- "Macroeconomic data"
- "Academic data at enrollment"

- “Academic data at the end of the 1st semester”
- “Academic data at the end of the 2nd semester”

A continuació es mostra una breu explicació del que significa cada variable i dels valors que pot tenir aquesta:

- **Dades demogràfiques:**
- Marital status: Estat civil. Categories numèriques d’1 al 6.
- Nationality: Nacionalitat. Categories numèriques d’1 al 21. Gran Majoria Portugal.
- Displaced: Fora del seu país. Valors de 0 o 1. (1 es si i 0 és no, aixi per tots els 0-1 de si/no)
- Gender: Gènere. Valors de 0 o 1.
- Age at enrollment: Edat quan s’apuntà. Va de 17 a 70, mean de 23, median de 20.
- International: Si es estudiant internacional o no. Valors 0-1
- **Dades socioeconòmiques:**
- Father’s/Mother’s qualification: Fins a quin nivell d’educació tenen els pares. Categories numèriques del 1 al 34. Són del tipus “Secondary Education” (1), “10th year” (12),…
- Father’s/Mother’s occupation: Ocupació dels pares. Categories numèriques 1-46.
- Educational special needs: Necessitats especials. 0 o 1.
- Debtor: Si deu diners o no. 0 o 1.
- Tuition fees up to date: Si té els pagaments de la matricula en ordre. 0 o 1.
- Scholarship holder: Si te beca o no. 0 o 1.
- **Dades macroeconòmiques:**
- Unemployment rate: taxa d’atur al pais al moment d’entrar a la carrera. Va de 7,6 a 16,2.

- Inflation rate: Taxa de inflació al pais al moment d'entrar a la carrera. -0,8 a 3,7.
- GDP: Variança de Producte Interior Brut del pais al moment d'entrar a la carrera respecte a l'any anterior. -4,1 a 3,5.
- **Dades acadèmiques al apuntar-se:**
- Application mode: Com s'apunten a la carrera. Categoríes numèriques 1-18.
- Application order: A quina universitat/carrera van entrar de les 9 que es posen per ordre de preferència.
- Course: Quina carrera cursen. Categories numèriques 1 a 17.
- Daytime/evening attendance: Classes al matí (1) o a la tarda (0).
- Previous qualification: Quina educació tenia l'alumne abans d'entrar. 1 a 17, majoria secundària.
- **Dades acadèmiques al final del 1r/2n semestre (separades)**
- Curricular units (credited): Numero de unitats curriculars (assignatures) acreditades al final de 1r o 2n semestre. Majoria de 0.
- Curricular units(enrolled) : Numero d'assignatures matriculades.
- Curricular units(evaluations): Numero d'assignatures que s'han evaluat.
- Curricular units(approves): Numero d'assignatures aprovades
- Curricular units(grade): Nota mitjana de les assignatures (0-18,9)
- Curricular units(without evaluations): Assignatures sense evaluar. (Majoria 0)

A la secció *4.3.1 Anàlisi dades PIP* s'explicarà que volen dir els valors, ja que sembla que tots estàn en format numèric.

Després d'observar una mica els continguts d'aquest dataset, podem començar a fer nosaltres mateixos un anàlisi de les dades de la manera que més ens convengui per així més endavant crear un model de predicció eficient. Amb l'anàlisi d'aquestes dades ens podrem

adonar de quines són les variables més rellevants, les quals es podrien començar a recollir a la UB per a poder tenir bases de dades més útils.

3.1.2 Descripció de les dades de la UB

Pel que fa a les dades de la UB, disposem de 5 taules de dades acadèmiques i 1 taula de dades referents a si l'alumne ha abandonat el grau o no. Els datasets de dades acadèmiques tenen com a variables les següents:

- ID alumne, on cada alumne està representat per un ID únic.
- ID ensenyament, on s'especifica l'ID del grau en que es troba l'alumne.
- Ensenyament, que ens diu el nom del grau.
- ID assignatura ens dona un ID únic de cada assignatura
- Assignatura, que ens dóna el nom de l'assignatura.
- Curs, es refereix a l'any en que es va cursar aquesta assignatura.
- Semestre diu el semestre en que es va cursar l'assignatura corresponent.
- Tipus, que ens diu si l'assignatura és ordinària o reconeguda, entre d'altres.
- Nota 1 i Nota 2, ens mostra si l'estudiant ha suspès, aprovat, tret un notable o excel·lent. 1 i 2 es refereix a la convocatòria de l'exàmen.
- Qualificació 1 i Qualificació 2, que ens diu la qualificació numèrica per a cada assignatura. 1 i 2 es refereix a la convocatòria de l'exàmen.

Per tant, aquests datasets tenen una fila per a cada assignatura cursada per cada alumne, amb les seves corresponents qualificacions i l'any en que la va cursar. Els datasets són els següents:

- Expedient 2009-2012, amb 705517 files de dades.
- Expedient 2013-2014, amb 547158 files de dades.

- Expedient 2015-2017 amb 771473 files de dades.

- Expedient 2018-2020 amb 739804 files de dades.

- Expedient 2021-2023 amb 720978 files de dades.

A banda d'aquests conjunts de dades tenim el conjunt "Resultat alumnes". Aquest conjunt disposa de les següents variables:

- ID, referint-se a l'ID únic de l'alumne.

- ID ensenyament, que ens diu l'ID del grau cursat o per cursar.

- Nom, que ens dona el nom del grau.

- Estat, que ens diu si l'alumne s'ha graduat, traslladat, ha abandonat, segueix matriculat, està sabàtic o si és adaptat.

Un cop sabem com són aquestes dades el que farem es processarles amb l'objectiu de obtenir un dataset equivalent al del PIP. Volem reduir el dataset conservant la màxima informació rellevant possible

3.2 Processament de les dades

En aquest punt es veurà el processament pel que passen les dades abans de fer l'anàlisi i visualització. Amb "processament" ens referim al procés seguit per a poder emprar aquestes dades dins un esquema de machine learning. Generalment es tracta de corregir errors (cleaning) del dataset, com treure valors nuls.

3.2.1 Processament de les dades del PIP

El primer dataset emprat, el del PIP, no ha requerit gaire preparació. Aquest dataset va ser publicat amb l'objectiu que diferents usuaris l'emprin precisament per fer un model de predicció d'abandonament universitari. Així doncs, les dades estan ja preparades per al seu ús. Encara així, a la secció *6.1 Experiments sobre el conjunt de dades del PIP*, es veuran les

diferents modificacions que es poden aplicar per a possiblement millorar l'eficiència del model futur. Si ens referim a errors que pugui tenir el dataset, com valors nuls, o valors fora del seu rang, aquest dataset no en té cap, i està pràcticament preparat per emprar-lo en el nostre problema.

L'únic que s'ha fet és modificar la variable "Target". És la única del dataset que no està representada amb nombres. És una de les més importants, ja que és la que ens diu si un alumne, després de la durada normal del grau, estarà graduat, haurà abandonat o si seguirà en els estudis. Per això ens interessa tenirla numèricament com la resta de variables, per poder així calcular millor la correlació i també per obtenir millors resultats al futur model de predicció.

Abans de passar aquesta variable a valors numèrics, s'ha decidit modificarla, així com es modificaràn les de la UB. Per a obtenir millor resultat transformarem les variables disponibles per a "Target", tant amb les dades del PIP com amb les de la UB, en únicament dos valors: Dropout o no dropout. És a dir, ha abandonat o no ha abandonat el grau que està cursant. En el cas de les dades del PIP tenim "dropout", "graduated" i "enrolled". Juntarem aquestes dues darreres, "graduated" i "enrolled" en un valor "no dropout". Amb aquesta modificació obtindrem millors resultats a l'hora d'entrenar el model. També podem considerar que si un alumne després de la durada natural del curs, encara està apuntat però sense estar graduat, és molt probable que acabi aquesta carrera. Aquesta és una consideració que s'ha pres per així obtenir millors resultats. El valor de la variable serà de "no dropout", és a dir, que no ha abandonat (el qual sabem que és 100% veritat). Que després d'això acabi la carrera o no, ho ha de considerar cadascú per la seva part.

3.2.2 Processament de les dades de la UB

Pel que fa a les dades de la UB si que s'ha requerit d'una preparació de les dades més laboriosa. L'objectiu d'aquest punt és unificar totes les dades que tenim disponibles de la UB en un únic dataset sòlid i preparat per emprar-se en un model de predicció. Degut a que l'objectiu final és millorar les dades de la UB emprant en un futur variables que puguin tenir les dades del PIP, es necessita preparar les dades de la UB amb el mateix format que tenen les del PIP.

Per a aconseguir això treballarem amb Python Notebook, concretament amb la llibreria "Pandas" [6], amb la qual ens serà suficient per a modificar i preparar les dades de la manera desitjada.

El primer pas serà llegir aquestes dades i guardar-les dintre de variables. Per a fer això necessitem saber el delimitador i la codificació dels arxius csv. El delimitador es pot veure d'una manera simple obrint l'arxiu amb el "Bloc de notes". Ens n'adonem que aquest delimitador és ";". Per a la codificació hem emprat la llibreria "Chardet" [7], i veiem que es tracta de una codificació ISO-8859-1. Un cop sabem això, ja és possible llegir els conjunts de dades i adjudicar-los a variables.

El que s'ha fet a continuació és concatenar els 5 conjunts de dades acadèmiques dins un mateix dataset. S'ha obtingut un dataset amb 3484925 files de dades amb ID's d'alumnes fins a 103241. El dataset té el següent format:

IDALUMNE	IDENSENYAMENT	ENSENYAMENT	IDASSIGNATURA	ASSIGNATURA	CURS	SEMESTRE	TIPUS	NOTA_1	NOTA_2	QUALIFICACIO_1	QUALIFICACIO_2
2	G1058	Ciències Polítiques i de l'Administració	362516	Sistema Polític Espanyol	2011	2n semestre	ORDINARI	Suspens	No consta	0.8	0.0
2	G1058	Ciències Polítiques i de l'Administració	362516	Sistema Polític Espanyol	2012	2n semestre	ORDINARI	No presentat	No consta	0.0	0.0
2	G1058	Ciències Polítiques i de l'Administració	362517	Ciència de l'Administració	2012	2n semestre	ORDINARI	No presentat	No consta	0.0	0.0
2	G1058	Ciències Polítiques i de l'Administració	362518	Sociologia General	2011	1r semestre	ORDINARI	Aprovat	No consta	6.0	0.0
2	G1058	Ciències Polítiques i de l'Administració	362519	Estructura Social	2011	2n semestre	ORDINARI	Suspens	No consta	3.2	0.0
...	...	...	...	...	...	...	...	...	...	...	...
103241	G1016	Història	361415	Món Actual	2021	2n semestre	ORDINARI	Notable	No consta	8.1	NaN
103241	G1016	Història	361417	Història d'Àfrica	2022	1r semestre	ORDINARI	Notable	No consta	8.1	NaN
103241	G1016	Història	361419	Història de la Cultura i de les Institucions E...	2023	1r semestre	ORDINARI	No consta	No consta	0.0	NaN
103241	G1016	Història	361421	Hispània Antiga	2023	2n semestre	ORDINARI	No consta	No consta	0.0	NaN
103241	G1016	Història	364013	Història de Grècia	2023	2n semestre	ORDINARI	No consta	No consta	0.0	NaN

Taula 1: Conjunt de dades complet de la UB

La primera modificació que s'ha realitzat sobre aquest dataset amb totes les dades, és quedar-nos només amb el primer any cursat per a cada alumne. Per a fer això primer hem de passar els valors de la variable "Curs" a numèrics, ja que poden estar com a valor de "string". Un cop fet això, definim una condició amb funcions de "Pandas" que ens permeti filtrar els alumnes segons el valor de "Curs" més petit que tengui cadascun. La resta de files les eliminarem aplicant aquesta condició al nostre dataset amb totes les dades. Després d'aplicar això, ens quedem amb un dataset de 1114554 files que mostren només el primer any cursat de cada alumne.

Després d'aquest pas, seguim tenint files que no ens interessa tenir. Aquestes files són les assignatures que tinguin com a "Tipus" valors del tipus "Reconegudes" o "Convalidades". Aquestes assignatures no ens interessen ja que no s'han cursat realment dins d'aquest grau, sino que l'alumne les té reconegudes d'un altre grau que hagi cursat. Aplicam el codi necessari per a fer això i ens quedem només amb les assignatures que tenguin com a valors:

“Ordinari”, “Convocatòria extraordinària” i “Itinerari doble-ordinari-doble titulació”. Un cop aplicats aquests canvis obtenim un dataset de 933610 files.

Per al següent pas s’ha començat a tenir en compte com està montat el conjunt de dades del PIP. Aquest dataset no té la variable “Assignatura”, sino que disposa d’una fila per a cada alumne on les variables referents a qualificacions tracten sobre la mitjana de les notes obtingudes i del nombre d’assignatures aprovades. A banda d’això, al dataset del PIP no hi és el concepte de “1 i 2”, el qual es refereix a la convocatòria en la qual l’alumne ha obtingut les seves qualificacions, per tant volem agafar només el valor màxim de cadascuna. Per exemple, si l’alumne té un 3 a “Qualificacio_1” i un 5 a “Qualificacio_2” només es tindrà en compte el 5 a l’hora de fer la mitjana.

Pel que fa a les variables “Nota_1” i “Nota_2”, el que es farà és convertirla a la columna “Num_Suspeses”, que ens dirà el nombre d’assignatures suspeses per cada alumne durant el seu primer curs. Recordem que totes aquestes dades encara estàn separades per la variable “Semestre” que pot ser 1r, 2n o anual.

Per obtenir aquest dataset s’han emprat funcions de la llibreria “Pandas”. Concretament el que s’ha fet és:

- Crear la columna “Avg_Qualificacio”, la qual per ara es tracta del màxim entre “Qualificacio_1” i “Qualificacio_2” per a cada fila.

- Crear la columna “Num_Suspeses” de la mateixa manera que la columna anterior, posant un 1 com a valor si el valor anterior de les dues notes és “Suspens”, “No presentat” o “No consta” i un 0 si és diferent a aquests valors a les columnes “Nota 1” o “Nota 2”, ja que si ha aprovat a una de les convocatòries ja es considerarà aprovat.

- Creem un nou dataset agrupant el dataset anterior per alumne, grau, semestre i curs

afegint la mitjana de la columna “Avg_Qualificacio” i el sumatori de la columna “Num_Suspeses”.

Després d’aplicar aquest mètode, obtenim un dataset reduït a 223156 files amb el següent format:

	IDALUMNE	ENSENYAMENT	IDENSENYAMENT	SEMESTRE	CURS	AVG_QUALIFICACIO	NUM_SUSPESES
0	1	Filosofia	G1060	1r semestre	2013	9.000000	0
1	1	Filosofia	G1060	2n semestre	2013	6.133333	0
2	2	Ciències Polítiques i de l'Administració	G1058	1r semestre	2011	4.666667	1
3	2	Ciències Polítiques i de l'Administració	G1058	2n semestre	2011	2.000000	2
4	3	Administració i Direcció d'Empreses	G1072	1r semestre	2016	6.700000	0
...	...	...	...	...	...	...	...
223151	103239	Dret	G1055	2n semestre	2019	8.133333	0
223152	103240	Treball Social	G1027	1r semestre	2014	6.550000	0
223153	103240	Treball Social	G1027	2n semestre	2014	8.066667	0
223154	103241	Història	G1016	1r semestre	2018	8.500000	0
223155	103241	Història	G1016	2n semestre	2018	8.750000	0

Taula 2: Dades UB modificades

Veiem que tenim 223156 files però només IDs d’alumnes fins a 103241. El motiu principal d’això és que en principi tenim dues files per a cada alumne, una per a cada semestre. Encara així, si només tenim en compte això, el nombre de files segueix sense quadrar. Hi ha alumnes que aparèixen més de dos cops. Amb un simple mètode miram quins són els alumnes que aparèixen més de dos cops. Obtenim que l’alumne 58764 apareix 6 cops o el 34427 apareix 5 cops, entre molts d’altres. S’han analitzat totes les files en les que aparèixen aquests alumnes i s’ha trobat que aquests alumnes s’han matriculat per més d’un grau, en el cas dels que aparèixen un nombre parell de vegades. En el cas dels imparells, el que passa és que a semestre tenen el valor “anual”, i per tant només aprèixen un cop per grau.

Sobre la columna semestre, s’ha implementat un petit canvi per millorar en memòria i rendiment del futur model. El que s’ha fet és canviar els valors “object” “1r semestre” per una variable de nombre enter de valor 1, “2n semestre” per una de valor 2 i “anual” per una

de valor 3. La quantitat d'aquests valors és de 106117 per a "1r semestre", 104496 per a "2n semestre" i 12543 per "anual".

Per últim, falta afegir al dataset obtingut fins ara la variable "Estat" del dataset mencionat abans "Resultat alumnes". Primer de tot s'ha llegit i adjudicat a una variable de python de la mateixa manera que els anteriors datasets. Després de treure la columna "Nom", que no ens interessa tenim el següent:

	IDALUMNE	IDENSENYAMENT	ESTAT
0	1	G1060	GRADUAT
1	2	G1058	TRASLLAT
2	3	G1072	GRADUAT
3	4	G1015	GRADUAT
4	5	G1036	ABANDO
...	...	...	...
8566	8101	G1025	GRADUAT
8567	8102	G1015	GRADUAT
8568	8102	G1016	ABANDO
8569	8103	G1035	ABANDO
8570	8104	G1028	NaN

Taula 3: Taula de dades de resultats

Veiem que només apareixen 8570 alumnes, comparat amb els 103241 que es tenien fins ara. Aquestes són les dades que s'han rebut de la UB i no ha estat possible aconseguir les dels 90 mil alumnes restants. L'única possibilitat aqui és emprar els 8104 (les 400 files que "sobren" son degudes a que un mateix alumne pot haver fet més d'un grau) alumnes per al nostre futur model de predicció, ja que la columna "Estat" és la etiqueta que s'emprarà i és indispensable.

Per a unir els dos datasets s'ha fet servir un "Left join" [8] depenent de les variables "ID alumne" i "ID ensenyament":

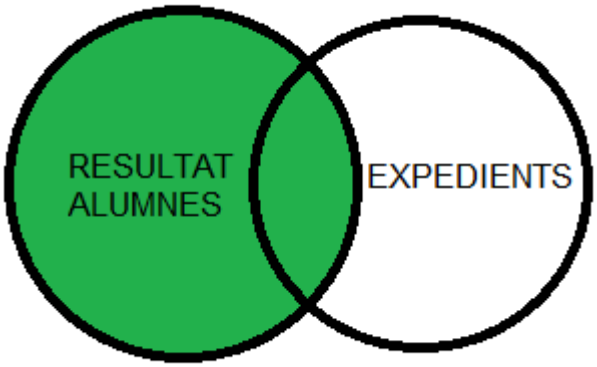


Figura 1: Left Join

El que fa aquest "merge" és agafar tots els "ID alumne" presents a la taula "Resultat alumnes". Aquests IDs es seleccionen de la taula que teniem fins ara, "Expedients", i els seus respectius valors i variables s'afegeixen a la taula "Resultat alumnes", obtenint així un conjunt de dades com el que teniem fins ara, però afegin la columna "Estat" a cada alumne corresponent d'entre aquests 8103 que teniem. Abans hem dit 8104, però aquest darrer tenia un valor nul a "Estat" i s'ha tret. Així doncs, el nostre conjunt de dades ja preparat ens queda de la següent manera:

	IDALUMNE	IDENSENYAMENT	ESTAT	ENSENYAMENT	SEMESTRE	CURS	AVG_QUALIFICACIO	NUM_SUSPESES
0	1	G1060	GRADUAT	Filosofia	1	2013.0	9.000000	0.0
1	1	G1060	GRADUAT	Filosofia	2	2013.0	6.133333	0.0
2	2	G1058	TRASLLAT	Ciències Polítiques i de l'Administració	1	2011.0	4.666667	1.0
3	2	G1058	TRASLLAT	Ciències Polítiques i de l'Administració	2	2011.0	2.000000	2.0
4	3	G1072	GRADUAT	Administració i Direcció d'Empreses	1	2016.0	6.700000	0.0
...	...	...	...	...	...	...	...	...
16936	8101	G1025	GRADUAT	Mestre d'Educació Infantil	3	2009.0	8.600000	0.0
16938	8102	G1016	ABANDO	Història	1	2009.0	0.000000	3.0
16939	8102	G1016	ABANDO	Història	2	2009.0	0.000000	2.0
16940	8103	G1035	ABANDO	Física	1	2009.0	0.000000	1.0
16941	8103	G1035	ABANDO	Física	2	2009.0	0.000000	1.0

Taula 4: Taula final UB

Tenim 16941 files, majoritàriament 2 per a cada alumne, una per a cada semestre. S'ha guardat aquest conjunt de dades a un arxiu "csv" i ja està preparat per a fer un anàlisi visual i el model de predicció, amb un format equivalent del conjunt de dades del PIP.

Finalment s'ha modificat la variable "Estat". Aquesta variable pot tenir els valors de:

- Graduat
- Abandó
- Trasllat
- Matriculat
- Adaptat
- Sabatic

Aquests valors s'han transformat al mateix format en que ho hem fet amb les dades del PIP, quedant-nos amb els valors "Dropout" i "No dropout". Per a fer això hem seguit la mateixa llògica d'abans. Els valors que s'han contat com a "No dropout" són "Graduat" i "Matriculat". Els valors que han passat a ser "Dropout", en canvi, son "Abandó" i "Trasllat".

Aquest darrer el considerem “Dropout” ja que si un alumne s’ha canviat de grau es pot considerar que no ha acabat el grau del qual es va matricular originalment, que és el que ens interessa. Els valors de “Adaptat” i “Sabàtic” son una mica més ambigus, i en total consten només de 323 files. És per aquests motius que s’ha decidit treure-los del conjunt de dades.

3.3 Anàlisi de les dades

Per a poder definir més endavant diferents models de predicció amb les dades del nostres datasets, ens interessa dur a terme un anàlisi d’aquestes dades per tal de veure quines són més o menys rellevants, per veure inclús si es poden eliminar algunes característiques poc importants, estalviant així amb memòria i temps sense alterar gaire el resultat, a banda d’ajudar al nostre objectiu de millorar les dades de la UB.

El llenguatge utilitzat per a fer l’anàlisi serà Python, i més concretament emprarem el format Jupyter Notebook. Aquest format sembla el més oportú per a realitzar l’anàlisi, ja que tindrem un document d’anàlisi de dades visual i fàcil d’entendre. A més, al mateix document més endavant hi podrem afegir els models de machine learning. Per a fer l’anàlisi de dades emprarem principalment les llibreries “Pandas” i “MatPlotLib” [9].

3.3.1 Anàlisi de les dades del PIP

Un cop tenim la nostra taula de dades a una variable de Python, podem començar a investigar el que més ens convengui. S’ha fet un primer anàlisi per comprovar que efectivament no hi ha valors nuls dins el dataset o dades en format extrany. Totes les dades estàn en format numèric. Per designar una variable del tipus “Course”, per exemple, s’empren números de l`1 al 17. Cada curs correspon a un número. Els valors d’aquests, els trobem a l’explicació del conjunt de dades realitzada pels autors [5] i són els següents:

Attribute	Values
Course	1—Biofuel Production Technologies
	2—Animation and Multimedia Design
	3—Social Service (evening attendance)
	4—Agronomy
	5—Communication Design
	6—Veterinary Nursing
	7—Informatics Engineering
	8—Equiniculture
	9—Management
	10—Social Service
	11—Tourism
	12—Nursing
	13—Oral Hygiene
	14—Advertising and Marketing Management
	15—Journalism and Communication
	16—Basic Education
	17—Management (evening attendance)

Taula 5: Categorització de variable “Course”

Ja que tot el conjunt està format per categoríes en format numèric, hi ha altres variables que han hagut de ser categoritzades amb el mateix sistema, i són les següents: *Marital status, Nationality, Application Mode, Previous qualification, Mother’s and father’s values, Mother’s and father’s occupation, Gender values, Attendance regime, Yes/No attributes.* Totes aquestes especificacions es poden trobar tant a l’annex com a la pàgina anomenada [5].

3.3.1.1 Correlacions

El que ens interessa veure a continuació són les correlacions de les variables amb el *Target*. Per a fer això, primer anem a explicar breument què és la correlació.

La correlació és una mesura estadística que expresa fins a quin punt dues variables estàn relacionades linealment, és a dir, com canvien juntes a un ritme constant. És una mesura que s'empra per descriure relacions senzilles sense fer una afirmació sobre causa i efecte [10]. En aquest cas la mesura que emprarem per calcular aquestes variables és un valor en el rang [-1,1]. Com més proper sigui aquest valor a 0, menys depen una variable de l'altre. Si el valor és proper a 1, les dues variables es relacionen equitativament. Si el valor és proper a -1, vol dir que les dues variables es relacionen en direccions contràries.

Feim el càlcul de la correlació de totes les variables amb la variable "Target" amb la funció "corr" de "Pandas" i ho fiquem a un gràfic de barres, el qual ens dona el següent resultat:

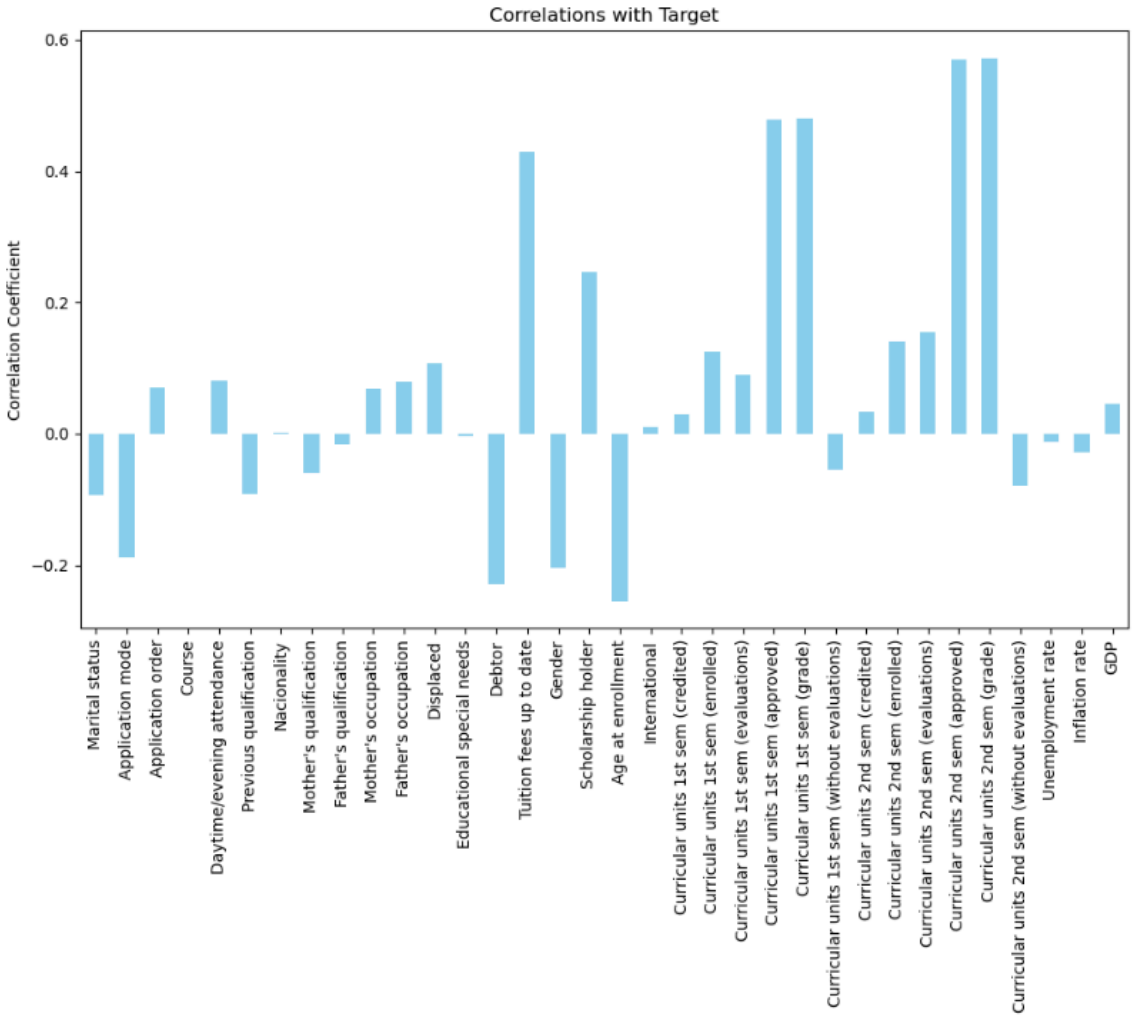


Figura 2: Correlacions amb target. Dades PIP

Més concretament les correlacions son les següents:

```
Marital status                                  -0.093712
Application mode                                -0.188908
Application order                                0.070485
Course                                           0.000083
Daytime/evening attendance                       0.080499
Previous qualification                          -0.091590
Nacionality                                      0.001571
Mother's qualification                          -0.059499
Father's qualification                          -0.016267
Mother's occupation                              0.069102
Father's occupation                              0.079753
Displaced                                        0.107232
Educational special needs                       -0.002806
Debtor                                          -0.229407
Tuition fees up to date                          0.429149
Gender                                          -0.203983
Scholarship holder                               0.245354
Age at enrollment                               -0.254215
International                                    0.010360
Curricular units 1st sem (credited)              0.029308
Curricular units 1st sem (enrolled)              0.124635
Curricular units 1st sem (evaluations)           0.090125
Curricular units 1st sem (approved)              0.479112
Curricular units 1st sem (grade)                 0.480669
Curricular units 1st sem (without evaluations)  -0.054230
Curricular units 2nd sem (credited)              0.033038
Curricular units 2nd sem (enrolled)              0.141515
Curricular units 2nd sem (evaluations)           0.154999
Curricular units 2nd sem (approved)              0.569500
Curricular units 2nd sem (grade)                 0.571792
Curricular units 2nd sem (without evaluations)  -0.079901
Unemployment rate                               -0.012980
Inflation rate                                  -0.027826
GDP                                              0.046319
```

Veiem que el rang de correlacions en aquest cas, si ho passem a valor absolut, va de 0.000083 per a la variable “Course” a 0.571792 per a la variable “Curricular units 2nd sem (grade).

Podem apreciar que hi ha clarament algunes variables més importants que altres. De fet, veiem algunes que pràcticament no tenen correlació amb *Target*, per tant, en principi no ens seràn de gran importància a l’hora de implementar el nostre model predictiu. Anem a veure el mateix gràfic però una mica més especific, on inclourem només el top 10 de variables amb més correlació amb “Target”. El gràfic és el següent:

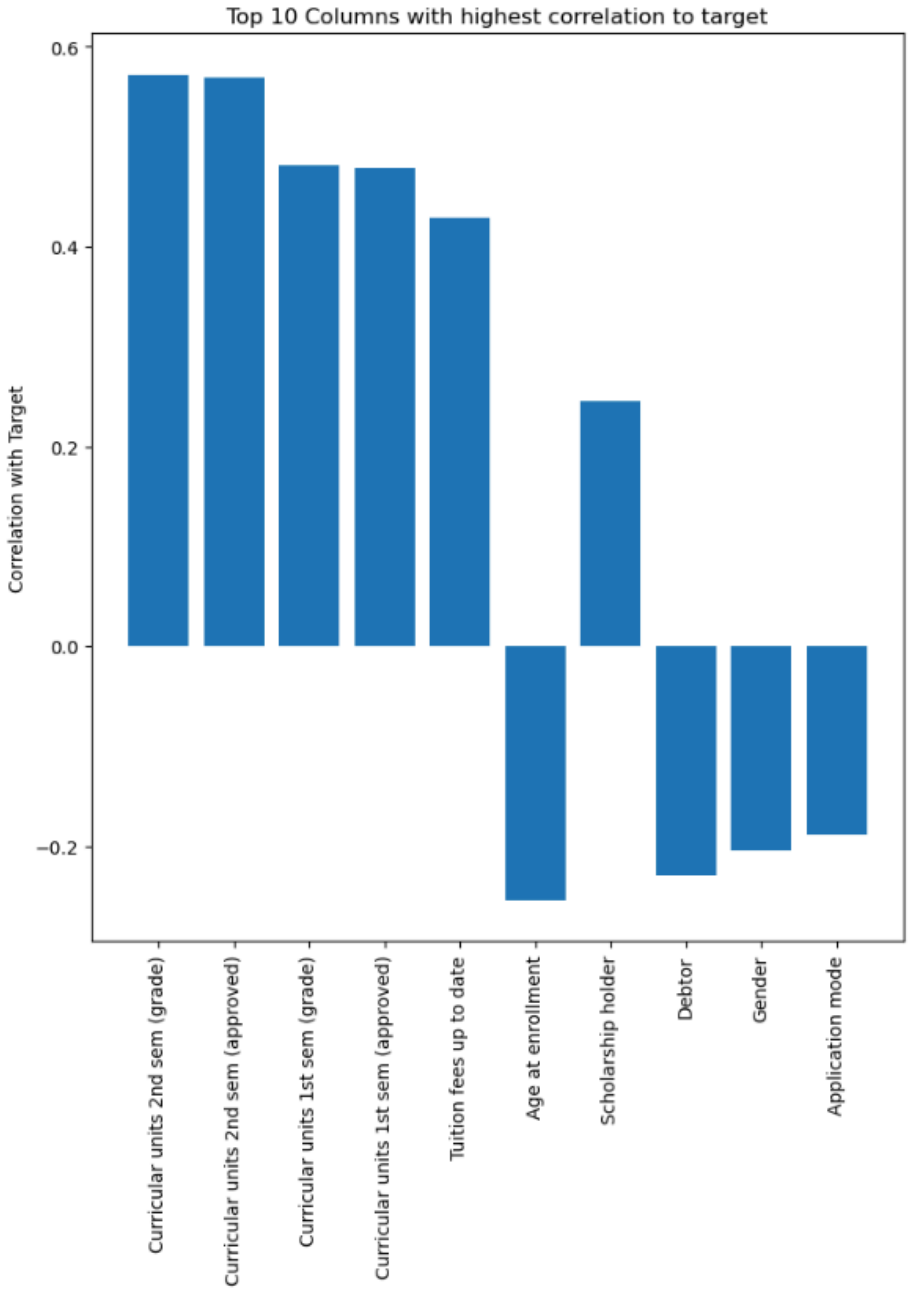


Figura 3: Top 10 correlacions amb “Target”. Dades PIP

Aqui tenim una primera impressió de quines seràn possiblement les variables més importants a l’hora de fer el nostre model de predicció. Veiem principalment les “Curricular Units, 1st i 2nd semester”, per “aprroved” (que diu quantes d’han aprovat) i “grade” (que diu la mitjana de notes que ha tret l’alumne al primer i segon semestre). Aquestes dades les tenim també ala taula de dades de la UB.

A part de les dades acadèmiques, també criden l'atenció altres variables amb alta correlació. Les dues que més capten l'ull són "Gender" i "Age at enrollment". És a dir, el gènere i l'edat quan un alumne es matricula ténen força importància per a determinar si es graduarà o no. Això és molt interessant, ja que la UB té aquestes dades a mà però no les inclou al conjunt de dades, quan possiblement ajudarien a fer millors models de predicció.

Finalment, veim unes altres variables dins aquest top, també interessants, "Tuition fees up to date" i "Debtor". Al que es refereixen aquestes variables és, respectivament, a si la matrícula està pagada i que si es deudor (dos valors de si/no). També tenen una gran correlació amb "Target", i s'estudiarà fins a quin punt afecten al model. Per a la UB, aquestes dades també es podrien recollir fàcilment.

També és interessant veure com aquestes variables es relacionen una amb l'altre. Per fer això, s'ha creat una "Scatter-matrix" [11] amb les 5 variables que més correlació tenen amb target, la qual és la següent:

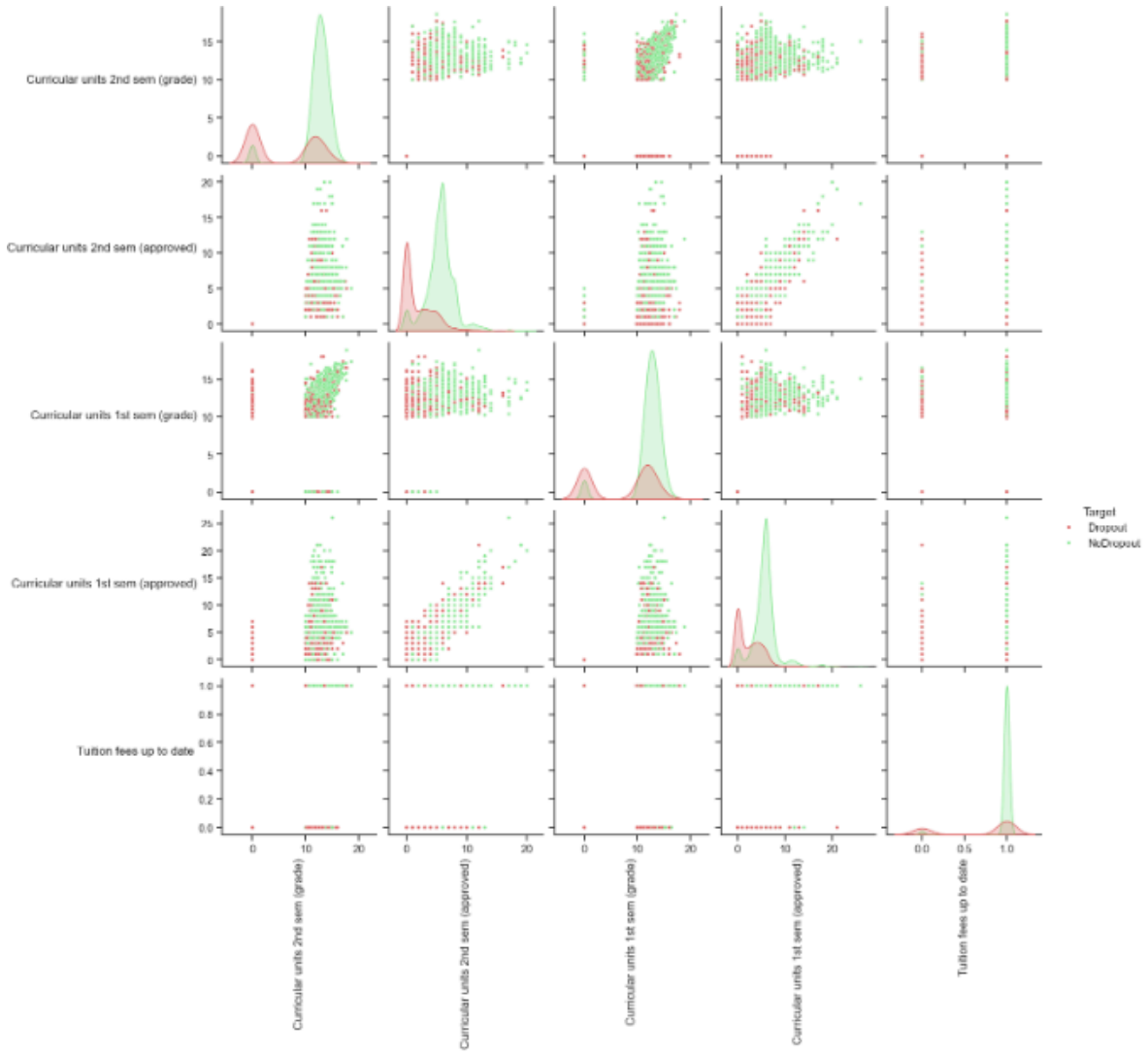


Figura 4: “Scatter-matrix”

El que es pot apreciar aquí és que les relacions entre aquestes variables segueixen una distribució força lineal, sobretot per les dades referents a les qualificacions. Això ens serà d’ajuda per a fer l’elecció de l’algorisme que s’emprarà per entrenar aquestes dades

Finalment s'ha creat un mapa de calor amb totes les variables del dataset per a veure la correlació entre elles:

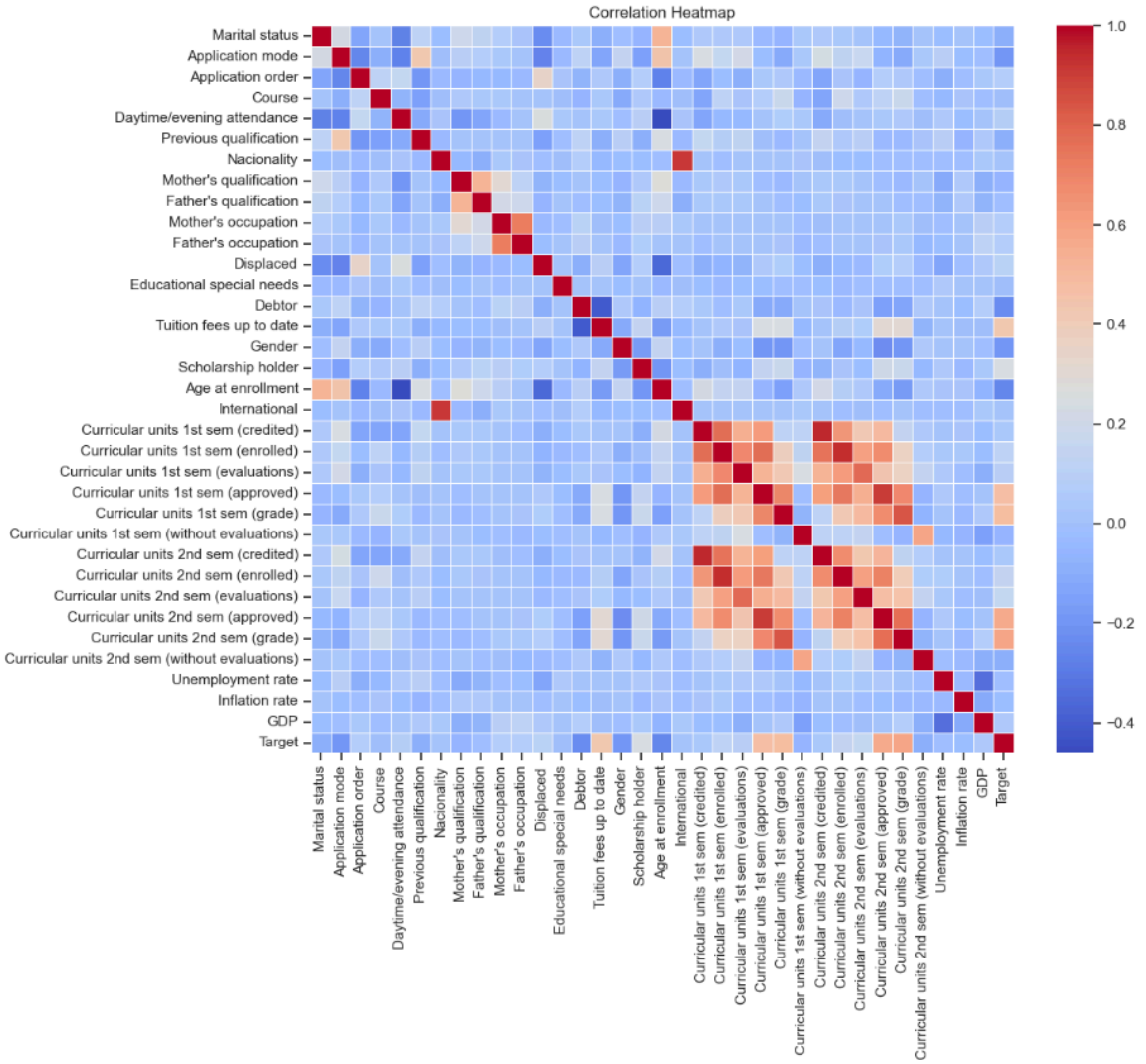


Figura 5: Mapa de calor - PIP

Es pot apreciar que, llògicament, les variables amb més correlació entre elles son les dades purament acadèmiques, és a dir, totes les del tipus "Curricular units". En aquest cas, hi ha poca diferència entre semestres. Les dades del primer semestre estàn correlacionades amb les del segon semestre pràcticament de la mateixa manera que amb el propi primer semestre i viceversa. Altres correlacions "calentes" aïllades que es poden veure son "International" amb

"Nacionality", el qual té bastanta llògica, "Marital status" amb "Age at enrollment" i, curiosament, "Mother's qualification" i "Mother's occupation" amb les variables equivalents per a "Father".

3.3.1.2 Visualització de les dades per estat final

Un cop estudiades les correlacions, anem doncs a analitzar i visualitzar les dades que considerem més importants. El primer que visualitzarem és un gràfic circular del total de "Dropout" i "No Dropout" que tenim:

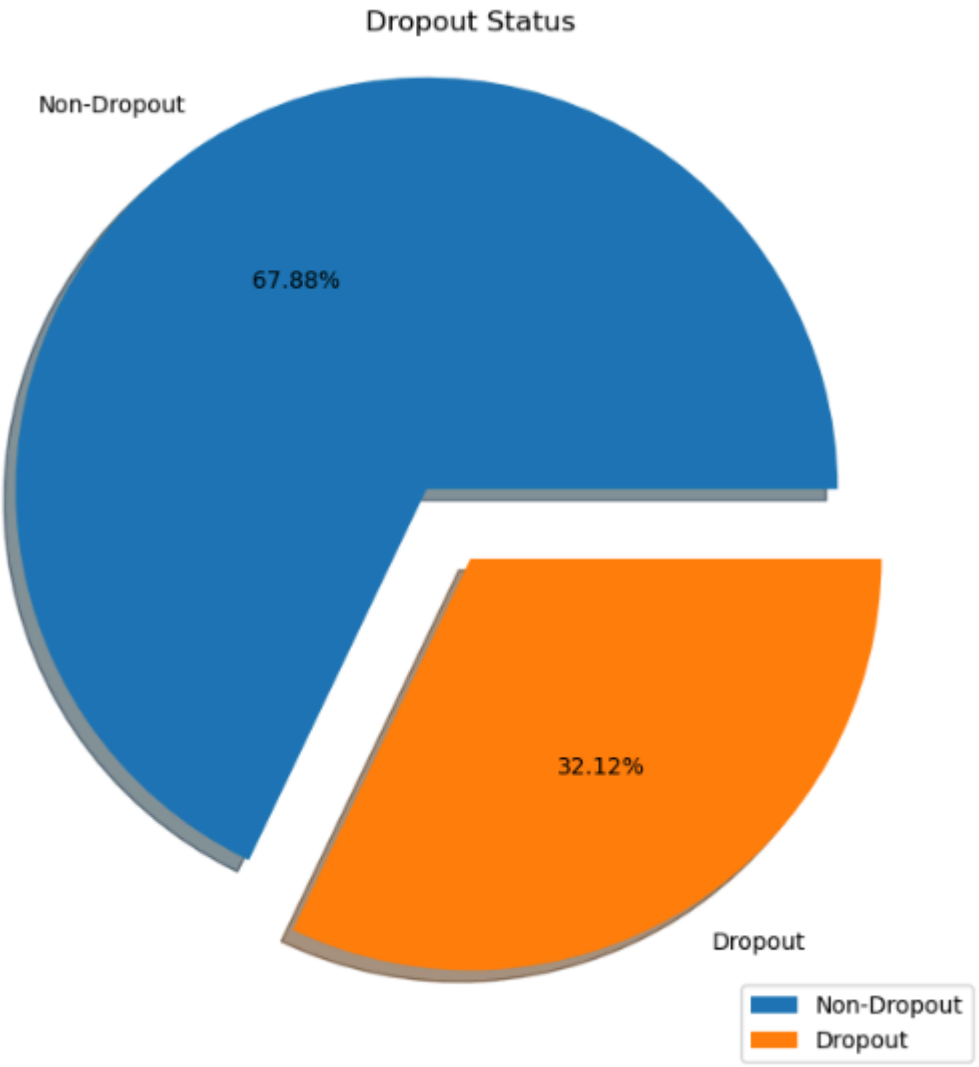


Figura 6: Gràfic "Dropout" vs "No dropout" - PIP

Veiem que tenim més casos d'alumnes graduats que d'alumnes que han abandonat. Concretament tenim 3003 "No Dropout" i 1421 "Dropout". Això ens fa veure que un terç dels alumnes que figuren dins aquest conjunt de dades han abandonat, el qual és una xifra considerablement alta.

A continuació el que farem serà un "Violin plot" de la variable "Target" amb la variable "Age at enrollment", per a visualitzar de quina manera afecta l'edat al apuntar-se a una carrera a l'hora d'abandonar o no:

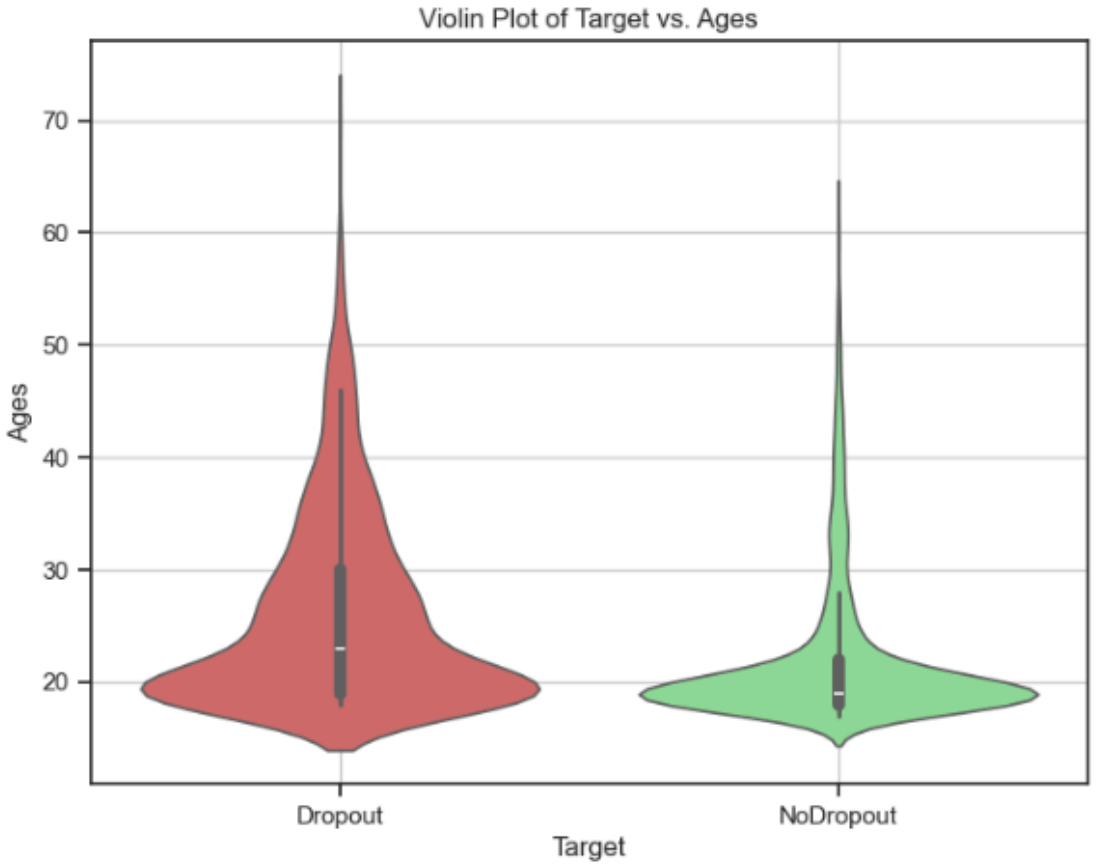


Figura 7: "Violin plot" edat vs "target" - PIP

Es veu que la majoria d'alumnes s'apunten a una carrera apoximadament als 18-20 anys. D'aquests rang d'edats, la proporció de "Dropout" i "No Dropout" és aproximadament la mateixa, seguint els percentatges totals d'abandó que s'han vist. La part interessant ve quan l'edat va pujant. A partir dels, aproximadament, 25 anys, la proporció d'abandó és

significativament superior a la de “No Dropout”. El que s’ha de tenir en compte en aquest gràfic, és que hi ha molts més casos d’alumnes al rang de 18-22 anys al dataset que alumnes amb edats superiors a aquest rang.

A continuació s’observarà com afecta el gènere a l’hora de abandonar la universitat. Això ho farem amb un “Bar plot”:

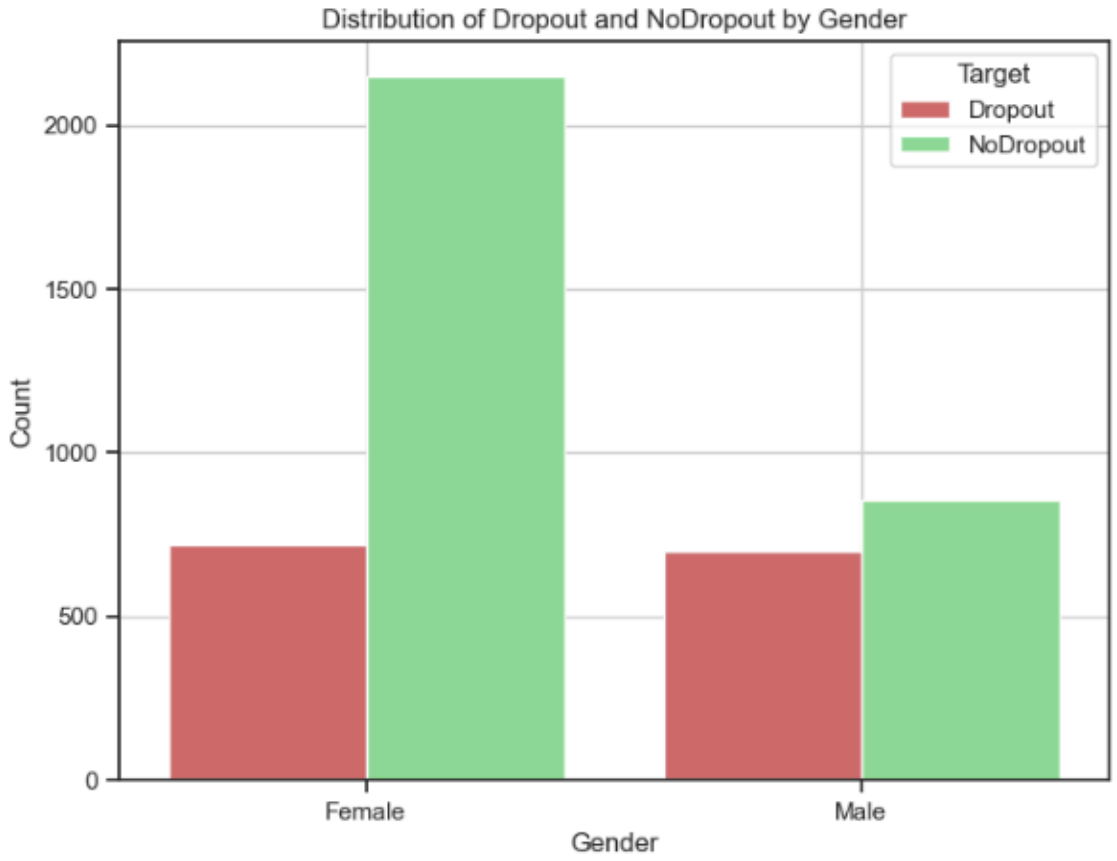


Figura 8: “Barplot” de gènere vs “Target” - PIP

Aquest gràfic és força interessant. En el cas dels alumnes masculins, el percentatge de “Dropout” és pràcticament del 50%. Per al gènere femení, en canvi, tenim uns 700 casos d’alumnes que han abandonat comparat amb més de 2000 casos d’alumnes que s’han graduat. Això dona un percentatge d’abandó d’aproximadament 30%, significativament inferior que en el cas d’alumnes masculins.

En el següent gràfic, també un “Bar Plot”, es veurà com afecten els pagaments de matrícula per saber si un estudiant abandonarà. S’emprarà la variable “Tuition fees up to date”, que recordem, és una variable booleana que indica si té els pagaments en ordre:

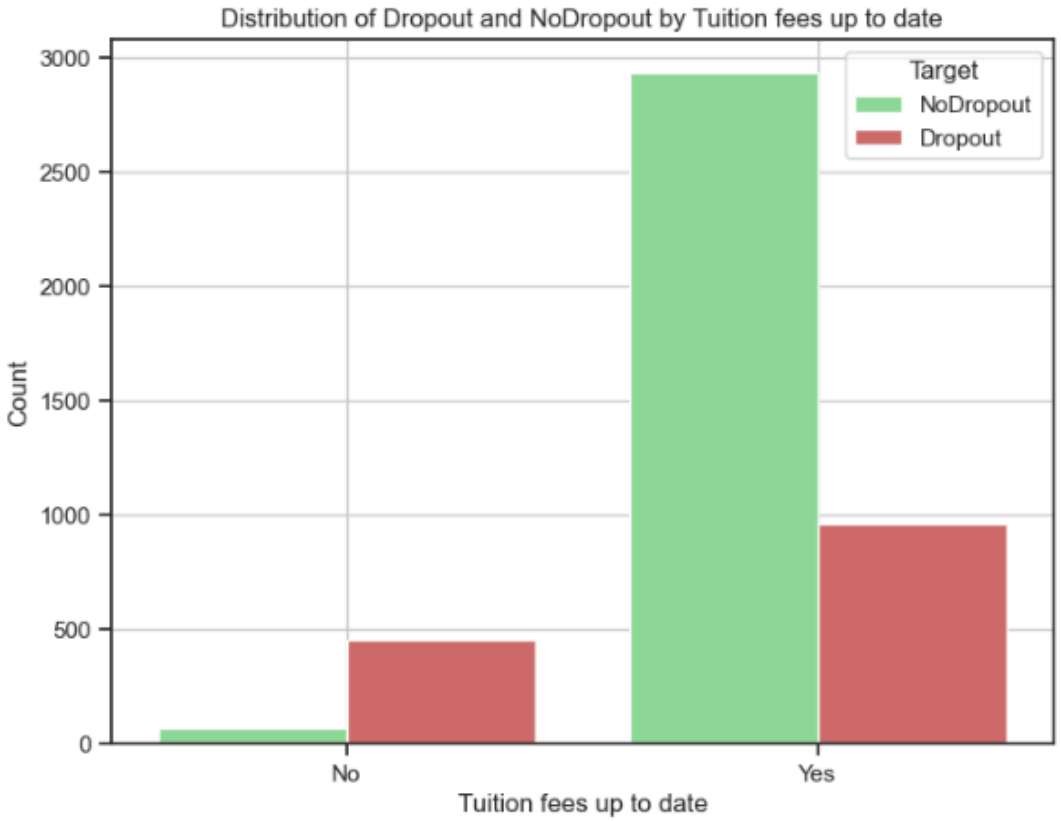


Figura 9: “Bar plot” Pagaments vs “Target” - PIP

Es pot apreciar que quan un alumne té els pagaments de matrícula en ordre, la distribució de “Dropout” vs “No Dropout” segueix pràcticament l’ordre de tot el dataset, no influeix molt. L’interessant ve quan un alumne no té els pagaments de la matrícula en ordre (recordem que aquest dataset té dades recollides després del primer any de carrera). Quan l’alumne no s’ha pogut permetre pagar la matrícula pràcticament tots abandonen el grau.

Durant aquesta secció s'han analitzat les variables més rellevants del dataset del PIP que no estàn disponibles a les dades de la UB. Afegir aquestes variables a les futures dades que es recolleixin a la UB, en prinicpi, no suposaria cap problema, ja que son dades que la UB té disponibles però no inclou als conjunts de dades acadèmiques. Com a conclusions d'aquest anàlisi podem extreure que aquestes variables tenen una gran rellevància a l'hora de saber si un alumne abandonarà el grau o no, i per això sembla important tenir-les en compte en el futur per així poder fer models de predicció més consistents.

Finalment s'ha estudiat com les diferents carreres que aparèixen dins aquest dataset afecten al resultat de l'alumne:

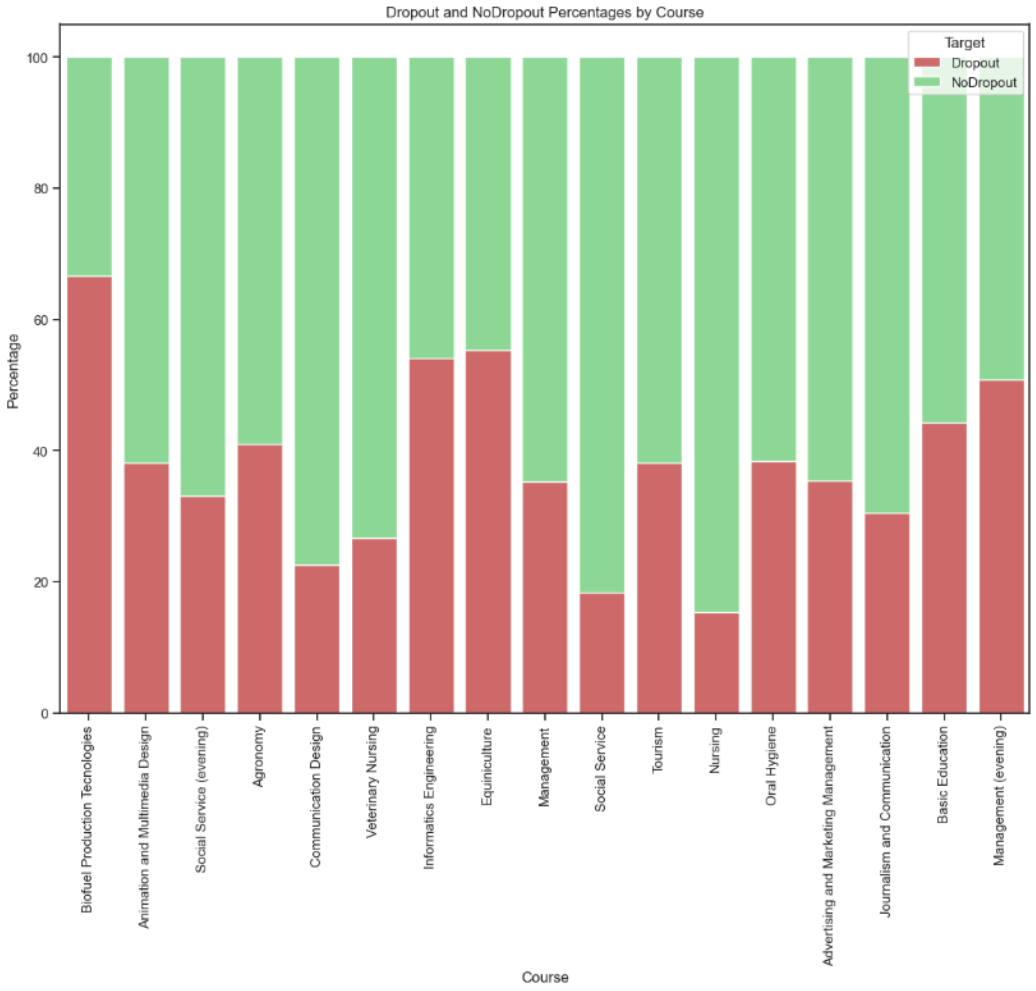


Figura 10: "Dropout" per assignatura - PIP

Es troba que encara que la correlació sigui força baixa, si que hi ha diferències entre cursos, haguent carreres més i menys propenses a l'abandonament. El motiu d'aquestes correlacions tan baixes, és que els graus que més dispersen de la proporció total d'abandó, disposen de considerablement menys mostres que les que segueixen aquesta proporció.

Ara que hem fet un anàlisi intensiu d'aquestes dades podem començar a plantejar el problema del model de Machine Learning. Ens hem fet una idea de les correlacions, de quines són més importants i hem visualitzat les dades per a tenir una idea de com són aquestes i com influeixen en els resultsts. Per tant, estem preparats per a començar a construïr el model de predicció. Abans de fer-ho, es veurà l'anàlisi que s'ha fet amb les dades de la UB.

3.3.2 Anàlisi de les dades de la UB

Com s'ha comentat abans, els conjunt de dades de la UB que hem preparat disposa de 8 columnes: "ID alumne", "ID ensenyament", "Estat", "Ensenyament", "Semestre", "Curs", "AVG qualificacio" i "Num. suspeses". Aixó doncs, l'anàlisi d'aquestes dades és més escàs que el de les dades del PIP, ja que disposem de bastantes menys variables.

3.3.2.1 Correlacions

El primer que es farà serà, com abans, estudiar les correlacions d'aquestes dades. Ja hem vist abans el significat de "correlació", per tant passem directament a veure com son aquestes.

Les correlacions d'aquestes variables amb la variable "Estat" son les següents:

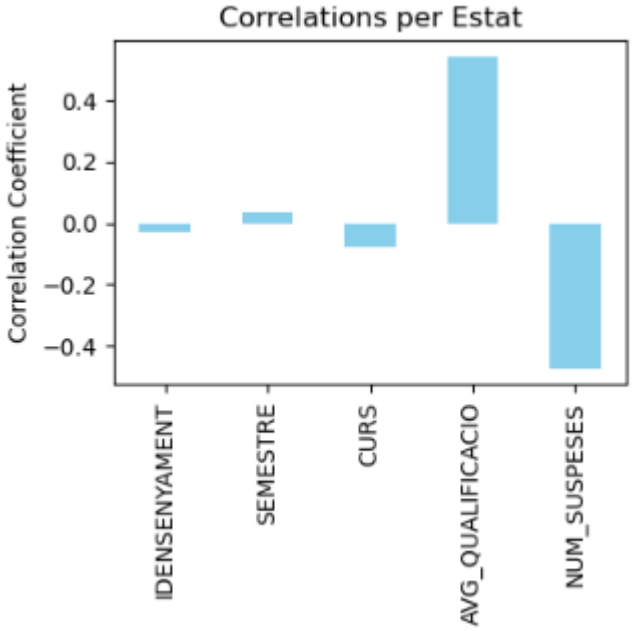


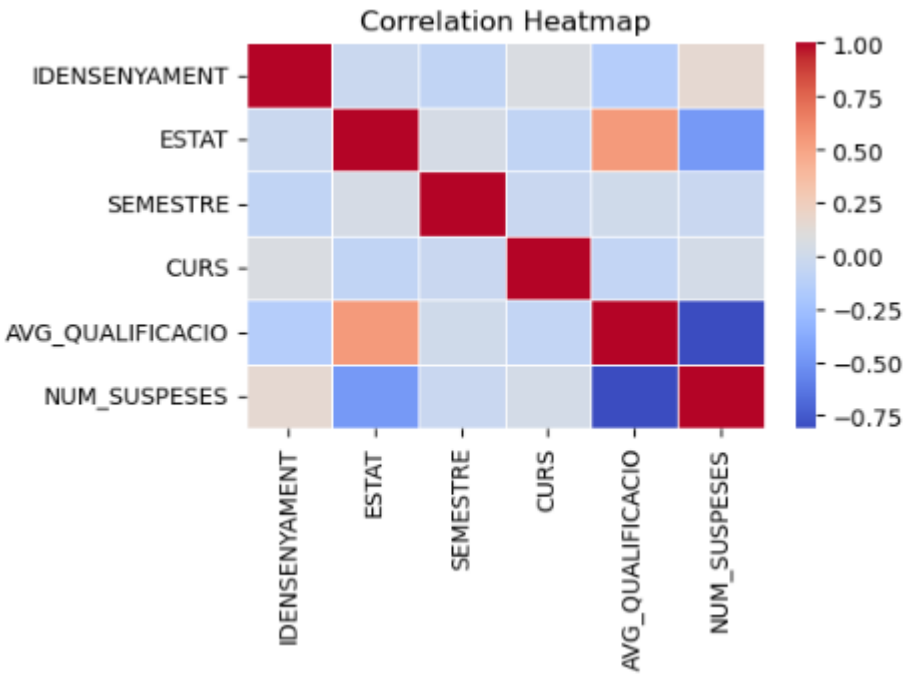
Figura 11: Correlacions amb “Estat” - UB

En aquesta gràfica s’aprecia que les correlacions més fortes, igual que amb el PIP són per part de les variables que tracten dades de qualificacions. Concretament les correlacions són les següents:

```
IDENSENYAMENT      -0.028217
SEMESTRE            0.038879
CURS               -0.078743
AVG_QUALIFICACIO    0.544377
NUM_SUSPESES       -0.475795
```

Les variables “ID ensenyament”, “Semestre” i “Curs” tenen correlacions de menys de 0.1, el qual és bastant baix. Les altres dues en canvi tenen una correlació significant amb “Estat”, per tant seran les més rellevants s l’hora d’entrenar el nostre model.

També s’ha vist com es correlacionen aquestes variables entre si, el qual ens dona el següent resultat:



Figuta 12: Mapa de calor - UB

Es veu que les correlacions més fortes, a banda de les que acabam d eveure, es donen entre "Num_Suspeses" i "Avg_qualificaio", com era d'esperar. La resta no semblen ser gaire dependents entre si.

3.3.2.2 Visualització de les dades per estat final

Un cop estudiades les correlacions, anem a visualitzar i analitzar una mica el conjunt de dades. Hem seguit el mateix procediment que per a les dades del PIP. El primer que s'ha fet és veure la proporció d'abandó total del nostre conjunt de dades:

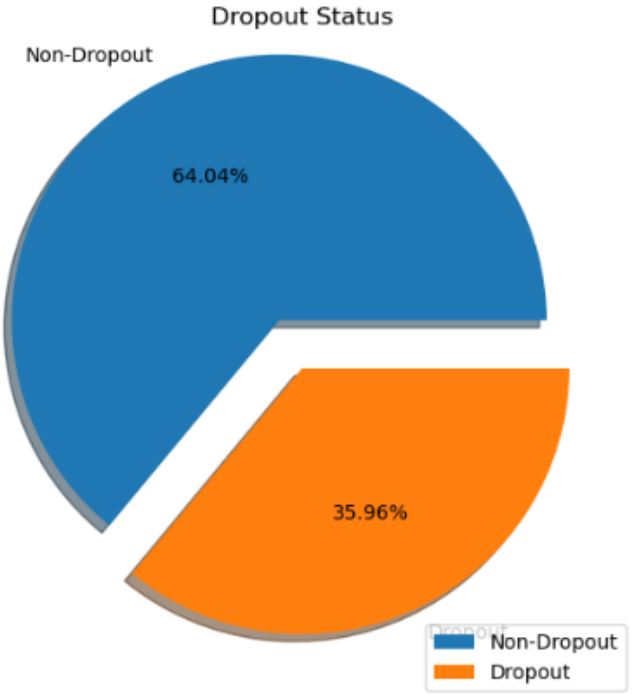


Figura 13: Proporció d'abandó - UB

Com es pot apreciar, aquesta gràfica té uns valors molt similars als del PIP, on els casos de "No dropout" son aproximadament 2 terços de totes les dades. La xifra de "Dropout", per tant, segueix sent bastant elevada.

Pel que fa a la variable “Curs”, s’ha fet un recompte dels cursos dels quals tenim dades, encara que no sigui gaire rellevant per el nostre futur model, i ens dona el següent:

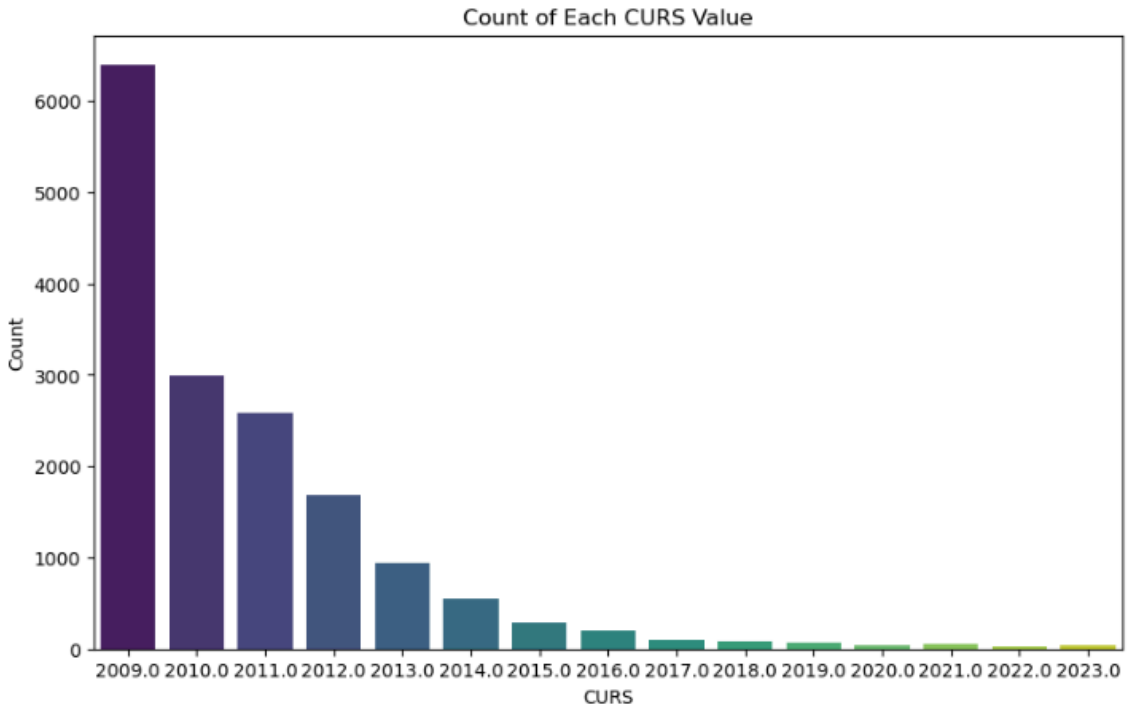


Figura 14: “Barplot” variable “Curs” - UB

Es veu que la majoria de dades que tenim disponibles son dels cursos 2009-2012. De la resta no s’ha pogut obtenir la variable “Estat”, encara que es disposés de la resta de dades, com s’ha explicat abans.

Referent als ensenyaments disponibles es ténen els següents:

ENSENYAMENT	
Dret	2009
Administració i Direcció d'Empreses	1989
Mestre d'Educació Primària	1481
Biologia	881
Relacions Laborals	850
Història	809
Història de l'Art	790
Química	746
Economia	568
Mestre d'Educació Infantil	555
Belles Arts	516
Física	435
Ciències Biomèdiques	405
Gestió i Administració Pública	400
Filosofia	399
Treball Social	357
Antropologia Social i Cultural	347
Enginyeria Informàtica	336
Matemàtiques	308
Educació Social	299
Pedagogia	259
Ciències Ambientals	247
Criminologia	190
Bioquímica	182
Enginyeria Química	168
Biotecnologia	147
Ciències Polítiques i de l'Administració	146
Geografia	126
Sociologia	117
Estadística	94
Conservació-Restauració de Béns Culturals	69
Arqueologia	52
Enginyeria Electrònica de Telecomunicació	39
Disseny	37
Enginyeria de Materials	18
Empresa Internacional	2

Es disposa d'un total de 36 ensenyaments, encara que s'ha de tenir en compte que per alguns d'aquests tenim molt poques mostres. Aquest fet influeix al resultat del gràfic que s'ha fet referent als ensenyaments, equivalent al que s'ha mostrat abans amb les dades del PIP:

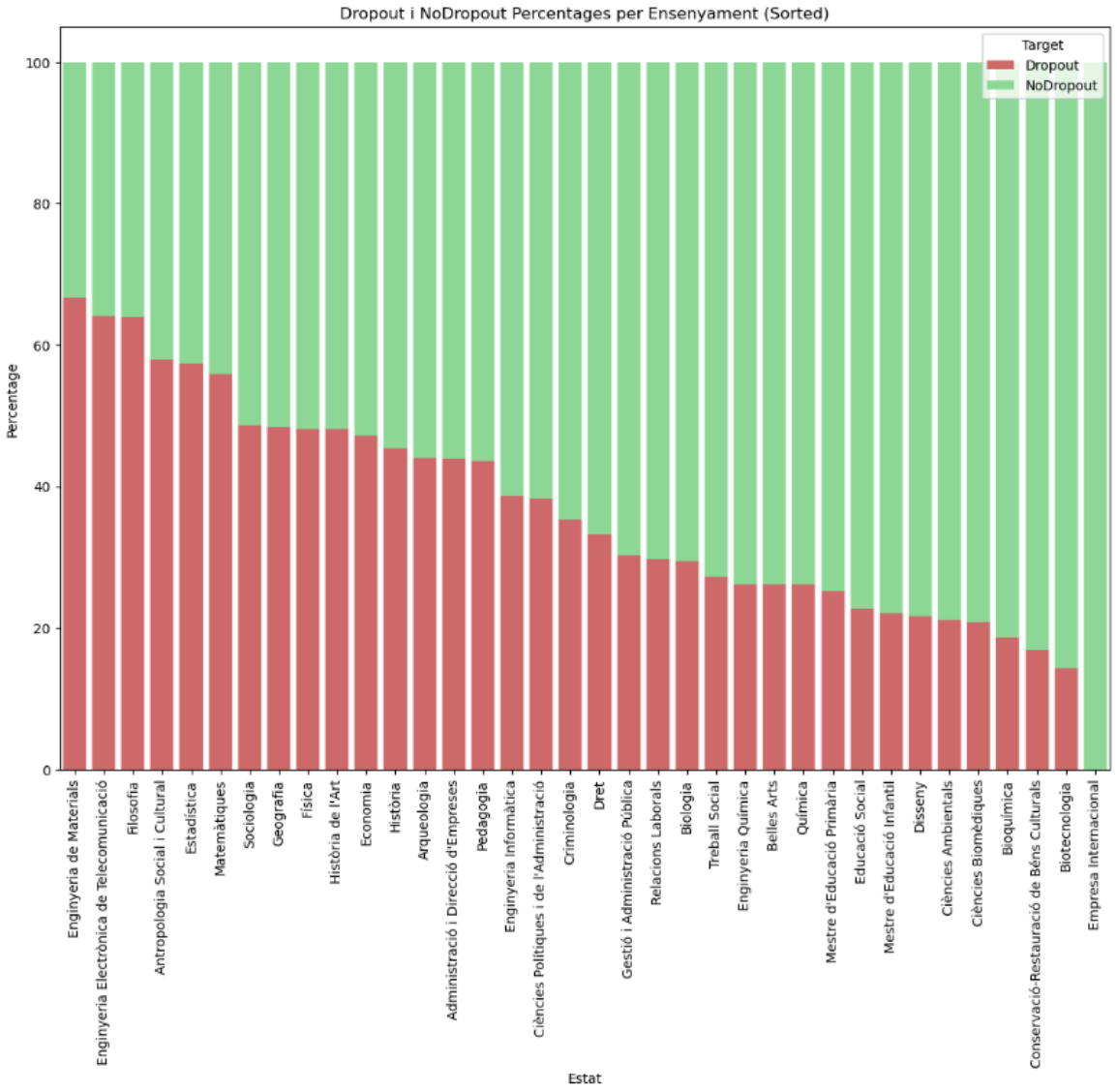


Figura 15: “Dropout” per assignatura - UB

Aqui es pot apreciar que les columnes que més difereixen dels resultats esperats (diguem que els resultats esperats són els valors de “Estat” que tenim, 64% “No dropout” contra 36% de “dropout”) son majoritàriament per les que tenim pocs exemples. Els casos més clars d’això son els dels graus “Empresa internacional” i “Enginnyeria de materials”, situats cadascun a un extrem del gràfic, i que disposen de 2 i 18 exemples respectivament. Amb aquest fet es podria explicar la baixa correlació amb “Estat”, ja que les que més aparèixen

com “Dret”, “Administració i direcció d’empreses” o “Mestre d’educació primària” estàn més propers al percentatge explicat abans.

A continuació, una vegada acabats aquests anàlisis, estudiarem diferents models de predicció per veure quins són els més adequats per emprar. Després d’això s’implementaràn models de predicció per als dos conjunts de dades amb l’objectiu de comparar els resultats i com afecta el tenir més variables rellevants a un dataset.

4. Tècniques d'aprenentatge automàtic per la predicció de l'abandonament

Abans de començar amb els experiments del futur model de predicció cal analitzar diferents models d'entrenament per a veure quin serà el més adequat per al nostre problema. S'analitzaràn breument els algorismes més coneguts i emprats en la ciència de dades, es veuràn els avantages i desavantatges de cadascún i en base a això s'elegirà un model, el qual s'estudiarà amb més detall i amb el qual farem els nostres experiments. Aquest procés de selecció es duu a terme degut al teorema "No Free Lunch" [12]. Aquest ens diu que no hi ha cap algorisme de "Machine Learning" que funcioni millor per a tots els problemes, i és especialment rellevant per a l'aprenentatge supervisat, que és el que tenim nosaltres. Supervisat es diu als problemes dels quals tenim les seves dades etiquetades, com en el nostre cas son les variables "Target" o "Estat".

S'han estudiat els avantatges i desavantatges dels següents algorismes: Regressió logísitica, Naive Bayes, Random Forest i XGBoost.

4. 1 Regressió logística

La regressió logística és un algorisme de classificació que s'empra per trobar la probabilitat d'un event. Es sol emprar en problemes que tenen resultats binaris, com és el nostre cas (Dropout, No Dropout). Els punts forts i dèbils més importants son els següents [13]:

Avantatges:

- Fàcil d'implementar i interpretar, a més de tenir alta eficiència.

- Funciona bé quan el conjunt de dades és linealment separable. Aquest concepte significa que podem “dibuixar una línea recta” entre les variables que son, en el nostre cas “Dropout” i “No Dropout”.

- Poca possibilitat de “Over-fitting” en datasets amb poques dimensions.

Desavantatges:

- Assumeix la linearitat entre la variable dependent i les independents, és a dir entre el “Target” i la resta. En el nostre cas tenim bastanta linearitat entre les dades de resultats acadèmics com les notes i el “Target”, però no tanta en la resta de variables.

- No s’ha d’emprar si el nombre d’observacions (files o valors) és menor que el nombre de característiques (variables o columnes). Per nosaltres, no és el cas.

- Només pot ser emprada per prediure funcions discretes. Per la predicció de dades contínues no és convenient. Tampoc és el nostre cas.

4.2 Naive Bayes

Naive Bayes és un classificador basat en el teorema de Bayes. Els punts forts i dèbils són els següents [14]:

Avantatges:

- Dona resultats molt bons si les variables del conjunt de dades son independents.

- Fàcil d’implementar ja que només es calcula la probabilitat.

- Funciona bé amb datasets de dimensions altes, és a dir, si “Target” pot tenir molts valors.

Desavantatges:

- Si les variables no són independents, no funciona gaire bé.

- S’ha de fer un procés de “Smoothing” si la probabilitat d’un resultat és cero.

- Pot tenir “Vanishing value”. Això vol dir que un valor molt petit “desapareix” per el fet de ser tan petit, quan s’hauria de tenir en compte.

4.3 Random Forest

Aquest algorisme és un model d’aprenentatge que empra múltiples arbres de decisió per fer les prediccions. Pot ser emprat per classificació i regressió. El que fa és crear un gran nombre d’arbres amb diferents mostres aleatories de les dades. S’entrenen tots aquests arbres i s’agrega el resultat de tots per obtenir el resultat final. Els avantatges i desavantatges són els següents [15]:

Avantatges:

- Funciona bé per relacions no lineals i complexes entre les característiques i la variable “Target”.

- Pot manejar dades numèriques i categòriques sense problemes, valors nuls, i no necessita escalar els valors.

- Dona informació sobre la importància de cada característica.

- Funciona amb datasets grans i amb altes dimensions.

Desavantatges:

- La interpretabilitat dels resultats és molt baixa.

- Pot tenir alts costos computacionals.

- Sensible a dades amb soroll.

4.4 XGBoost

Aquest és un algorisme que implementa arbres de decisió potenciats amb el gradient, dissenyat per tenir bona velocitat i rendiment que tracta de obtenir un resultat a través de combinar resultats d'un conjunt de models més simples. Els avantatges i desavantatges són els següents [16]:

Avantatges:

- Disposa d'una regularització que impideix l'overfitting.
- Treballa bé amb valors faltants.
- Paral·lelització, bon rendiment.

Desavantatges:

- Si els parametres no estàn ben ajustats, si que es posible l'overfitting.
- Per datasets grans pot trigar molt.

4.5 Elecció de model d'aprenentatge

Un cop hem vist els avantatges i desavantatges dels diferents models estudiats hem de decidir quin estudiarem més en detall i emprarem per al nostre problema. Per a fer això tindrem en compte les següents característiques de les nostres dades:

- Tenim conjunts de dades relativament petits, un amb 4000 files i 34 columnes i un altre amb 16000 files i 7 columnes. Això fa que no sigui molt important el rendiment de cada model a l'hora d'elegir-lo.
- Mescla de dades categòriques i numèriques. Això "suma punts" per al Random Forest.
- Variable "Estat" binària. Aquest fet fa que la regressió logística sigui bona elecció.
- Les relacions entre característiques són linears en alguns casos i en altres no. Les més importants, que son les variables de "notes", tenen alta linealitat, però tota la

resta, en les dades del PIP, no molta. Això fa que per el primer cas regressió logística s'adapti bé, i en el segon el Random Forest sería l'adequat, ja que tenim una mescla, Si totes fóssin independents, aniria bé Naive Bayes, però no és el cas.

- Volem tenir una bona interpretabilitat dels resultats, ja que és interessant que siguin fàcils d'analitzar. Per aquest cas, regressió logística té el millor comportament.

Tinguent això en compte, ens en podem adonar que el model que més compleix amb les nostres expectatives és el de Regressió logística. Ésta dissenyat específicament per problemes amb resultats binaris, té resultats interpretables, bon rendiment, no tenim relacions amb poca linearitat a les variables importants. Anem a estudiar amb més detall doncs com funciona el model de Regressió logística. També cal mencionar que un experiment realitzat al Kaggle [19] demostra que tant la regressió logística com el "Random forest", que ens sembla el segon més convenient, obtenen resultats molt semblants en quant a precisió, per tant per temes de rendiment i interpretabilitat la regressió logística serà l'algorisme més adequat.

4.6 Model escollit vist en detall

La regressió logística es defineix com a un algorisme d'aprenentatge automàtic supervisat que realitza tasques de classificació binària a través de prediure la probabilitat d'un event [17]. El model analitza la relació entre una o més variables independents i classifica dades a classes discretes.

Aquest algorisme empra una funció logística anomenada "sigmoid function" per mapejar prediccions i les seves probabilitats. Aquesta funció es visualitza com una curva en forma de S que converteix qualsevol valor real a un rang d'entre 0 i 1.

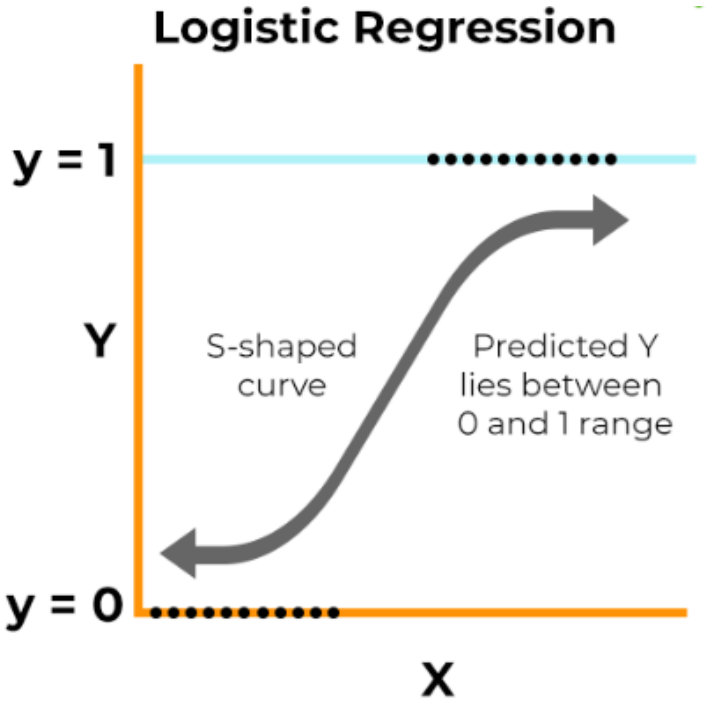


Figura 16: Regressió logística

Si el resultat de la funció és major que un llindar predefinit, el model prediu que la instància pertany a una classe. Si el resultat és més petit que el llindar, la instància no pertany a aquesta classe. L'equació d'aquesta funció és la següent:

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

On la e es refereix a la base de logaritmes naturals i el valor x al valor numèric que es vol transformar al rang de 0 a 1.

L'equació de la regressió logística és la següent:

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

On x és el valor que es reb, y la predicció, b0 un valor de cal·libració i b1 el coeficient per x.

4.7 Tècnica de validació

El que també s'ha tengut en compte a l'hora de l'entrenament és com separarem aquestes dades en conjunts d'entrenament i testeig. La resolució tradicional és separar les dades en un 80% per entrenar i un 20% per testejar. El problema d'això, és que a l'hora de fer l'entrenament no es té en compte aquest 20%, el qual pot suposar problemes de pèrdues de dades importants, models mal entrenats o resultats incorrectes. Per a arreglar això s'ha decidit emprar una altra tècnica, "K-fold cross validation" [20].

"K-fold cross validation" és un mètode emprat per separar les dades d'una manera que totes les dades siguin emprades per entrenar i testejar. El que fa és primerament mesclar les dades aleatòriament. Després separa la taula de dades en k grups. Un cop separades, entrena i testeja les dades k vegades, emprant un grup com a testeig cada cop. Després suma tots els resultats obtinguts als tests, i fa una mitjana per obtenir al precisió. En el nostre cas, hem emprat un valor de k=5.

Ara que s'han escollit les tècniques que s'empraràn per a l'entrenament de les dades, és hora de posar la feina feta en pràctica i començar amb els experiments, vegent quin rendiment podem treure dels nostres models i quines variables es podrien afegir a les dades de la UB per tal de millorar aquestes i ajudar a futurs models de predicció d'abandonament universitari.

5. Experiments i resultats

Un cop preparades i analitzades les dades i els models d'entrenament, ja estem preparats per a començar amb els experiments i a crear un model d'aprenentatge per Machine Learning que ens ajudi a saber si un estudiant abandonarà els seus estudis després del primer any cursat. Comparant els resultats entre els dos conjunts de dades disponibles veurem si realment les variables que no es tracten resultats acadèmics ajuden a prediure el comportament de l'estudiant, i es faràn decisions i propostes sobre quines variables caldria afegir en un futur a les dades de la Universitat de Barcelona.

Es crearan diversos datasets, escollint diferents subconjunts de característiques a partir de les originals, amb els quals farem experiments per tal de trobar-ne el subconjunt de característiques més eficient per predir l'abandonament.

5.1 Experiments sobre el conjunt de dades PIP

A continuació es mostraràn els diferents experiments que s'han fet amb les dades disponibles del PIP. Cal recodar que l'objectiu d'aquests experiments no és aconseguir un model de predicció amb la màxima precisió possible, sino que el que es vol és analitzar com afecten les diferents dades al rendiment del nostre model. És per això que el que s'ha fet no és provar diferents algorismes per a un conjunt de dades, sino que el que s'ha fet és provar un mateix algorisme per a diferents subconjunts de dades.

5.1.1 Subconjunts de les dades PIP

El primer que s'ha fet és separar el conjunt de dades per diferents subconjunts basats en la correlació de les variables amb la variable "Target". Per a fer això hem fet 4 subconjunts del

nostre conjunt principal. Aquests són els següents:

- El primer amb el top 10 de variables amb més correlació → "Top10"
- Un amb les variables amb una correlació major que 0.01 → "0.01"
- Un altre amb les variables amb una correlació major que 0.05 → "0.05"
- Per útlim, un altre amb les variables amb una correlació major que 0.1 → "0.1"

Al capítol 4.2.1 Anàlisi dades PIP - correlacions es poden trobar aquestes per a veure quines són les variables que s'han quedat a cada subconjunt.

Després s'ha creat una taula simulant el conjunt de dades que tenim a la UB. Aquesta taula tindrà només les variables equivalents a les de la UB de "Ensenyament", "Num_Suspeses" i "Avg_qualificacio", que en el conjunt de dades del PIP són: "Course", "Curricular units 1st sem (approved)", "Curricular units 2nd sem (approved)", "Curricular units 1st sem (grade)", "Curricular units 2nd sem (grade)". Aquest subconjunt de dades l'anomenarem "Marks".

Finalment s'ha creat un últim subconjunt de dades, que per als nostres objectius és de gran importància. L'anomenarem "Extras", i tracta del subconjunt descrit anteriorment, "Marks", que és l'equivalent a la UB, però afegint les variables que s'han estudiat abans com a possibles variables per a afegir a la UB. Aquestes son "Gender", "Age at enrollment", "Debtor" i "Tuition fees up to date". S'analitzaràn les diferències de rendiment entre el dataset "Marks" i aquest per simular com realment afectaria afegir aquestes variables a les dades de la UB.

5.1.2 Resultats dels experiments PIP

La visualització d'aquests experiments es farà amb una taula que resumeixi els resultats, per així tenir una visió clara i ràpida d'aquests resultats. A l'annex d'aquest document, es podràn veure visualitzacions més detallades, amb les respectives matrius de confusió de cada experiment.

Els resultats han estat els següents:

Acc.: Accuracy o exactitud

Des. T.:Desviació Típica

Temps: Temps que ha trigat per l'entrenament en segons

DP: Dropouts predits

DMP: Dropouts mal predits

%D: Percentatge encert dropouts

NDP: No Dropouts predits

NDMP: No Dropouts mal predits

%ND: Percentatge encert no dropouts

Dropouts totals: 1421 i No dropouts totals: 3003

Dades	Acc.	Des. T.	Temps	DP	DMP	%D	NDP	NDMP	%ND
Totals	0.8749	0.0075	1.22	1037	384	72.97	2834	169	94.37
Top10	0.8517	0.0082	0.29	984	437	69.24	2784	219	92.7
0.01	0.8759	0.01	0.83	1039	382	73.11	2836	167	94.43
0.05	0.872	0.01	0.64	1007	414	70.08	2851	152	94.49
0.1	0.8736	0.007	0.49	1006	415	70.79	2859	144	95.2
Marks	0.8239	0.01	0.08	867	554	61.01	2778	225	92.5
Extras	0.849	0.01	0.09	980	441	68.96	2776	227	92.44

Taula 5: Resultats - PIP

5.1.3 Anàlisi dels resultats PIP

Veiem una taula amb resultats que semblen bastant interessants. Anem primer a analitzar els resultats del dataset total comparat amb els datasets reduïts segons la correlació. Els conjunts de dades amb millors resultats en quant a accuracy i desviació típica són els que més variables tenen, el dataset complet i el dataset "0.01". Aquest és un punt important a tenir en

compte tenint en compte els objectius d'aquest projecte d'analitzar com afectaría afegir més variables a les dades de la UB.

Un altre fet que es pot apreciar és que el percentagge d'encert dels "No dropout" és significativament major que el percentatge d'encert dels "Dropout". Principalment això es deu a que les dades no estàn del tot balancejades, i tenim bastants més casos de "No dropout" que de "Dropout".

A banda d'això, hi ha un altre punt interessant en aquests resultats. Encara que l'accuracy general millori com més variables tenim, podem observar un comportament contrari en la predicció de "No dropout". Com es pot veure, el dataset que millor rendiment té en aquest sentit és el "0.1" amb un 95.2% d'encert, a diferència del 94.37% d'encert del dataset amb totes les dades. És una diferència poc rellevant en aquest cas, però el realment interessant ve quan passem de 14 ("0.1) a 10 ("Top10") variables i aquest percentatge baixa al 92.7%. Les quatre variables que hi son a "0.1" però no a "Top10" són: "Displaced", "Curricular units 1st sem (enrolled)", "Curricular units 2nd sem (enrolled)", "Curricular units 2nd sem (evaluations)". Això implica que aquestes variables ajuden a saber si un estudiant serà "No dropout".

Però per el que es realment important, saber si un estudiant abandonarà, el model té el comportament esperat, com més dades millor. L'única excepció es dona en que el percentatge és molt lleugerament millor en el conjunt "0.01" que en el complet. Les variables que s'han tret del "0.01" són: "Course", "Nacionality", "Educational special needs". Ja que la desviació típica és també una mica major en els resultats de "0.01", es donarà per fet que aquestes variables no tenen importància i que els resultats son els mateixos tant si hi son com si no.

Anem a veure ara l'experiment més interessant d'aquest conjunt de dades, la comparació entre les dades "Marks", que son les equivalents a les dades de la UB, i les dades "Extras" que son aquestes mateixes però afegint gènere, edad, si es deudor i si té la matrícula pagada.

Com hem dit aquestes dades semblen bastant fàcils de reunir per a la UB, i els resultats són bastant millors amb elles. L'accuracy puja de 0.823 a 0.849, casi 0.03 punts. Això no sembla molt, però si mirem els percentatges la percepció canvia. El percentatge d'encert de la predicció dels "Dropout" puja d'un 61.01% a un 68.96%. Gairebé un 9%, el qual és una xifra molt bona.

Així doncs, queda clar que afegint unes poques variables a les dades de la UB, el model de predicció millora considerablement. A continuació s'analitzaràn els resultats obtinguts dels experiments que s'han fet amb les dades de la UB.

5.2 Experiments sobre el conjunt de dades UB

Amb els experiments realitzats sobre el conjunt de dades de la UB s'ha seguit el mateix patró que amb els experiments anteriors amb les dades del PIP, és a dir, s'han realitzat diferents subconjunts de dades per a un mateix algorisme d'aprenentatge de dades.

5.2.1 Subconjunts de les dades UB

Per a fer això, igual que abans s'han fet diferents suconjunts de dades per a saber quines varibles ajuden més a fer prediccions. La correlació és una bona mesura per agafar de base, però fins que no es facin els experiments no es pot saber amb certesa quines van millor que altres. S'han fet experiments amb els següents subconjunts de dades:

- El dataset al complet → "Total"
- El dataset deixant només "Avg_Qualificacio" i "Num_Suspeses" → "Notes"
- Deixant només "Avg_Qualificacio" → "Avg"
- Deixant només "Num_Suspeses" → "Suspeses"
- Un per cada semestre i un per les assignatures anuals → "semX" , "anual"

5.2.2 Resultats dels experiments UB

L'algorisme emprat igual que en les altres dades serà el de Regressió logística. Es visualitzaràn els resultats amb una taula amb el mateix format i a l'annex es podràn trobar les matrius de confusió. Els resultats són els següents:

Acc.: Accuracy o exactitud

Des. T.:Desviació Típica

Temps: Temps que ha trigat per l'entrenament en segons

DP: Dropouts predits

DMP: Dropouts mal predits

%D: Percentatge encert dropouts

NDP: No Dropouts predits

NDMP: No Dropouts mal predits

%ND: Percentatge encert no dropouts

Dropouts totals: 5772 i No dropouts totals: 10278

Dades	Acc.	Des. T.	Temps	DP	DMP	%D	NDP	NDMP	%ND
Totals	0.785	0.004	0.24	3376	2396	58.48	9233	1045	89.83
Notes	0.783	0.004	0.1	3361	2411	58.22	9219	1059	89.69
Avg	0.786	0.003	0.08	3328	2444	57.65	9286	992	90.34
Susp.	0.737	0.01	0.05	3412	2360	59.11	9471	807	92.14
sem1	0.769	0.005	0.14	1577	1236	56.06	4281	525	89.07
sem2	0.796	0.009	0.15	1703	1064	61.54	4358	485	89.98
anual	0.86	0.01	0.06	106	86	55.2	603	26	95.86

Taula 6: Resultats - UB

5.2.3 Anàlisi dels resultats UB

Es poden veure patrons semblants als que hem obtingut amb els experiments de les dades del PIP. Hi ha significativament menys encert en els "Dropout" que en els "No dropout", ja

que es tenien les dades imbalancejades de la mateixa manera.

Els resulats de “anual” no els tindrem en compte ja que el subconjunt és massa petit com per tenir credibilitat. Entre “sem1” i “sem2” si que es veu una diferència rellevant. El percentatge d’encert per als “Dropout” de “sem2” és un 4% major, el qual és bastant. Això ens indica que es més fàcil encertar un alumne amb els resultats del segon semestre del seu primer any. Aquests resultats, són inclus lleugerament majors que els de les dades totals, tinguent la meitat de samples. És un punt a tenir en compte a l’hora de treure conclusions finals.

Pel que fa a les diferències d’emprar tot el dataset o només els resultats acadèmics, es pot veure que no hi ha pràcticament diferència. Això ens indica que l’ensenyament, i l’any no tenen gaire rellevància. Es podria dir que el semestre tampoc, però ja hem comprovat al paràgraf anterior que si té importància.

Finalment podem comparar els resultats del PIP amb els de la UB. Bàsicament tindrem en compte el subconjunts del PIP “marks” i “extras”, ja que són l’equivalent a la UB i la possible millora de les dades de la UB. Ens centrarem en el percentatge de “Dropouts” ben predits que és el que ens interessa. La mitjana de dropouts ben predits entre tots els subconjunts de la UB, sense contar “anual” és de 58,51%. El dataset “marks” té un percentatge d’encert per els dropouts de 61.01%. Aquestes dues xifres son bastant semblants, per tant ens posarem en el cas de que són iguals.

Així doncs, si a la UB hi afegíssim les variables que hem afegit a “extras”, podriem arribar a una precisió per l’encert de “Dropouts” de quasi 70%. Aquesta precisió podría ser encara més elevada si només es tenen en compte dades del segon semestre, i sobretot si es crea un projecte centrat en obtenir la màxima precisió possible. Semblen unes millores considerables tinguent en compte que l’únic que s’hauria de fer és afegir variables a les dades de les quals la universitat ja disposa.

6. Conclusions

L'objectiu principal d'aquest treball ha estat fer un anàlisi de les dades disponibles al conjunt de dades de la UB en la tasca de predicció de l'abandonament d'un estudiant i proposar una millora d'aquestes dades amb altres característiques rellevants. a. Mitjançant l'estudi i comparació de dos conjunts de dades, el de la UB i el del PIP, s'han extret les característiques més importants del PIP, que en disposa de moltes més que la UB per a estudiar si valdria la pena incloure-les al conjunt de dades de la UB.

Per a fer això primer s'han recollit totes aquestes dades així com es menciona al punt *3.Obtenció de les dades.* Un cop obtingudes, s'han processat amb l'objectiu de tenir dos conjunts de dades equivalents, en la mesura del possible. S'ha transformat el conjunt de dades de la UB per fer-ho equivalent al del PIP, per així poder estudiar millor com afecten les característiques que té el PIP i no té la UB. Aquest processament s'ha fet modificant les dades de la UB en direcció a les del PIP. Fer-ho en l'altra direcció no era possible, ja que al PIP disposavem d'una mitjana de les qualificacions i, a la UB, de totes les notes, és a dir, no es poden obtenir totes les notes a partir de la mitjana però si viceversa.. Un cop processades aquestes dades, s'han analitzat amb l'objectiu de trobar quines característiques són més importants a l'hora de definir un model de predicció. Per a fer això, s'ha fet un estudi de les correlacions i s'han visualitzat diferents gràfics de característiques sobre la variable "Estat". D'aquest anàlisi se'n poden treure les conclusions de que, efectivament, hi ha característiques que semblen inlfuir en el resultat final d'un alumne a banda de les qualificacions. Les tres variables que més rellevància tenen segons aquest anàlisi de les dades, són el Gènere de l'alumne, l'edad d'un alumne al apuntar-se a un grau, i si l'alumne té els pagaments de matrícula en ordre. De les dades que es disposen de la UB, només les variables referents a qualificacions influeixen en el resultat final. La resta de variables que hi ha no semblen tenir la suficient correlació amb "Estat".

Finalment, s'han fet els experiments necessàris i els resultats reafirmen el què s'ha extret de l'anàlisi de les dades Per a fer això, primer s'ha escollit un algorisme d'aprenentatge automàtic d'entre 4 estudiats. S'ha escollit la Regressió Logística, principalment degut a la interpretabilitat dels seus resultats i que és una predicció binària, on "Estat" és del tipus si/no. S'han realitzat diferents experiments per a cada conjunt de dades. Ja que l'objectiu era analitzar la importància de les característiques en comptes d'obtenir un model de predicció el més precís possible, el que s'ha fet és fer l'experiment amb diferents subconjunts de dades emprant el mateix model d'aprenentatge automàtic, per veure com canvia el seu rendiment només canviant les dades que se li introdueixen com a input. Els resultats han confirmat que com més dades es tenen disponibles, més elevada és la precisió del model. El més interessant que s'ha extret d'aquests experiments, és el següent:

- El model de predicció entrenat amb el conjunt de dades complet del PIP (34 característiques) aconsegueix un 72.97% d'encert per als alumnes "Dropout".

- El model de predicció entrenat amb el subconjunt de dades del PIP, que consta només de les qualificacions (el que es té a la UB) aconsegueix només un 61.01% d'encert per als alumnes "Dropout".

Vol dir això, que per pujar un 12% el percentatge d'encert per als alumnes que abandonen, s'han de recollir 30 característiques de cada alumne a la UB? No, això es deu a que:

- El model de predicció entrenat amb el subconjunt de dades del PIP, que consta de les qualificacions, el gènere, l'edat i el pagament de matrícula ha obtingut un 68.96% d'encert per als "Dropout".

Per tant, afegint només aquestes tres variables, gènere, edat i pagament de matrícula en ordre, es puja el percentatge d'encert en un 8%. Això ja sembla més interessant.

Per confirmar que aquests resultats també es donarien a la UB, s'han fet experiments amb les seves dades, i com era d'esperar, el percentatge obtingut ha estat d'un 60%. Pràcticament

el mateix que l'obtingut amb el subconjunt de dades del PIP que "simula" el de la UB. Per tant, com a conclusió final d'aquest projecte, es pot extreure que:

Afegint 3 característiques a les dades de la UB: gènere, edat i pagament de matrícula en ordre es podrien millorar les prediccions de si un alumne abandonarà al voltant de un 10%.

Cal destacar que el percentatge d'encert per als "No dropout" és bastant més elevat, d'aproximadament un 90%, però el que ens interessa es predir amb exactitud si un alumne abandonarà, no si es graduarà.

Per tant, des d'aquest projecte, la proposta que es fa a la UB és afegir aquestes tres característiques mencionades al conjunt de dades acadèmiques que es recolleix anualment, per així poder ajudar a futurs projectes dedicats a construir models de predicció altament precisos i sòlids.

6.1 Treball futur

Aquest projecte es pot ampliar en un futur de diferents maneres. Encara que s'hagin extret unes conclusions i s'hagi fet una proposta interessant, amb més treball i estudis sobre aquest tema es podrien extreure més fets interessants que ajudin a entendre la importància de les característiques d'un alumne sobre els seus resultats finals de la carrera. Algunes de les tasques que es podrien realitzar són:

- Analitzar conjunts de dades de més universitats, amb variables diferents de les que es disposàven a les dades del PIP.

- Experimentar amb diferents algorismes d'aprenentatge automàtic, per descobrir si canviant aquests algorismes, canvia també la importància de les variables.

- Treballar en l'obtenció d'un model d'aprenentatge el més precís possible, per a veure si d'aquesta manera la importància de les variables augmenta o disminueix.

- Treballar amb models de dades més voluminosos, per així tenir en compte també l'eficiència dels models de predicció en termes de temps i memòria, i si encara així val la pena afegir característiques.

- Fer més subconjunts amb altres variables a banda de les mencionades per a trobar si n'hi ha més que tinguin una gran importància.

Bibliografia

[1] Esteve, M. (2022). L'11% dels estudiants abandonen la universitat a l'Estat. *Diari ARA*. <https://www.ara.cat/societat/l-11-dels-estudiants-abandonen-universitat-l_1_4303864.html> [08/10/2023]

[2] Pérez, F. y Aldás, J. (2019). *Indicadores Sintéticos de las Universidades Españolas,* Informe U-Ranking FBBVA-Ivie, <https://www.fbbva.es/wp-content/uploads/2022/06/Informe-U-Ranking-FBBVA-Ivie-2022.pdf> [08/10/2023]

[3] Fernández-Mellizo, M. (2022). *Análisis del abandono de los estudiantes de grado en las universidades presenciales en España.* Madrid: Universidad Complutense de Madrid. <https://www.universidades.gob.es/wp-content/uploads/2022/11/EAU_Informe_abandono.pdf> [08/10/2023]

[4] @The Devastator. (Febrer 2023). *Predict students' dropout and academic success*. Kaggle. <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data> [16/10/2023]

[5] Realinho, V; Machado, J; Baptista, L; Martins, M.V. (2022). Predicting Student Dropout and Academic Success. *MDPI. 7*(11), 146. <https://www.mdpi.com/2306-5729/7/11/146#app1-data-07-00146> [21/10/2023]

[6] *Pandas documentation*. Pandas. <https://pandas.pydata.org/docs/> [03/11/2023]

[7] *Chardet documentation.* Chardet <https://chardet.readthedocs.io/en/latest/> [14/12/2023]

[8] *SQL JOIN Types Explained.* (2023, 29 de novembre) Coursera. <https://www.coursera.org/articles/sql-join-types> [27/12/2023]

[9] *Matplotlib 3.8.2 documentation.* (). Matplotlib. <https://matplotlib.org/stable/index.html> [05/11/2023]

[10] Correlation. JMP statistical discovery.
<https://www.jmp.com/en_au/statistics-knowledge-portal/what-is-correlation.html>
[10/11/2023]

[11] *Seaborn documentation.* Seaborn
https://seaborn.pydata.org/examples/scatterplot_matrix.html

[12] *What is no free lunch theorem*. (2021, 25 de juny). Geeksforgeeks. Recuperado de:
https://www.geeksforgeeks.org/what-is-no-free-lunch-theorem/ [16/11/2023]

[13] Jain, A. (2020) Advantages and Disadvantages of Logistic Regression in Machine Learning. *Medium.*
<https://medium.com/@akshayjain_757396/advantages-and-disadvantages-of-logistic-regression-in-machine-learning-a6a247e42b20> [18/11/2023]

[14] Soni, A. (2020). Pros and Cons Of Naive Bayes Classifier :-. *Medium.*
<https://medium.com/@anuuz.soni/pros-and-cons-of-naive-bayes-classifier-40b67249ae8> [18/11/2023]

[15] *What are the advantages and disadvantages of Random Forest?* (2023, 3 d'octubre). AIML.com.
<https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/>
[18/11/2023]

[16] Parasaniya, V. XGBoost. Machine Learning Book.
<https://vatsalparsaniya.com/ML_Knowledge/XGBoost/Readme.html> [18/11/2023]

[17] Kanade V. (2022) *What is Logistic Regression? Equation, Assumptions, Types, and Best Practices.* Spiceworks.com
https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/ [21/11/2023]

[18] Portal d'Estadístiques de l'Àmbit Acadèmic

http://www.ub.edu/dades_academiques/index.php

[19] Subhajeet, D. (2023) *Student dropout prediction.* Kaggle.com

https://www.kaggle.com/code/subhajeetdas/student-dropout-prediction#Logistic-Regression [19/11/2023]

[20] Brownlee, J. (2023). *A gentle introduction to k-fold Cross Validation.*

machinelearningmastery.com

https://machinelearningmastery.com/k-fold-cross-validation/ [23/11/2023]

Annex

Secció *3.3.1 Anàlisi de dades del PIP,* categorització de les diferents variables en format de categoríes numèriques:

Attribute	Values
Marital status	1—Single
	2—Married
	3—Widower
	4—Divorced
	5—Facto union
	6—Legally separated

Taula A1: Valors de “Marital status”

Attribute	Values
Gender	1—male
	0—female

Taula A2: Valors de “Gender”

Attribute	Values
Daytime/evening attendance	1—daytime
	0—evening

Taula A3: Valors de “Daytime/evening attendance”

Attribute	Values
Nationality	1—Portuguese
	2—German
	3—Spanish
	4—Italian
	5—Dutch
	6—English
	7—Lithuanian
	8—Angolan
	9—Cape Verdean
	10—Guinean
	11—Mozambican
	12—Santomean
	13—Turkish
	14—Brazilian
	15—Romanian
	16—Moldova (Republic of)
	17—Mexican
	18—Ukrainian
	19—Russian
	20—Cuban
	21—Colombian

Taula A4: Valors de Nacionalitat

Attribute	Values
Displaced	
Educational special needs	
Debtor	1—yes
Tuition fees up to date	0—no
Scholarship holder	
International	

Taula A5: Valors de variables de si/no

Attribute	Values
Application mode	1—1st phase—general contingent 2—Ordinance No. 612/93 3—1st phase—special contingent (Azores Island) 4—Holders of other higher courses 5—Ordinance No. 854-B/99 6—International student (bachelor) 7—1st phase—special contingent (Madeira Island) 8—2nd phase—general contingent 9—3rd phase—general contingent 10—Ordinance No. 533-A/99, item b2) (Different Plan) 11—Ordinance No. 533-A/99, item b3 (Other Institution) 12—Over 23 years old 13—Transfer 14—Change in course 15—Technological specialization diploma holders 16—Change in institution/course 17—Short cycle diploma holders 18—Change in institution/course (International)

Taula A6: Valors de “Application mode”

Attribute	Values
Course	1—Biofuel Production Technologies
	2—Animation and Multimedia Design
	3—Social Service (evening attendance)
	4—Agronomy
	5—Communication Design
	6—Veterinary Nursing
	7—Informatics Engineering
	8—Equiniculture
	9—Management
	10—Social Service
	11—Tourism
	12—Nursing
	13—Oral Hygiene
	14—Advertising and Marketing Management
	15—Journalism and Communication
	16—Basic Education
	17—Management (evening attendance)

Taula A7: valors de “Course”

Attribute	Values
Previous qualification	1—Secondary education
	2—Higher education—bachelor's degree
	3—Higher education—degree
	4—Higher education—master's degree
	5—Higher education—doctorate
	6—Frequency of higher education
	7—12th year of schooling—not completed
	8—11th year of schooling—not completed
	9—Other—11th year of schooling
	10—10th year of schooling
	11—10th year of schooling—not completed
	12—Basic education 3rd cycle (9th/10th/11th year) or equivalent
	13—Basic education 2nd cycle (6th/7th/8th year) or equivalent
	14—Technological specialization course
	15—Higher education—degree (1st cycle)
	16—Professional higher technical course
	17—Higher education—master's degree (2nd cycle)

Taula A8: Valors de "Previous qualification"

Attribute	Values
Mother's qualification Father's qualification	1—Secondary Education—12th Year of Schooling or Equivalent 2—Higher Education—bachelor's degree 3—Higher Education—degree 4—Higher Education—master's degree 5—Higher Education—doctorate 6—Frequency of Higher Education 7—12th Year of Schooling—not completed 8—11th Year of Schooling—not completed 9—7th Year (Old) 10—Other—11th Year of Schooling 11—2nd year complementary high school course 12—10th Year of Schooling 13—General commerce course 14—Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent 15—Complementary High School Course 16—Technical-professional course 17—Complementary High School Course—not concluded 18—7th year of schooling 19—2nd cycle of the general high school course 20—9th Year of Schooling—not completed 21—8th year of schooling 22—General Course of Administration and Commerce 23—Supplementary Accounting and Administration 24—Unknown 25—Cannot read or write 26—Can read without having a 4th year of schooling 27—Basic education 1st cycle (4th/5th year) or equivalent 28—Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent 29—Technological specialization course 30—Higher education—degree (1st cycle) 31—Specialized higher studies course 32—Professional higher technical course 33—Higher Education—master's degree (2nd cycle) 34—Higher Education—doctorate (3rd cycle)

Taula A9: Valors de "Mother's qualification" i "Father's qualification"

Attribute	Values
Mother's occupation Father's occupation	1—Student 2—Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 3—Specialists in Intellectual and Scientific Activities 4—Intermediate Level Technicians and Professions 5—Administrative staff 6—Personal Services, Security and Safety Workers, and Sellers 7—Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry 8—Skilled Workers in Industry, Construction, and Craftsmen 9—Installation and Machine Operators and Assembly Workers 10—Unskilled Workers 11—Armed Forces Professions 12—Other Situation; 13—(blank) 14—Armed Forces Officers 15—Armed Forces Sergeants 16—Other Armed Forces personnel 17—Directors of administrative and commercial services 18—Hotel, catering, trade, and other services directors 19—Specialists in the physical sciences, mathematics, engineering, and related techniques 20—Health professionals 21—Teachers 22—Specialists in finance, accounting, administrative organization, and public and commercial relations 23—Intermediate level science and engineering technicians and professions 24—Technicians and professionals of intermediate level of health 25—Intermediate level technicians from legal, social, sports, cultural, and similar services 26—Information and communication technology technicians 27—Office workers, secretaries in general, and data processing operators 28—Data, accounting, statistical, financial services, and registry-related operators 29—Other administrative support staff 30—Personal service workers 31—Sellers 32—Personal care workers and the like

33—Protection and security services personnel

34—Market-oriented farmers and skilled agricultural and animal production workers

35—Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence

36—Skilled construction workers and the like, except electricians

37—Skilled workers in metallurgy, metalworking, and similar

38—Skilled workers in electricity and electronics

39—Workers in food processing, woodworking, and clothing and other industries and crafts

40—Fixed plant and machine operators

41—Assembly workers

42—Vehicle drivers and mobile equipment operators

43—Unskilled workers in agriculture, animal production, and fisheries and forestry

44—Unskilled workers in extractive industry, construction, manufacturing, and transport

45—Meal preparation assistants

46—Street vendors (except food) and street service providers

Taula A10: Valors de “Mother’s occupation” i “Father’s occupation”

Resultats experiments

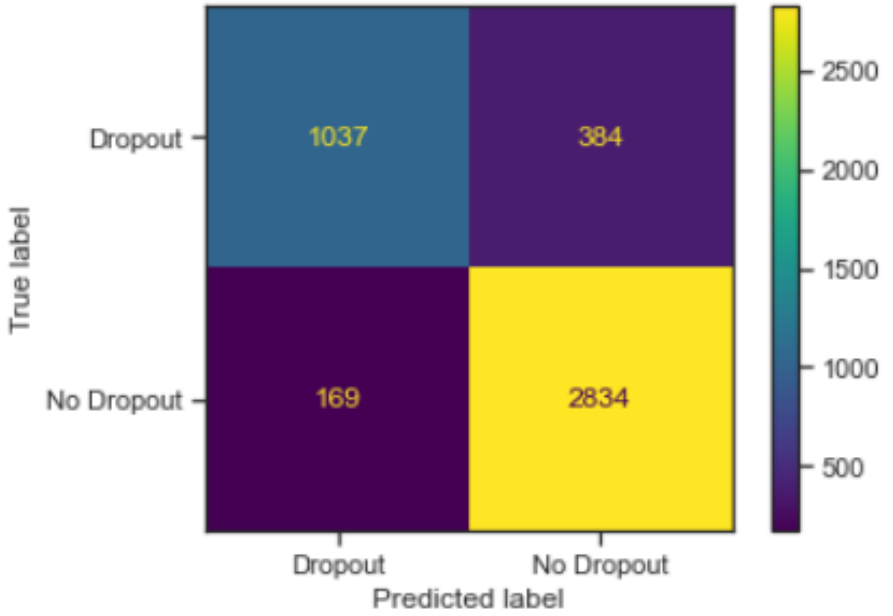


Figura A1: Resultats dataset complet - PIP

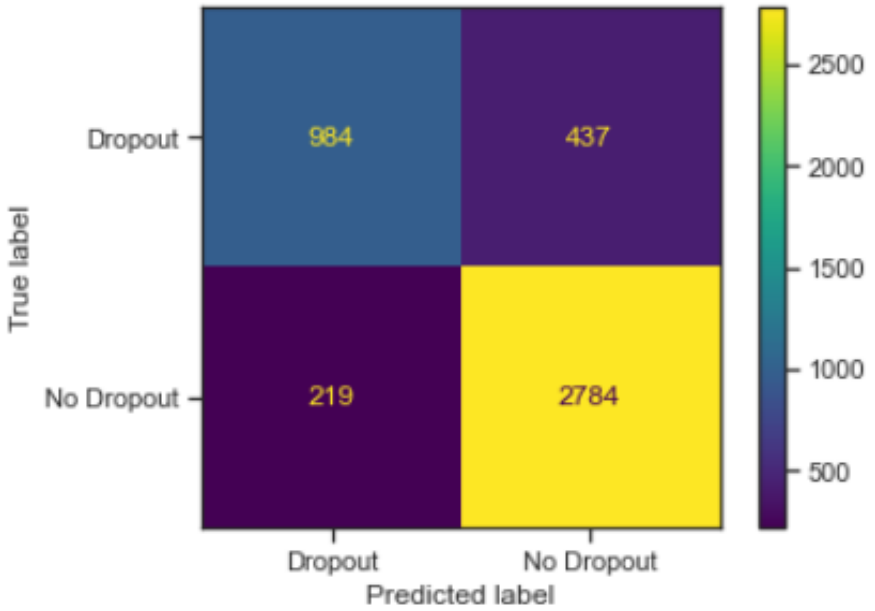


Figura A2: Resultats “top10” - PIP

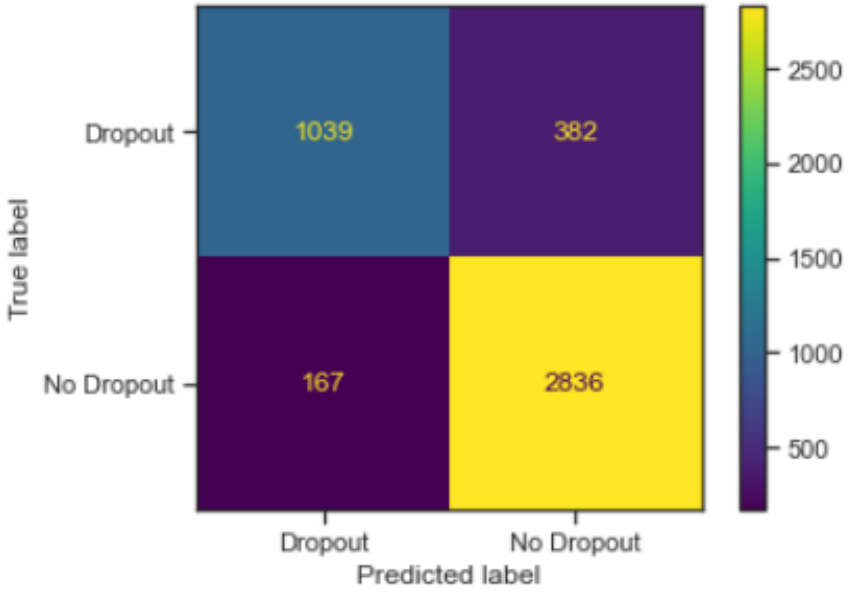


Figura A3: Resultats “0.01” - PIP

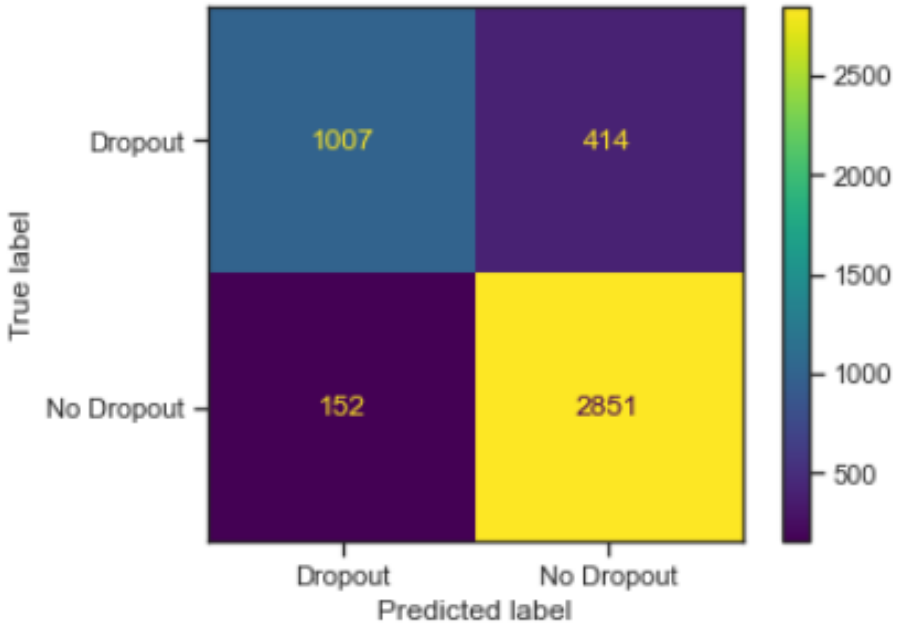


Figura A4: Resultats “0.05” - PIP

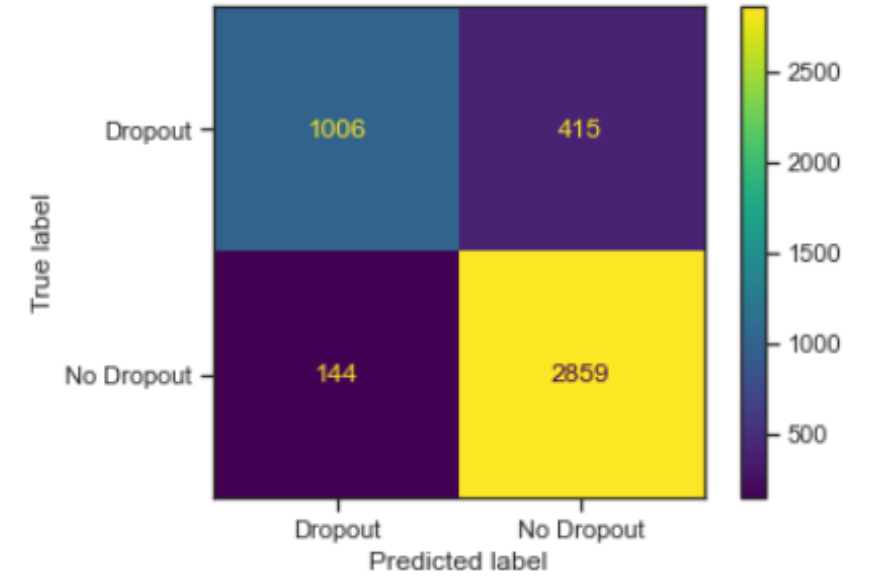


Figura A5: Resultats “0.1” - PIP

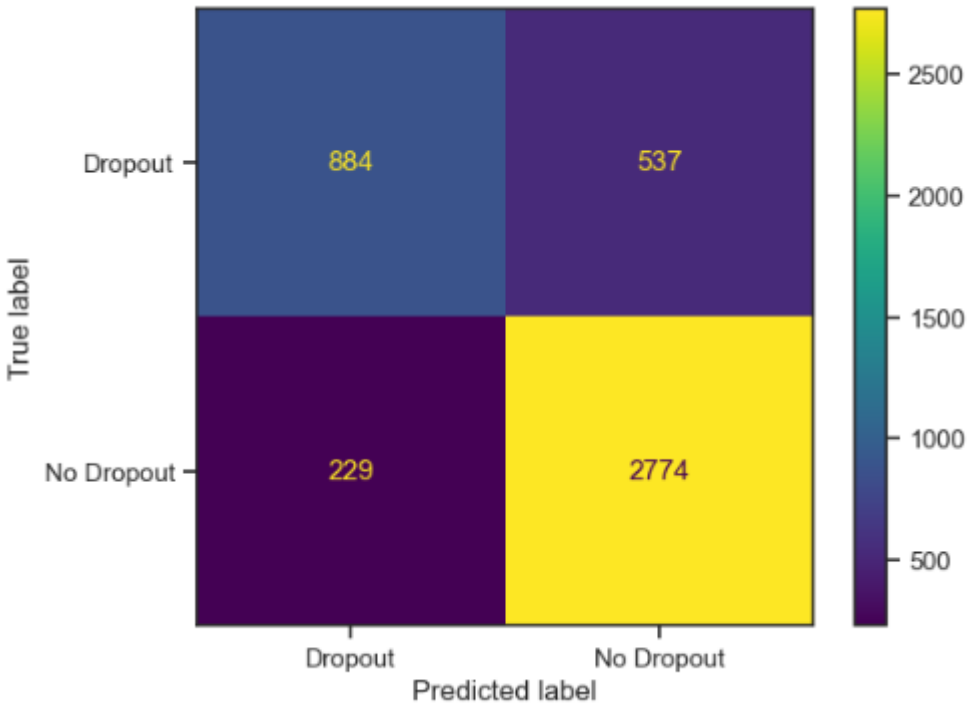


Figura A6: Resultats “Marks” - PIP

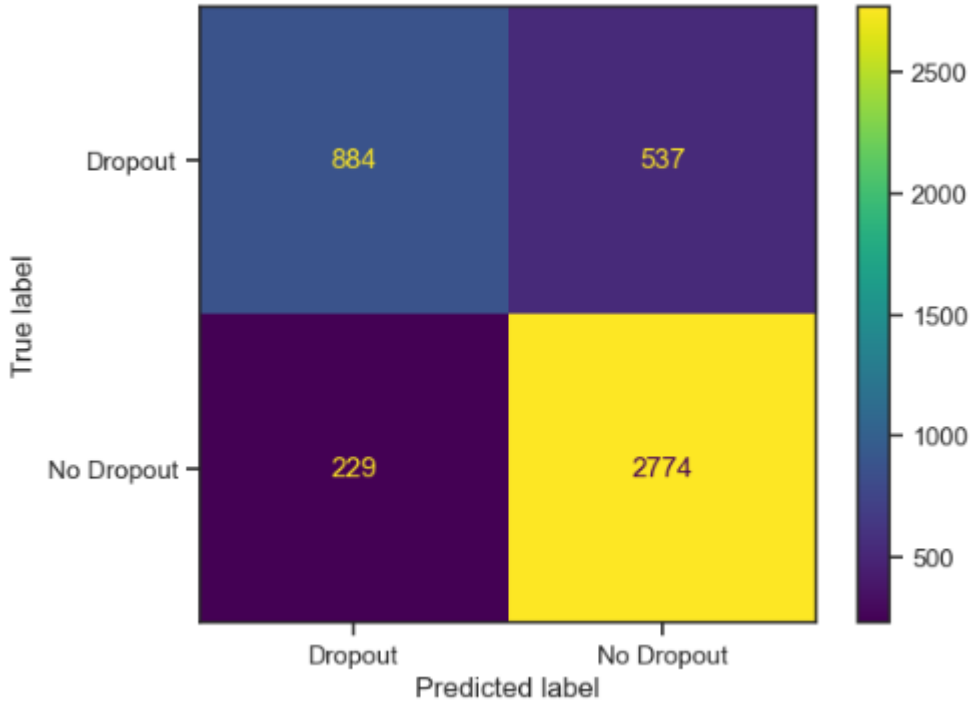


Figura A7: Resultats “Extras” - PIP

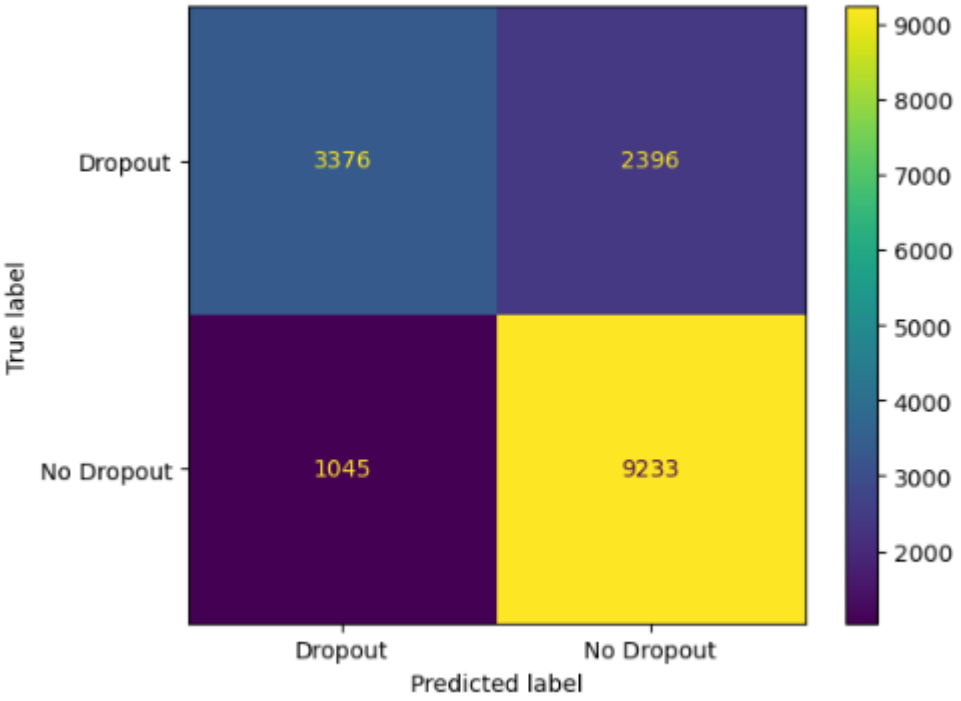


Figura A8: Resultats dataset complet - UB

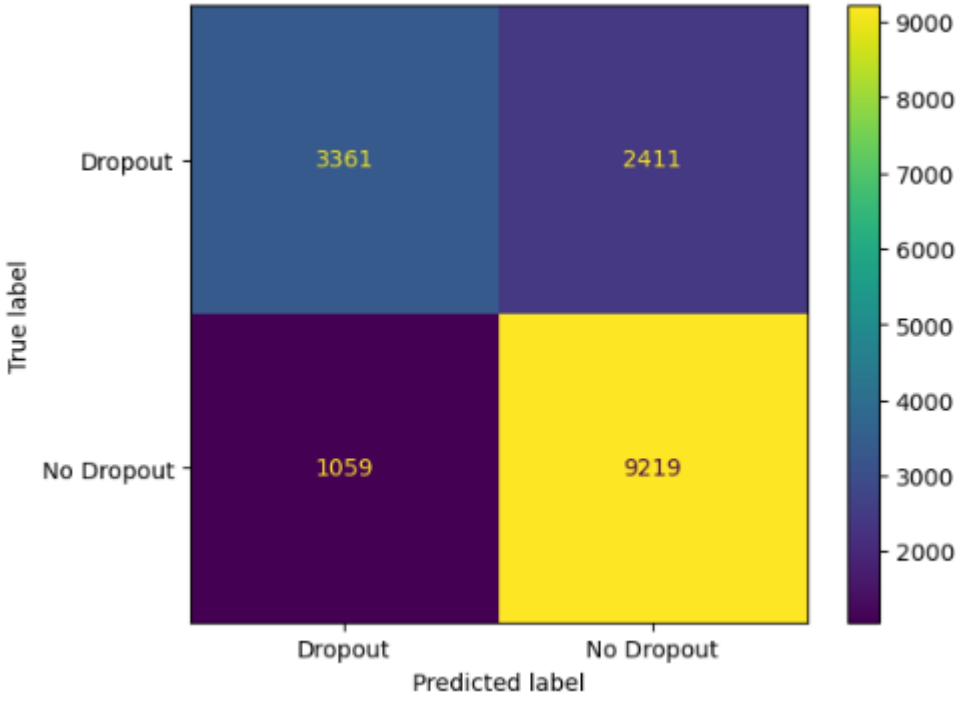


Figura A9: Resultats “notes” - UB

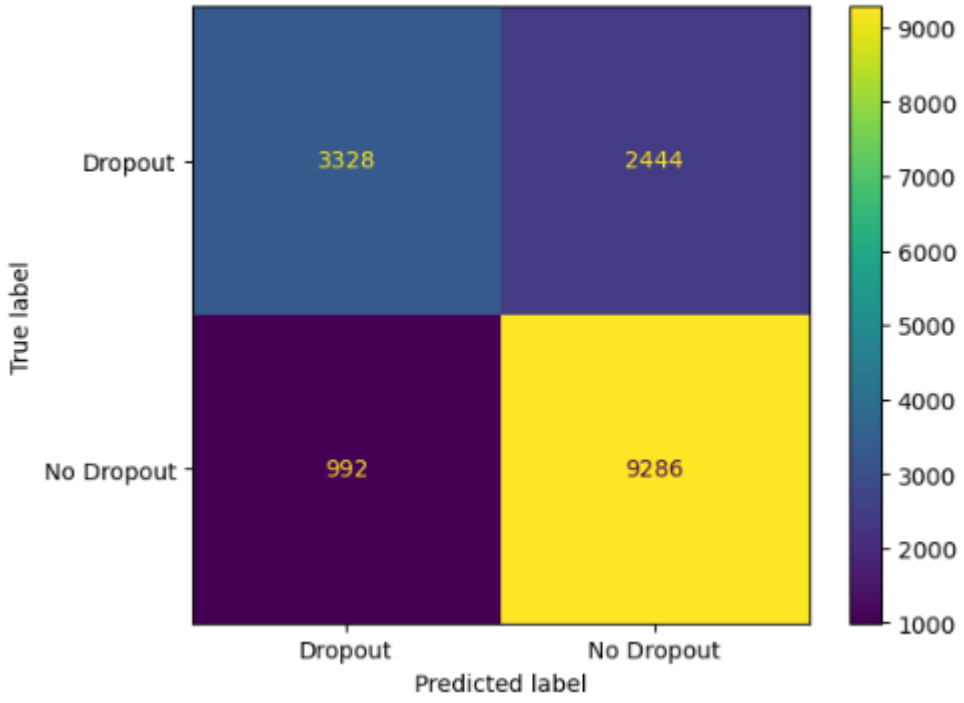


Figura A10: Resultats “avg” - UB

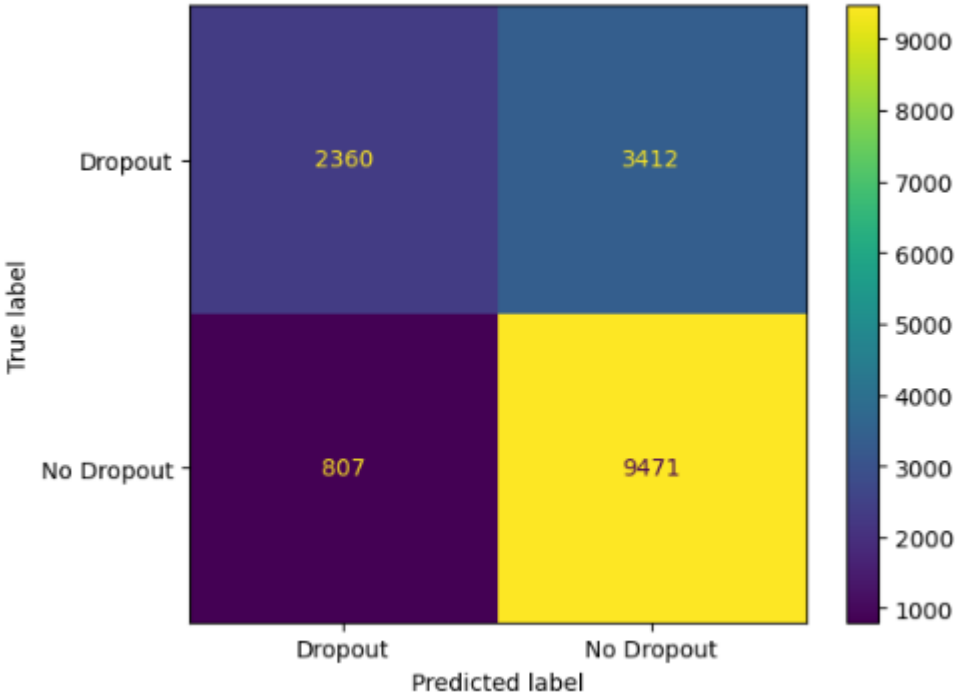


Figura A11: Resultats “susp” - UB

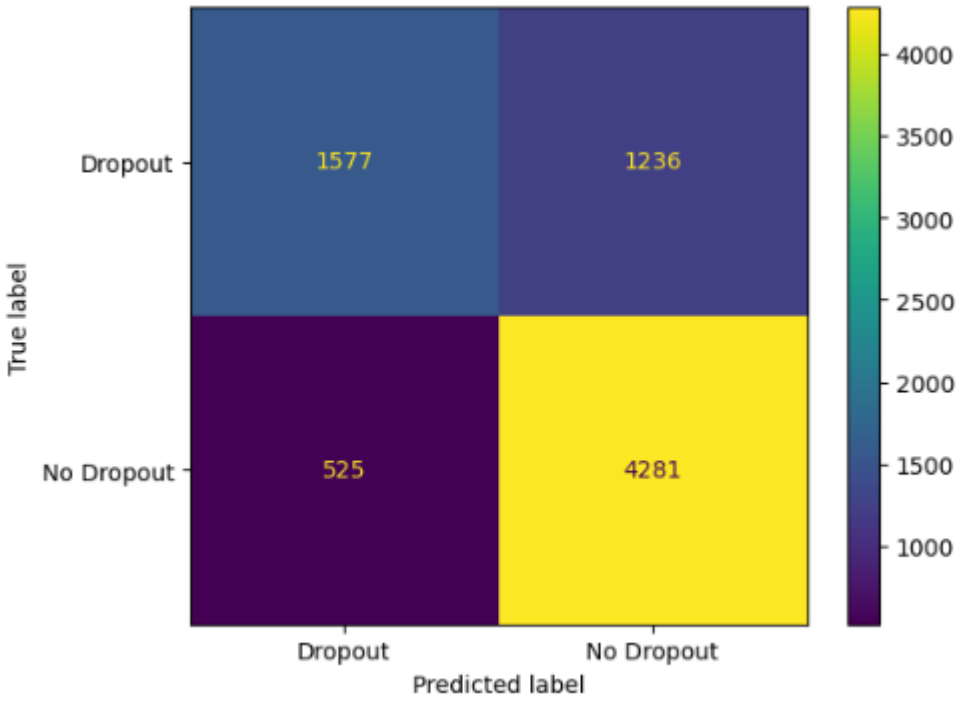


Figura A12: Resultats “sem1” - UB

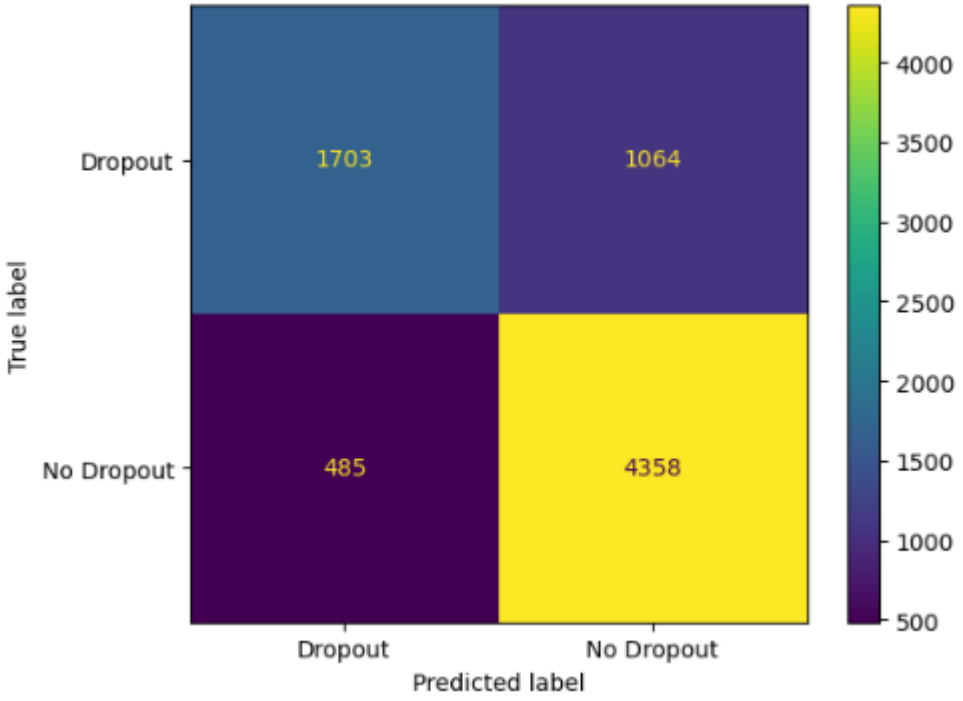


Figura A13: Resultats “sem2” - UB

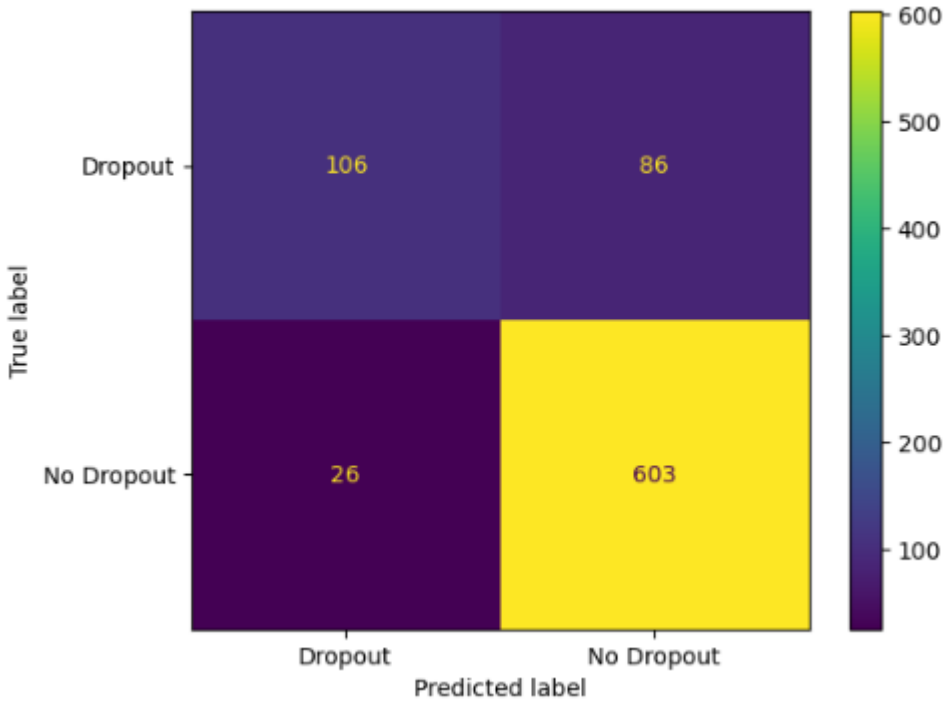


Figura A14: Resultats “anual” - UB