Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Aplicació de tècniques de potenciació del gradient a la predicció de resultats acadèmics

MEMÒRIA

Autor: Sara Provost Ferrer
Director: Luis José Talavera Méndez

Convocatòria: Gener 2022

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resum

Aquest document és la memòria escrita sobre l'estudi predictiu dels resultats acadèmics dels estudiants del Grau en Enginyeria en Tecnologies Industrials de l'ETSEIB. Concretament, s'han utilitzat tècniques de potenciació del gradient (Gradient Boosting) per predir si els estudiants del Grau, un cop superat el primer curs, aprovaran o suspendran les assignatures del tercer quadrimestre del Pla d'estudis.

Aquest estudi s'ha pogut realitzar gràcies al gran volum de dades acadèmiques disponibles sobre els estudiants de l'ETSEIB. Per tal d'aprofitar el màxim potencial de les dades, s'ha emprat la metodologia CRISP-DM, que especifica quines són les diferents fases d'un projecte de mineria de dades i les tasques corresponents a realitzar. Una gran part de les tasques han sigut de programació i s'han realitzat mitjançant la plataforma Anaconda, amb l'entorn de programació Spyder. S'ha emprat el llenguatge Python i, principalment, s'han utilitzat les llibreries Pandas i Sklearn.

Les principals conclusions que s'han tret al realitzar l'estudi són que el Gradient Boosting aconsegueix fer millors prediccions que altres models més simples i encara obtindria un rendiment predictiu superior si no fos per la baixa qualitat de les dades amb què s'ha treballat i el fet que estan desequilibrades. S'hauria de millorar la base de dades per poder fer prediccions més encertades.

Sumari

1. Introducció

L'any 1962 l'estadístic americà John W. Tukey [1], conegut pel desenvolupament del famós diagrama de caixa o *Box Plot,* va ser el primer que va parlar de la ciència de dades (*Data Science)* al reflexionar sobre l'evolució de l'estadística matemàtica i declarar que l'anàlisi de dades és una ciència empírica. Més endavant, el 1974, el científic danès Peter Naur va aconseguir una gran difusió del concepte 'ciència de dades' amb la publicació del llibre *Concise Survey of Computer Methods*, on parlava de la ciència de dades com un substitut de les ciències computacionals.

El 1996 va sorgir el dilema *Statistics = Data Science?* en una conferència al Japó i finalment el 2001 l'informàtic i estadístic americà William S. Cleveland va introduir la ciència de dades com una disciplina independent a l'estadística. Cleveland va definir sis àrees tècniques que, segons ell, conformaven la ciència de dades: els models i mètodes per a dades, les investigacions multidisciplinàries, la computació amb dades, la pedagogia, l'avaluació d'eines i la teoria.

Actualment, la ciència de dades és coneguda com la convergència de moltes disciplines d'anàlisi de dades i es basa en les matemàtiques, l'estadística i les tecnologies de la informació (IT). Té com a objectiu principal manipular i transformar les dades per extreure'n informació que aporti coneixements útils i rellevants. Es poden distingir tres tipus d'anàlisi de dades, en funció de les característiques, funcionalitats i objectius de l'estudi: l'anàlisi descriptiva, predictiva i prescriptiva [2].

L'anàlisi descriptiva, com bé el seu nom indica, s'utilitza per descriure un context o situació determinats de l'objecte d'anàlisi. Permet descriure patrons i realitzar un diagnòstic a partir de les dades disponibles, siguin del present (temps real) o del passat (temps històric). Serveix per obtenir informes d'estadística descriptiva, com per exemple percentatges, mitjanes, medianes, mínims, màxims o qualsevol indicador que permeti descriure l'objecte d'estudi. També proporciona visualitzacions avançades mitjançant eines de *reporting* que ofereixen informes personalitzats.

L'anàlisi predictiva de dades està basada en models estadístics i matemàtics. S'utilitzen tècniques més avançades, com el *Data Minin*g i el *Machine Learning.* La seva funció principal és predir esdeveniments futurs a partir d'un conjunt de dades històriques. Per exemple, pot ser útil per estimar vendes futures d'una empresa a curt termini o predir la situació meteorològica dels dies a venir.

L'anàlisi prescriptiva ens indica la millor decisió a prendre, o accions que s'han de dur a terme, en funció dels resultats obtinguts en l'anàlisi. S'utilitzen les tècniques més avançades utilitzades en el *Data Science* per poder fer recomanacions d'entre diferents opcions dins un escenari complex.

1.1. Machine Learning

L'aprenentatge automàtic o *Machine Learning* [4] engloba un conjunt de tècniques que, mitjançant algorismes complexos i mètodes matemàtics, permeten automatitzar la identificació de patrons o tendències a partir de l'anàlisi de les dades. És un subcamp de la intel·ligència artificial que, amb el suport de sistemes computacionals avançats, aconsegueix l'aprenentatge automàtic dels algorismes, sense la necessitat de programació explícita. L'aprenentatge pot ser supervisat o no supervisat [5].

En l'aprenentatge supervisat, els algorismes treballen amb dades etiquetades (*labeled data*) i pretenen trobar una funció que, donades unes variables d'entrada (*input data*), proporcioni l'etiqueta de sortida corresponent. Els algorismes s'entrenen amb dades històriques i així es fan robustos, aconseguint predir el valor de sortida adequat. El valor a predir pot ser una variable quantitativa, com en el cas de problemes de regressió, o qualitativa, com en el cas de problemes de classificació.

L'aprenentatge no supervisat es basa en algorismes d'agrupació o *clustering* i és de caràcter exploratori. En aquest cas no es disposa de dades etiquetades per a l'entrenament de l'algorisme. Només es coneixen les dades d'entrada però no existeixen valors de sortida associats als *inputs*. Per tant, només es pot intentar organitzar l'estructura de les dades per tal de simplificar l'anàlisi.

Actualment existeix una gran varietat d'algorismes de *Machine Learning,* alguns més complexos que altres, com per exemple la Regressió Lineal, la Regressió Logística, els arbres de decisió, les Xarxes Neuronals, el Support Vector Machine (SVM), el Gradient Boosting, etc.

La tècnica d'aprenentatge automàtic més avançada s'anomena *Deep Learning*. És una variant de la tècnica de Xarxes Neuronals, ja que pretén imitar el funcionament del cervell humà. L'algorisme, que pot ser de tipus supervisat o no supervisat, està format per un conjunt de capes neuronals artificials interconnectades entre si. La sortida d'una capa és l'entrada de la següent, d'aquesta manera es transmet la informació. El *Deep Learning* és molt útil si es volen tractar dades no estructurades com ara imatges, vídeos, àudios, etc. De fet, avui en dia molts sistemes

de reconeixement de veu i de visió artificial utilitzen aquesta tecnologia.

També destaquen els mètodes d'*ensemble*, que consisteixen a crear un model robust a partir de la combinació de múltiples models individualment febles (*weak learners*). S'anomenen models febles perquè obtenen prediccions lleugerament millors que si es generessin de manera aleatòria.

L'objectiu dels mètodes d'*ensemble* és aconseguir un bon equilibri entre el biaix i la variància, obtenint així molt bones prediccions. El biaix fa referència a com s'allunyen les prediccions d'un model respecte als valors reals. Reflecteix la capacitat del model per aprendre la relació real que hi ha entre les variables d'entrada i les variables resposta. Per altra banda, la variància indica com de diferent serà el model en funció de quines dades s'utilitzin en l'entrenament d'aquest.

Els dos mètodes d'*ensemble* [6] més utilitzats són el Bagging i el Boosting. El Bagging està format per models entrenats amb diferents subconjunts de dades i, per tant, cadascun fa una predicció diferent. S'agafa com a valor final la mitjana de totes les prediccions, si les variables són contínues, o la classe més freqüent, en el cas de treballar amb variables categòriques. En canvi, el Boosting [7] executa els models febles de forma seqüencial. Així cada model aprèn dels models anteriors i minimitza l'error de predicció. Els principals algorismes de Boosting són el AdaBoost i el Gradient Boosting. En aquest treball s'utilitzarà el Gradient Boosting amb arbres de decisió com a *weak learners,* que es descriuen en detall més endavant, als apartats 3.1 i 3.2.

1.2. Metodologia CRISP-DM

Per tal d'aprofitar el màxim potencial de les dades, s'utilitzen metodologies que especifiquen les diferents fases del procés i les tasques corresponents a realitzar. Les més conegudes són la metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) i la metodologia SEMMA (*Sample, Explore, Modify, Model and Assess*). La metodologia que se seguirà durant tot el projecte és la CRISP-DM i a continuació s'explica en què consisteix.

La metodologia CRISP-DM [3] proporciona una descripció normalitzada del cicle de vida d'un projecte estàndard d'anàlisi de dades, que està format per sis fases (vegeu *Figura 1.2.1*):

1. Business Understanding: En aquesta fase inicial s'han d'identificar els objectius del projecte, avaluar la situació i definir un problema de mineria de dades. També es dissenya un pla preliminar per assolir els objectius marcats.
2. Data Understanding: En primer lloc es recopilen les dades disponibles i s'exploren per tal de familiaritzar-s'hi i verificar-ne la qualitat. És possible detectar subconjunts de

dades interessants que poden donar lloc a hipòtesis sobre la informació que s'obtindrà.

3. Data Preparation: Partint de les dades inicials, es duen a terme les activitats necessàries (p.e. reorganitzar, transformar i netejar) per tal de construir el conjunt final de dades, preparades per ser utilitzades a la següent fase.
4. Modeling: S'apliquen les tècniques de modelatge adequades. Algunes tècniques tenen requeriments específics pel que fa a la forma de les dades. Per això, sovint s'ha de tornar a la fase de preparació de dades.
5. Evaluation: Un cop tenim diversos models amb qualitat suficient, s'han d'avaluar i comparar amb els objectius establerts a l'inici. També s'ha de decidir quina aplicació tindran els resultats obtinguts en el procés d'anàlisi de les dades.
6. Deployment: Gràcies als models creats, obtenim informació rellevant sobre les dades i pot ser que confirmem hipòtesis prèvies o que fem algun descobriment. Els resultats obtinguts s'han d'organitzar i presentar perquè el client pugui utilitzar-los.

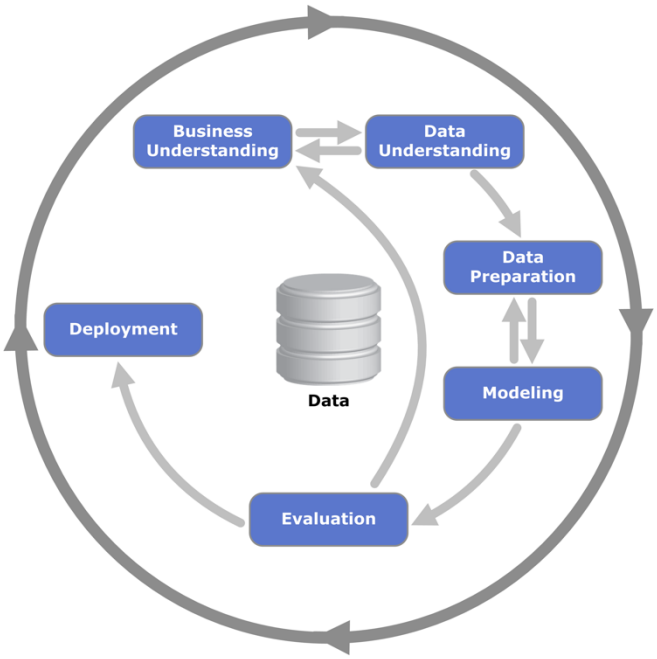


Figura 1.2.1. Procés CRISP-DM.

1.3. Objectius del projecte

L'objectiu principal d'aquest projecte és analitzar el rendiment de les prediccions obtingudes emprant el Gradient Boosting, un model predictiu molt popular en la ciència de dades. A partir d'una gran quantitat de dades acadèmiques d'estudiants de l'ETSEIB, es pretén predir si els estudiants del Grau en Enginyeria en Tecnologies Industrials, un cop superat el primer curs, aprovaran o suspendran les assignatures del tercer quadrimestre del Pla d'estudis.

Un altre objectiu és comprovar si la qualitat de les prediccions obtingudes aplicant l'algorisme de Gradient Boosting és superior a la que proporcionen altres models més simples, com ara la Regressió Logística. Per això, es farà una comparació dels resultats d'aquests dos models.

Aquest projecte també s'ha desenvolupat amb l'objectiu de proporcionar una ajuda als professors de l'ETSEIB. Els models predictius podrien ser una eina útil per determinar quins estudiants són propensos a suspendre i el professorat podria utilitzar aquesta informació a l'inici del curs per proposar suport acadèmic a aquests estudiants. Així, es podria aconseguir que hi hagués menys suspesos a les assignatures.

1.4. Abast del projecte

Com s'ha comentat prèviament, aquest projecte s'implementarà seguint la metodologia CRISP-DM. Es duran a terme totes les fases d'aquesta metodologia excepte l'última, la fase de Deployment, ja que queda fora dels límits establerts del projecte.

Durant la part experimental, d'anàlisi dels models predictius, es duran a terme tots els estudis necessaris per comprovar el rendiment del model i es faran els estudis addicionals pertinents, en la mesura que ho permetin els límits de duració del projecte.

2. Comprensió i preparació de dades

A continuació s'exploraran les dades disponibles per tal de familiaritzar-se amb els continguts i verificar-ne la qualitat. Posteriorment es descriuran les diferents operacions que s'han dut a terme per preparar les dades i els aspectes que s'han considerat rellevants de cara a la futura anàlisi.

2.1. Descripció de dades

Les dades disponibles per a l'estudi estan repartides en tres fitxers Excel. Per tal de comprendre aquestes dades s'han examinat els fitxers i detectat subconjunts amb més o menys potencial.

En primer lloc, es disposa d'un fitxer anomenat 'qfaseini19.xlsx' que conté dades corresponents a la fase inicial, és a dir, al primer curs del Grau, que consta de 10 assignatures. Apareixen les notes obtingudes a les diferents convocatòries d'un gran nombre estudiants d'entre el 2010 i el 2018.

També es disposa del fitxer 'qfasenoini19.xlsx' on es troben dades corresponents a les assignatures del segon, tercer i quart curs. Hi ha registrats els resultats que van obtenir estudiants d'entre el 2010 i el 2018.

Els dos fitxers mencionats tenen les dades distribuïdes en 11 columnes, que es descriuen a continuació (vegeu *Figura 2.1.1* i *Figura 2.1.2*):

- Codi Programa: Són els codis associats als Graus dels quals disposem dades. El codi que correspon al GETI és el '752'.
- Codi Expedient: Són codis de 6 dígits associats als estudiants, hi ha un codi diferent per a cada estudiant.
- Codi UPC: Són codis de 6 dígits associats a les assignatures del Grau. A la *Taula 2.1.1* es pot observar quins són els codis associats a les assignatures dels tres primers quadrimestres del Grau. Per altra banda, s'han detectat una sèrie de codis amb un format diferent del de les assignatures del pla d'estudis i, per tant, no identificats.
- Crèdits: Indica el nombre de crèdits de cada assignatura (veure *Taula 2.1.1*).
- Curs: Indica l'any en què els estudiants han cursat cada assignatura.
- Quad: Indica el quadrimestre en què s'ha cursat cada assignatura. El valor pot ser 0, 1 o 2. El valor 0 correspon a un quadrimestre previ a començar la carrera que serveix com a reforç, no s'ha fet sempre i no compta la nota. Cal destacar que el valor de

'Quad' no té per què correspondre al quadrimestre natural del pla d'estudis. Per exemple, si un estudiant ha cursat dues vegades una assignatura del primer quadrimestre, tindrà dues files associades a aquesta assignatura: una fila amb valor 'Quad' igual a 1 (primera convocatòria) i una altra fila amb valor 'Quad' igual a 2 (segona convocatòria).

- Supera: Indica si les convocatòries s'han aprovat o no, prenent el valor S o N respectivament.
- Nota Prof: És la nota que determina el professor en un primer lloc. Pot ser modificada posteriorment, per exemple si es va a revisió de l'examen final.
- Nota Num Aval: Aquesta nota coincideix gairebé sempre amb l'anterior i no se sap en què es diferencien.
- Nota Num Def: És la nota definitiva que apareix a l'acta.
- Grup Classe: Indica el grup al qual s'han matriculat els estudiants. Al primer curs hi ha 10 grups i al segon curs passa a haver-n'hi 5. La primera meitat dels grups acostumen a correspondre a horaris de matí i la segona meitat a horaris de tarda. En alguns casos, apareix el valor 'CONV' en lloc d'un grup. Això indica que l'estudiant ha convalidat l'assignatura en qüestió.

Taula 2.1.1. Crèdits i codis associats a les assignatures dels quadrimestres 1, 2 i 3.

Codi	Assignatura	Crèdits ECTS	Codi	Assignatura	Crèdits ECTS
240011	Àlgebra Lineal	6	**240024**	Química II	4,5
240012	Càlcul I	6	**240025**	Expressió Gràfica	7,5
240013	Mecànica Fonamental	6	**240031**	Electromagnetisme	6
240014	Química I	6	**240032**	Mètodes Numèrics	4,5
240015	Fonaments d'Informàtica	6	**240033**	Materials	4,5
240021	Geometria	6	**240131**	Equacions Diferencials	6
240022	Càlcul II	6	**240132**	Informàtica	4,5
240023	Termodinàmica Fonamental	6	**240133**	Mecànica	6

CODI_PROGRAMA	CODI_EXPEDIENT	CODI_UPC_UD	CREDITS	CURS	QUAD	SUPERA	NOTA_PROF	NOTA_NUM_AVAL	NOTA_NUM_DEF	GRUP_CLASSE
752	322167	240013	6	2015	2	N	2.7	2.7	2.7	30
752	327165	240023	6	2016	1	S	4.5	4.5	5	60
752	327165	240024	4.5	2016	1	S	8.3	8.3	8.3	20
752	327170	240025	7.5	2016	1	S	6	6	6	32

Figura 2.1.1. DataFrame corresponent al fitxer 'qfaseini19.xlsx'.

CODI_PROGRAMA	CODI_EXPEDIENT	CODI_UPC_UD	CREDITS	CURS	QUAD	SUPERA	NOTA_PROF	NOTA_NUM_AVAL	NOTA_NUM_DEF	GRUP_CLASSE
752	304101	240162	4.5	2018	1	N	4.6	4.6	4.6	20
752	265086	240162	4.5	2018	1	N	3.3	3.3	3.3	30
752	246396	240162	4.5	2018	1	N	4.1	4.1	4.1	40
752	276984	240401	3	2015	2	S	5.3	5.3	5.3	10

Figura 2.1.2. DataFrame corresponent al fitxer 'qfasenoini19.xlsx'.

Per últim, es disposa d'un fitxer anomenat 'dpersnomespreins19esc.xlsx' en què estan recopilades dades personals dels estudiants i relacionades amb el seu passat acadèmic, d'abans de començar la carrera. A continuació es detallen els diferents tipus de dades, que estan repartits en 8 columnes (vegeu *Figura 2.1.3*):

- Codi Expedient: Són codis de 6 dígits associats als estudiants, hi ha un codi diferent per a cada estudiant.
- Sexe: Indica si els estudiants són dones o homes, prenent el valor D o H respectivament.
- CP Familiar: És el codi postal associat a la residència dels estudiants a l'inici de la carrera.
- Any Accés: Indica l'any en què els estudiants van iniciar la carrera.
- Tipus Accés: Aquesta columna pren valor igual a 1 i indica que els estudiants han accedit al Grau havent fet les Proves d'Accés a la Universitat (PAU).
- Nota Accés: És la nota que va permetre als estudiants accedir al Grau. Aquesta nota depèn de la mitjana del Batxillerat i dels resultats obtinguts a les PAU.
- Centre Secundaria: Indica el centre on els estudiants van cursar el Batxillerat.
- CP Centre Sec: És el codi postal associat als centres de secundària.

CODI_EXPEDIENT	SEXE	CP_FAMILIAR	ANY_ACCES	TIPUS_ACCES	NOTA_ACCES	CENTRE_SECUNDARIA	CP_CENTRE_SEC
276951	H	25003	2013	1	11.022	INST.SAMUEL GILI I GAYA	25002
337243	H	08340	2017	1	9.85	INST.PERE RIBOT	8340
337246	D	08970	2017	1	9.71	SAGRAT COR-SARRIÀ	8034
337266	H	08021	2017	1	10.074	INFANT JESÚS	8006

Figura 2.1.3. DataFrame corresponent al fitxer 'dpersnomespreins19esc.xlsx'.

Havent valorat el tipus de dades del fitxer 'dpersnomespreins19esc.xlsx', s'ha decidit que de cara a l'anàlisi només s'utilitzaran les notes d'accés a la universitat (Nota Accés), que són les que poden aportar molta informació pel que fa al rendiment dels estudiants. Les dades de codis postals, tant 'CP Familiar' com 'CP Centre Sec', i les dades dels centres de secundària són massa variades, disperses i és difícil que puguin ser determinants.

2.2. Preparació de dades

Un cop familiaritzats amb les dades disponibles, s'han de reorganitzar per tal d'obtenir una estructura pràctica i còmoda a l'hora de dur a terme l'anàlisi. Aquesta estructura final de les dades és el resultat de fer operacions de neteja, filtratge i transformació, entre altres.

2.2.1. Operacions de neteja i filtratge

En primer lloc, s'han detectat grups de dades dels fitxers 'qfaseini19.xlsx' i 'qfasenoini19.xlsx' que, per motius concrets, s'han de descartar:

- S'han eliminat totes les files que contenen cel·les buides (NaN), ja que una de les biblioteques de Python amb què es treballarà posteriorment, anomenada Sklearn, no admet valors buits.
- S'han esborrat les files corresponents a assignatures cursades al 'quadrimestre 0' descrit anteriorment perquè no compta per nota, només serveix per reforçar els coneixements dels estudiants abans de començar el Grau.
- També s'han eliminat les files que tenen com a 'Grup Classe' el valor 'CONV'. Al ser assignatures convalidades, potser perquè s'han cursat en una altra facultat o grau, no aportaran informació.
- S'ha filtrat la columna 'Codi Programa' amb el valor '752' per tal de seleccionar només les dades corresponents al grau GETI.

A més, el fitxer 'qfasenoini19.xlsx' conté dades de tots els cursos posteriors a la fase inicial i per a l'estudi només interessen les dades relacionades amb el tercer quadrimestre (quadrimestre de tardor del segon curs). Per això, s'han eliminat totes les files no relacionades amb assignatures del tercer quadrimestre del pla d'estudis.

2.2.2. Operacions de transformació

Havent fet la neteja necessària de dades, s'ha creat una taula (DataFrame) fusionant les dades dels fitxers 'qfaseini19.xlsx' i 'qfasenoini19.xlsx'. Com s'ha pogut veure, als fitxers hi ha una fila

per a cada convocatòria a la qual s'han presentat els estudiants. Així doncs, s'han hagut de fer operacions de pivotatge per tal d'agrupar totes les dades de cada estudiant en una sola fila. D'aquesta manera, cada fila conté dades d'un estudiant diferent.

En primer lloc s'han generat les columnes Classe (Y), que corresponen a les dades de sortida que haurà de predir el model. En total són 6 columnes, una per cada assignatura del tercer quadrimestre del Grau, que contenen la nota obtinguda pels estudiants a la primera convocatòria (vegeu *Figura 2.2.1*). Com que es tracta d'un problema de classificació i, per tant, les variables de sortida han de ser categòriques, s'han codificat aquestes columnes assignant el valor 1 en cas de tractar-se d'una nota inferior a 5 i assignant un 0 en cas contrari (vegeu *Figura 2.2.2*).

)DI_EXPEDIE	ELECTROMAG	METNUM	MATS	EDOS	INFO	MEC
228576	6.8	9	4.4	5.6	7	0
229400	6.1	8.5	7.1	6.5	8.1	4.3
231418	2.7	3.1	5	3.9	3.5	2.2
230066	7.1	6.7	7	7.3	6.3	5

Figura 2.2.1. Columnes amb les notes de les assignatures del tercer quadrimestre.

)DI_EXPEDIE	ELECTROMAG	METNUM	MATS	EDOS	INFO	MEC
228576	0	0	1	0	0	1
229400	0	0	0	0	0	1
231418	1	1	0	1	1	1
230066	0	0	0	0	0	0

Figura 2.2.2. Columnes amb les notes codificades de les assignatures del tercer quadrimestre.

Seguidament, s'han creat les columnes predictores (X), les quals contenen el conjunt de dades d'entrada que ajudaran el model a predir correctament els valors de sortida. Primer s'ha generat una columna per cada assignatura del primer i segon quadrimestre, que sumen un total de 10 assignatures. Aquestes columnes contenen la nota que els estudiants van obtenir a l'aprovar les assignatures en qüestió, corresponent a l'última convocatòria que es van presentar (vegeu *Figura 2.2.3*). A més, s'han afegit 10 columnes, les quals contenen la mitjana dels estudiants en cada assignatura, calculada a partir de les notes que han obtingut en les diferents convocatòries cursades (vegeu *Figura 2.2.4*). Aquestes columnes són indicadores del rendiment de l'estudiant però no permeten distingir entre estudiants repetidors i no repetidors. Com que interessa fer aquesta distinció, s'ha decidit afegir una columna per cada assignatura que indiquin si els estudiants les han cursat un sol cop (prenent valor 0) o més d'un cop (prenent valor 1). Aquestes columnes es poden observar a la *Figura 2.2.5.*

)DI_EXPEDIE	ALG	CALC1	MECFON	QUIM1	FINFO	GEO	CALC2	TERMOFON	QUIM2	EXPRE
228576	5.4	5	5.7	6.4	8.4	7.3	5.8	5	7.8	7.7
229400	8.1	7.6	7.1	6.8	8.4	6	7	6.2	8.6	8.2
231418	5	7	5	7.3	6.4	6	5.6	5	7.1	5.3
230066	6.4	7.8	6.9	7.8	6.8	6.3	7.2	6.4	7.8	6.2

Figura 2.2.3. Columnes amb les notes aprovades de les assignatures del primer curs.

)DI_EXPEDIE	mALG	mCALC1	mMECFON	mQUIM1	mFINFO	mGEO	mCALC2	mTERMOFON	mQUIM2	mEXPRE
228576	5.4	4.85	5.7	5.45	8.4	5.65	5.8	3.2	7.8	7.7
229400	8.1	7.6	7.1	6.8	8.4	6	7	6.2	8.6	8.2
231418	5	5.65	4.4	5.8	3	3	5.6	2.75	5.05	5.3
230066	6.4	7.8	6.9	7.8	6.8	6.3	7.2	6.4	7.8	6.2

Figura 2.2.4. Columnes amb les mitjanes de les assignatures del primer curs.

)DI_EXPEDIE	rALG	rCALC1	rMECFON	rQUIM1	rFINFO	rGEO	rCALC2	rTERMOFON	rQUIM2	rEXPRE
228576	0	1	0	1	0	1	0	1	0	0
229400	0	0	0	0	0	0	0	0	0	0
231418	0	1	1	1	1	1	0	1	1	0
230066	0	0	0	0	0	0	0	0	0	0

Figura 2.2.5. Columnes indicant si els estudiants han repetit les assignatures del primer curs.

Per altra banda, s'ha considerat si el grup en el qual es matriculen els estudiants podria influir en les notes que obtenen. Més concretament, es planteja la idea que els estudiants que van a grups de matí tenen més bona nota que els que segueixen horaris de tarda. A l'hora de fer la matrícula, com millor és l'expedient acadèmic dels estudiants, més prioritat tenen per triar els grups classe. Els primers grups que s'omplen són els d'horari de matí i, per tant, els estudiants amb pitjor expedient acadèmic acaben triant grups d'horari de tardes, que són els únics que queden lliures.

Per comprovar que els grups classe estan directament relacionats amb les notes del primer curs, s'ha generat un gràfic per cada assignatura del primer curs. Els grups classe es troben a l'eix d'abscisses i les notes de la primera convocatòria dels estudiants a l'eix d'ordenades. A la *Figura 2.2.6* es mostra un dels gràfics generats, corresponent a l'assignatura Química I. Com es pot observar, la distribució de les dades és bastant homogènia. Es distingeix entre grups de matí, compresos entre els grups 10 i 70, i grups de tarda, que corresponen als grups 80, 90 i 100. Tot i que es detecten més suspesos als grups de tardes, no hi ha una diferència prou significativa de notes entre grups. A tots els grups hi ha una gran varietat de notes. També cal tenir en compte

que els estudiants molt sovint assisteixen a classe de grups diferents d'aquells als quals s'han matriculat i, per tant, no seria representativa la informació aportada per aquestes dades. Amb la resta de gràfics es treuen les mateixes conclusions i, per tant, es descarta afegir les dades referents als grups classe d'assignatures del primer curs, ja que no aportaran informació en l'estudi.

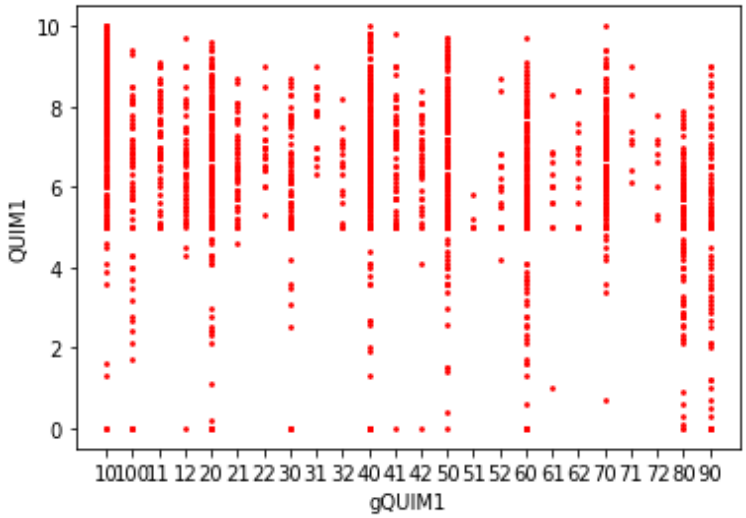


Figura 2.2.6. Gràfic amb els Grups Classe vs. Notes de Química I.

Per altra banda, es considera que les notes d'accés a la universitat ('Nota Accés') del fitxer 'dpersnomespreins19esc.xlsx' són de valor i per això s'ha afegir a la taula una columna amb aquestes dades.

3. Modelatge i Avaluació

En aquest apartat es descriurà l'algorisme de Gradient Boosting, que serà l'utilitzat per crear els models predictius dels diferents estudis. També és important conèixer el funcionament dels arbres de decisió, ja que són la base del Gradient Boosting.

Per altra banda, es definiran les diferents mètriques d'avaluació que es poden utilitzar per calcular el rendiment predictiu dels models. També s'escollirà el mètode d'avaluació que s'implementarà per aconseguir més robustesa en els models.

3.1. Decision Tree

Un arbre de decisió o Decision Tree és un model de predicció simple anomenat també classificador simple. És fàcil d'interpretar fins i tot quan la relació entre els predictors és complexa. A més, es pot aplicar en problemes de regressió i també de classificació: pot treballar tant amb predictors quantitatius com qualitatius [8].

Aquest model, a base de dividir les dades en subgrups, genera un diagrama amb forma d'arbre, tal com el seu nom indica (vegeu *Figura 3.1.1).* Està format per nodes units entre ells amb arestes (branques). Es distingeix entre tres tipus de nodes:

- Primer node o node arrel: Es troba a la part superior del diagrama i correspon a la primera divisió que es fa de les dades segons la variable d'entrada més significativa. Es fan tantes divisions com valors pot prendre la variable i correspon al nombre de branques que surten del node.
- Nodes interns o intermedis: Es tornen a dividir les dades en subgrups en funció de les variables, de manera que va augmentant el nombre de ramificacions del diagrama.
- Nodes terminals: Es troben a l'extrem inferior del diagrama i s'associen a les fulles dels arbres. Indiquen la classificació definitiva.

Concretament, a partir de regles binàries recursives el model analitza a cada node la millor variable per ramificar. Es van agrupant les dades en conjunts homogenis segons les variables d'entrada, de manera que a cada divisió augmenta el nombre de ramificacions del diagrama. La profunditat de l'arbre es pot fixar determinant el nombre màxim de nodes que es desitja que tinguin les branques.

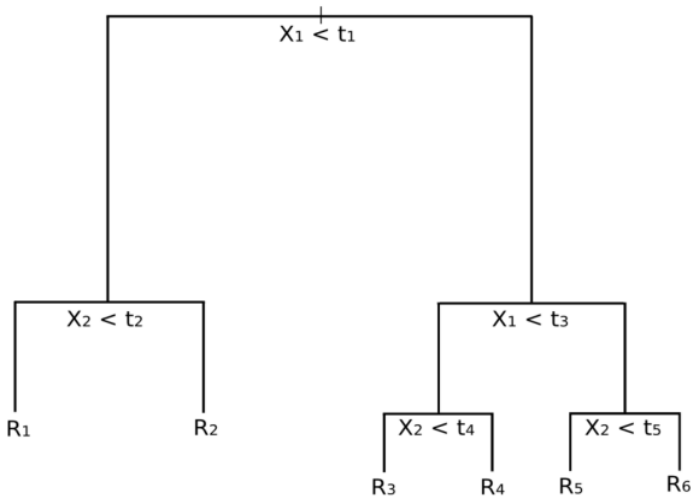


Figura 3.1.1. Arbre de decisió amb sis regions.

Els arbres de decisió es poden representar mitjançant un gràfic 2D (vegeu *Figura 3.1.2)* quan es disposa de només dues variables predictores (X_1 i X_2). Aquest tipus de gràfic permet visualitzar les diferents regions de decisió que genera un arbre en funció del valor d'entrada. Aquestes regions mai se solapen.

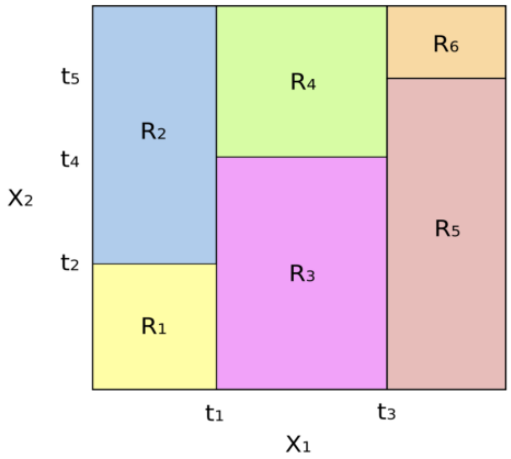


Figura 3.1.2. Representació gràfica 2D.

3.2. Gradient Boosting

El Boosting és un mètode d'*ensemble* que consisteix a crear un model predictiu a partir d'un conjunt de models predictius febles o *weak lerners*, els quals acostumen a ser arbres de decisió. El primer algorisme que va permetre aplicar el mètode de Boosting va ser l'AdaBoost (Adaptative Boosting), que es va inventar el 1995. Més endavant va aparèixer el Gradient Boosting *[9],* una generalització de l'algorisme AdaBoost que, al poder-se aplicar tant a problemes de regressió com de classificació, ha tingut molt èxit.

El model s'entrena de forma seqüencial, de manera que cada arbre minimitza l'error de predicció de l'arbre anterior. El fet de voler reduir l'error a cada iteració fa que el model sigui susceptible de patir *overfitting [10]*. Aquest fenomen es produeix quan el model s'ajusta tant a les dades d'entrenament que és incapaç de predir correctament noves observacions, ja que ha acabat memoritzant les dades d'entrenament sense aprendre la relació real entre variables. Això es pot detectar mirant si el rendiment del model és molt més elevat amb les dades d'entrenament que amb les dades d'avaluació.

Abans d'aplicar l'algorisme s'han de determinar una sèrie de paràmetres, també anomenats hiperparàmetres, que podran ajudar a evitar l'*overfitting* i augmentar el rendiment de les prediccions. A continuació es descriuen els més destacats:

- Nombre d'arbres (*n_estimators*): Serveix per determinar el nombre de *weak learners* que es desitja que tingui el model. Com més gran sigui aquest nombre, major serà el risc de patir *overfitting* i per això es tracta d'un paràmetre crític.
- Profunditat (*max_depth*): És el nombre màxim de divisions que es pot fer a cada arbre i, per tant, la profunditat màxima que podran tenir. Permet controlar el nivell de complexitat del model. Com més baix sigui aquest nombre, menor serà la complexitat del model. Normalment s'utilitzen valors petits, com ara 1, 2 o 3. Si per exemple s'escull un valor igual a 1, cada arbre farà una única divisió i, per tant, no tindran nodes intermedis.
- Paràmetre de penalització (*learning_rate*): Controla la velocitat d'aprenentatge del model. Com més gran sigui el valor del paràmetre escollit, més ràpid aprendrà el model, però augmentarà el risc d'*overfitting*. Per això, s'acostumen a utilitzar valors petits, com ara 0,01 o 0,001, encara que comporti utilitzar un major nombre d'arbres per tal d'aconseguir bons resultats.

Per tal de trobar els valors dels hiperparàmetres que optimitzen el model, es defineix una llista de valors per a cada un d'ells. A continuació, es fan totes les combinacions possibles dels hiperparàmetres agafant cada valor de les llistes definides i es comparen els rendiments del model obtinguts per a cada cas. Així, s'acaba escollint el millor model i els valors òptims de cada hiperparàmetre. Aquest procés es du a terme mitjançant la Validació Creuada, que requereix disposar d'un equip de computació potent i tarda molt a executar-se.

En cas de voler resoldre problemes de gran escala, en lloc d'aplicar l'algorisme Gradient Boosting es recomana utilitzar el XGBoost (Extreme Gradient Boosting) [11], ja que és més potent i eficient. L'algorisme es pot executar en paral·lel, fet que permet utilitzar més d'un

computador per entrenar el model si la complexitat del problema ho requereix. A més, té una velocitat d'execució 10 vegades superior a la del Gradient Boosting.

Tant si el problema que es vol estudiar és de classificació com si és de regressió, l'algorisme XGBoost el tracta internament com un cas de regressió. Només admet dades numèriques i per això és tan eficient. En cas de disposar de dades categòriques, s'han de transformar abans d'entrenar el model, per exemple, mitjançant un procés de codificació instantània. Per altra banda, no suposa un problema que faltin valors (NaN) del conjunt de dades, ja que XGBoost és capaç de manipular aquests valors buits. En canvi, XGBoost no funciona gaire bé si les dades són no estructurades, com ara en el reconeixement d'imatges i la visió per computador. Com s'ha comentat prèviament, en aquests casos és millor aplicar el *Deep Learning*.

3.3. Mètriques d'avaluació

És necessari escollir una mètrica d'avaluació per poder comprovar numèricament la fiabilitat del model un cop entrenat, és a dir, la qualitat de les prediccions que fa. A continuació es descriuen les diferents mètriques d'avaluació que se solen utilitzar en models de classificació [12].

La Matriu de Confusió és una mètrica molt utilitzada que permet inspeccionar i avaluar visualment les prediccions del model. La matriu té el mateix nombre de files i columnes, que correspon al nombre de classes (valors que pot prendre la variable resposta). A les files i es troben els valors reals i a les columnes j hi ha els valors que s'han predit. Així doncs, cada element de la matriu C_{ij} és el nombre d'observacions verdaderes del grup i classificades al grup j. A la *Figura 3.3.1* es veu un exemple amb dues classes i, per tant, amb matriu de confusió d'ordre 2. Els quatre elements de la matriu tenen la següent definició:

- Verdader positiu (VP): S'ha predit que el valor és Positiu i correspon amb el valor real. Per tant, la predicció és correcta.
- Fals positiu (FP): S'ha predit que el valor és Positiu i el valor real és Negatiu. Per tant, la predicció és errònia (conegut com a error Tipus 1).
- Fals negatiu (FN): S'ha predit que el valor és Negatiu i el valor real és Positiu. Per tant, la predicció és errònia (conegut com a error Tipus 2).
- Verdader negatiu (VN): S'ha predit que el valor és Negatiu i correspon amb el valor real. Per tant, la predicció és correcta.

En aquest treball els positius corresponen a la classe 'suspesos' i els negatius a la classe 'aprovats'.

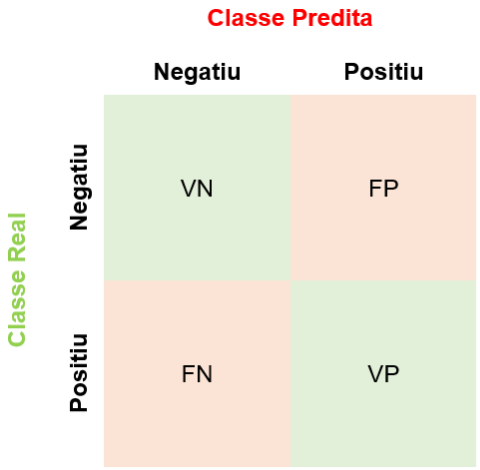


Figura 3.3.1. Matriu de confusió 2x2.

L'Exactitud (E) o Accuracy indica l'habilitat del model per predir de forma correcta totes les classes. Es calcula dividint el nombre de prediccions correctes entre el nombre total de prediccions realitzades i es representa en forma de percentatge o amb un valor entre el 0 i 1. És molt útil quan les dades són equilibrades, és a dir, quan es disposa més o menys del mateix nombre de dades de cada classe.

$$E = \frac{VP + VN}{Total} \qquad \textit{(Eq. 3.1)}$$

Per altra banda, la Precisió (P) o Precision se centra en els valors positius que prediu el model i indica l'habilitat que té per trobar verdaders positius (VP) i no assignar falsos positius (FP). Es calcula amb la següent equació:

$$P = \frac{VP}{VP + FP} \qquad \textit{(Eq. 3.2)}$$

En canvi, la Sensibilitat (S) o Recall indica l'habilitat que té el model per encertar positius (VP) respecte al nombre total de positius reals. Es calcula amb la següent equació:

$$S = \frac{VP}{VP + FN} \qquad \textit{(Eq. 3.3)}$$

Existeix una mètrica anomenada Puntuació F1 o F1 Score que s'utilitza sobretot quan les dades disponibles no estan equilibrades. És una combinació de la Precisió i la Sensibilitat i es calcula fent la mitjana harmònica de les dues mètriques. Per al cas en què es tenen només dues classes, l'equació és la següent:

$$F1_{score} = 2 \cdot \frac{P \cdot S}{P + S} \qquad \textit{(Eq. 3.4)}$$

Per tant, s'obtindrà una bona Puntuació F1 quan el nombre de positius predits (VP) sigui gran respecte el nombre de positius reals i el model no s'equivoqui al predir positius i, per tant, el nombre de falsos positius (FP) sigui petit.

Tenint en compte que les dades que s'utilitzaran per a l'estudi no estan equilibrades, es descarta utilitzar la mètrica Exactitud. Per comprovar el rendiment dels models s'utilitzarà l'F1 Score. Com més gran sigui el valor de F1 Score, més elevat serà el nombre de suspesos predits correctament (VP) i, per tant, el rendiment predictiu serà superior. Per tenir més informació sobre el rendiment dels models, és calcularà també el Recall. Un Recall elevat indicarà que el nombre de suspesos predits correctament (VP) és molt superior al nombre d'aprovats predits incorrectament (FN). També es contemplarà la Precisió, ja que interessa que el nombre de suspesos predits incorrectament (FP) sigui mínim. No seria òptim dedicar esforços per reforçar l'aprenentatge d'estudiants que es creu que suspendran però que, en realitat, no necessiten ajuda per aprovar.

3.4. Mètodes d'avaluació

Els mètodes d'avaluació s'utilitzen per entrenar el model i posteriorment avaluar-lo. L'avaluació del model permet esbrinar quin seria el seu rendiment davant de dades noves, que no ha vist. Els mètodes d'avaluació també són útils per detectar si el model pateix *overfitting*. Si a l'avaluació s'obté que el model té un rendiment molt més baix respecte a l'obtingut durant l'entrenament, vol dir que el model s'ha ajustat a les dades d'entrenament sense trobar la relació real entre les variables i, per tant, pateix *overfitting*.

A continuació s'escollirà el mètode d'avaluació que s'utilitzarà en el conjunt d'experiments que es duran a terme en aquest treball. Es farà una comparació entre el Hold-out i la Cross Validation [13], que són els mètodes més populars.

En el cas del Hold-out se separen les dades en dos grups, un per l'etapa d'entrenament i l'altre per l'etapa d'avaluació. El més freqüent és utilitzar un 70% o 80% de les dades per entrenar el model i utilitzar la resta, un 30% o 20%, per avaluar-lo i comprovar-ne el rendiment (vegeu *Figura 3.4.1*). Aquest mètode funciona molt bé quan es tenen grans quantitats de dades.

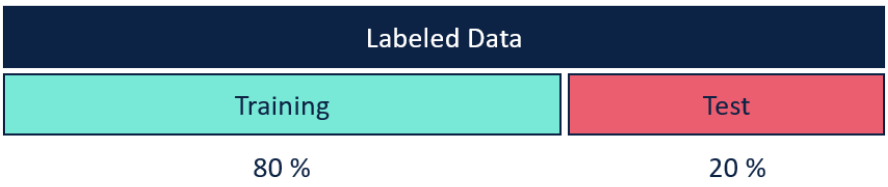


Figura 3.4.1. Mètode Hold-out.

En canvi, quan s'aplica el mètode de Validació Creuada o Cross Validation se separen aleatòriament les dades en k grups. Un dels grups s'utilitza per fer l'avaluació i la resta per a l'entrenament. Aquest procés es fa k vegades, de tal manera que tots els grups s'hauran utilitzat tant en l'etapa d'entrenament com en la d'avaluació. El procés descrit es pot observar a la *Figura 3.4.2* per a un cas concret amb $k = 5$. Aquest mètode s'utilitza quan es disposa de poques dades, ja que un sol grup d'entrenament no seria suficient per obtenir un model robust que contempli tots els casos.

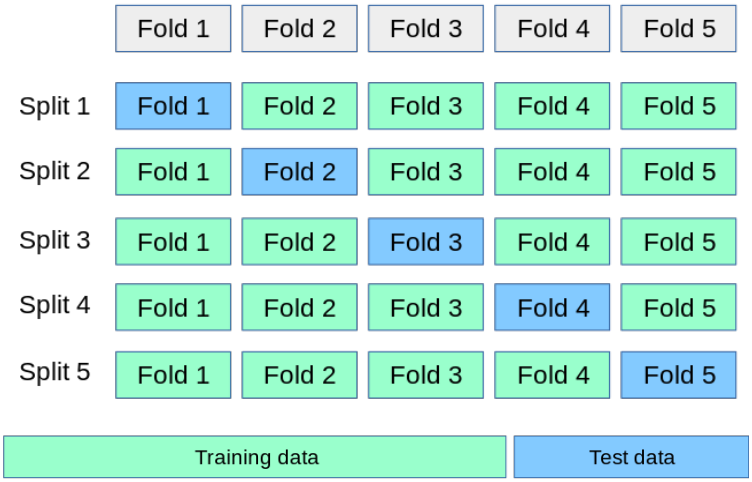


Figura 3.4.2. Mètode Cross Validation.

Havent vist les diferències entre els dos mètodes d'avaluació, s'ha decidit utilitzar el Hold-out. S'ha considerat que el volum de dades disponibles és prou gran per fer només un grup d'entrenament i un altre d'avaluació. També s'ha de tenir en compte que el mètode de Cross Validation, al ser una versió més sofisticada del Hold-out, requereix més temps d'execució i un equip més potent.

4. Avaluació dels resultats

A continuació es durà a terme la construcció dels models de Gradient Boosting que permetran predir si els estudiants, un cop superat el primer any de la carrera, suspendran o aprovaran les assignatures del tercer quadrimestre. En total es construiran sis models predictius diferents, ja que cada assignatura s'ha de tractar com un problema de classificació diferent. Les dades que es disposen de cada assignatura poden estar més o menys equilibrades. A més, les assignatures tenen mètodes d'avaluació diferents i poden ser més o menys complexes.

Per tal de trobar el millor model predictiu, s'han de trobar els valors òptims dels hiperparàmetres descrits prèviament. La Validació Creuada és el mètode que permet trobar aquests valors òptims. Però, com s'ha comentat anteriorment, tarda molt a executar-se i és necessari disposar d'un equip de computació potent. Com que per a aquest treball no es disposa d'un equip prou potent, s'ha aplicat un mètode alternatiu més simple que, conseqüentment, no garanteix trobar el model òptim. Tot i així, amb les limitacions esmentades, s'esbrinarà quins poden ser els millors models a aplicar.

El mètode alternatiu que s'utilitzarà consisteix a creuar diferents valors dels hiperparàmetres 'nombre d'arbres' i 'profunditat' i comparar els models resultants. Els valors de profunditat considerats són 1, 2 i 3, ja que se solen utilitzar valors petits, com s'explica a l'apartat teòric (el valor per defecte en el Gradient Boosting és 3). Per veure com influeix el 'nombre d'arbres' en el model, s'ha utilitzat la següent llista de valors: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300. S'han estudiat més en detall els valors inferiors a 100 (valor per defecte en el Gradient Boosting) perquè interessa que el model sigui al més simple possible. També s'han considerat els valors 200 i 300, ja que no es descarta que s'obtinguin millors resultats quan el nombre d'arbres és superior a 100. Si fos el cas, s'ampliaria el rang de valors de l'interval entre 100 i 300 per estudiar-los en detall. Per altra banda, s'ha decidit fixar l'hiperparàmetre de penalització o *learning_rate*, agafant el valor per defecte de la llibreria Sklearn, que és 0,1. Així es redueix el nombre de combinacions a valorar, i per tant, el nombre de models diferents a comparar entre ells.

Per a cada assignatura, s'ha decidit utilitzar dos grups de dades diferents per crear els models. El primer grup és el que s'anomenarà Dataset A i conté les columnes predictores corresponents a les notes que van obtenir els estudiants al aprovar les assignatures del primer curs. El segon grup de dades és el Dataset B i conté totes les columnes predictores que s'han descrit a l'apartat 2.2. L'objectiu de treballar amb aquests dos grups de dades diferents és veure com varia l'efectivitat dels models en funció de la quantitat de dades predictores utilitzades. Cal destacar que el Dataset A només conté 10 columnes predictores i, en canvi, el Dataset B en conté 31.

A continuació, es farà un estudi detallat dels millors models per a cada assignatura. En tots els casos s'utilitzarà el 75% de les dades per a l'entrenament del model i el 25% restant per a l'avaluació del model. En primer lloc s'aplicarà la Regressió Logística [14], que és un model molt simple comparat amb el Gradient Boosting. Els resultats obtinguts amb la Regressió Logística serviran de referència per, quan a continuació s'apliqui el Gradient Boosting Classifier [15], poder comprovar que al ser aquest de més complexitat aconsegueix fer millors prediccions. Per comparar la qualitat dels models s'observaran els valors de F1 Score, Recall i Precision obtinguts en l'avaluació. A més, serà molt útil visualitzar la matriu de confusió per acabar de prendre la millor decisió, buscant minimitzar els errors Tipus 1 i Tipus 2.

4.1. Electromagnetisme

Abans d'iniciar l'estudi, es comprovarà quina és la proporció d'aprovats i suspesos de l'assignatura d'Electromagnetisme, per tal de veure si es disposa de dades equilibrades o no. Parlaríem de dades equilibrades si disposéssim d'un 50% d'aprovats i un 50% de suspesos. Però s'ha trobat que un 65% de les dades corresponen a estudiants que van aprovar Electromagnetisme a la primera convocatòria i un 35% correspon a estudiants que la van suspendre. Així doncs, com que es disposa de menys dades sobre els estudiants que suspenen, serà més difícil predir els suspesos que els aprovats. Això també passarà en la resta d'assignatures en més o menys mesura.

Primer de tot s'estudiaran els models obtinguts a partir del Dataset A. El model de Regressió Logística, que servirà de referència, té un valor de F1 Score igual a 0.50, una Precision de 0.56 i un Recall de 0.45. A la *Figura 4.1.1* es pot observar la matriu de confusió d'aquest model. A més, per comprovar si hi ha *overfitting,* s'han comparat els valors de F1 Score del grup d'entrenament amb el d'avaluació i s'ha trobat que són iguals. Per tant, aquest model no presenta *overffiting*. Totes aquestes característiques es compararan posteriorment amb les obtingudes aplicant el Gradient Boosting.

A continuació s'aplicarà el Gradient Boosting Classifier i s'avaluaran les diferents combinacions que s'han fet dels hiperparàmetres 'nombre d'arbres' i 'profunditat'. A la *Figura 4.1.2* es mostra com varia l'F1 Score en funció dels hiperparàmetres.

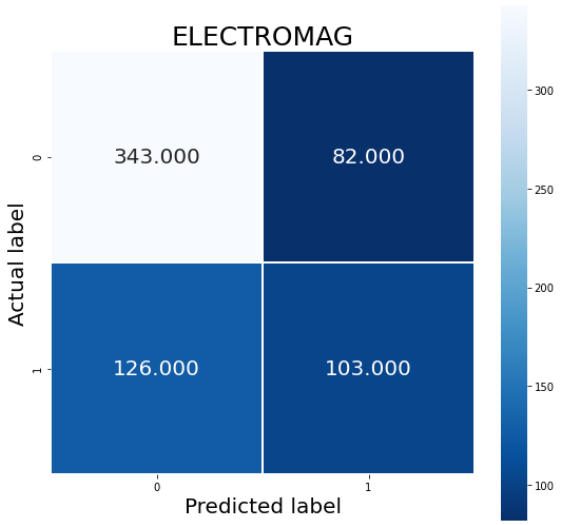


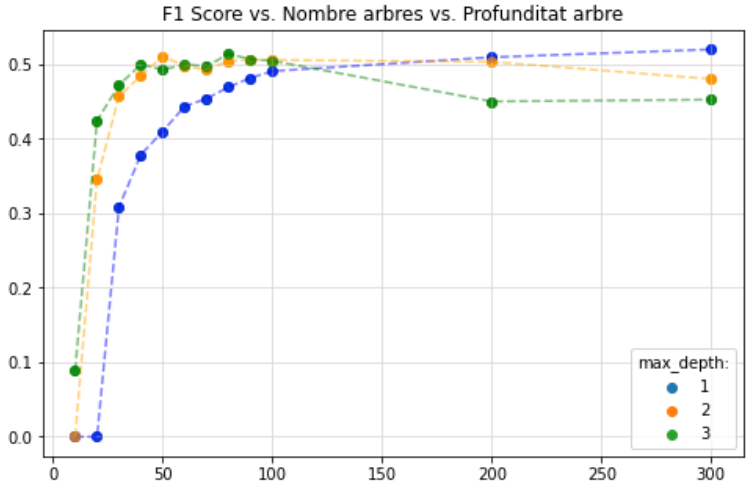
Figura 4.1.1. Matriu de confusió del model de Regressió Logística (Dataset A).

Havent observat la *Figura 4.1.2,* concretament la *(b)* i la *(c),* s'han recollit a la *Taula 4.1.1* les combinacions d'hiperparàmetres que han obtingut els millors valors de F1 Score. Per comprovar l'existència d'*overfitting*, s'han comparat els valors de F1 Score del grup d'entrenament (*train*) amb el d'avaluació (test) i s'ha pogut observar que la diferència entre aquests dos valors augmenta a mesura que augmenta la profunditat. Per tant, com més profunditat tinguin els arbres més *overfitting* patirà el model.

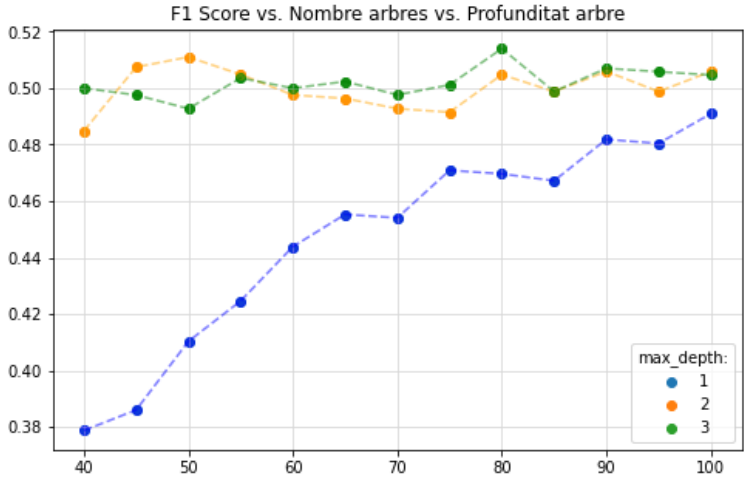
Taula 4.1.1. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	300	0.52	0.54	0.48	0.57
	275	0.52	0.54	0.48	0.56
	250	0.51	0.54	0.47	0.56
2	125	0.52	0.61	0.49	0.55
	175	0.52	0.64	0.49	0.55
	150	0.51	0.63	0.48	0.55
3	80	0.51	0.68	0.48	0.55
	90	0.51	0.69	0.47	0.55
	95	0.51	0.69	0.47	0.55

(a)



(b)



(c)

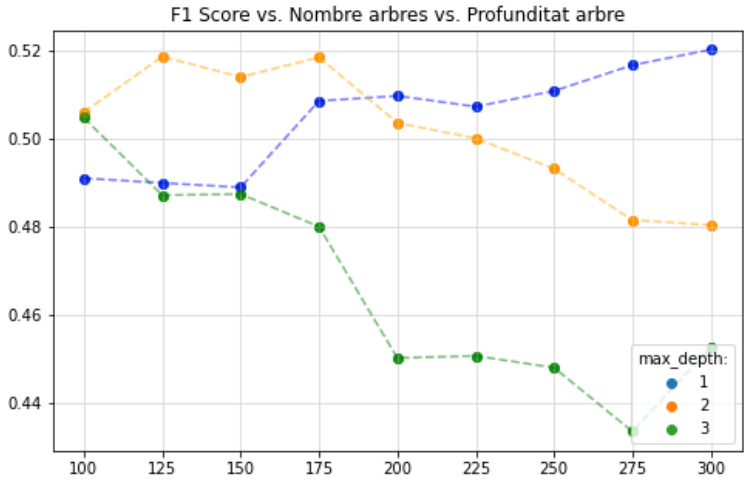


Figura 4.1.2. F1 Score en funció del nombre d'arbres i la profunditat (Dataset A). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 40 i 100. (c) Interval del nombre d'arbres ampliat entre 100 i 300.

A la *Taula 4.1.1* es pot observar que tots els models seleccionats són pràcticament idèntics pel que fa al seu rendiment, ja que les mètriques d'avaluació utilitzades ens mostren que tots tenen la mateixa qualitat de predicció. Pel que fa als hiperparàmetres, s'observa que quan augmenta la profunditat dels arbres de decisió, el model necessita menys arbres per obtenir els millors resultats.

Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió. Amb profunditat 1, es recomana utilitzar entre 275 i 300 arbres, ja que s'obté el menor nombre de falsos positius (FP) i un nombre elevat de verdaders positius (VP). En el cas d'escollir profunditat 2, es recomana utilitzar entre 125 i 175 arbres, ja que aconsegueixen un petit augment dels VP i una reducció dels FN respecte el model amb profunditat 1. Però el nombre de FP és lleugerament superior i, a més, presenta *overfitting*. No es recomana utilitzar una profunditat igual a 3 perquè té més *overfitting* i els resultats són pitjors. A la *Taula 4.1.2* es troba un resum dels valors dels hiperparàmetres recomanats.

Taula 4.1.2. Combinacions recomanades d'hiperparàmetres (Dataset A).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	[275, 300]	0.52	0.48	0.56
2	[125,175]	0.52	0.49	0.55

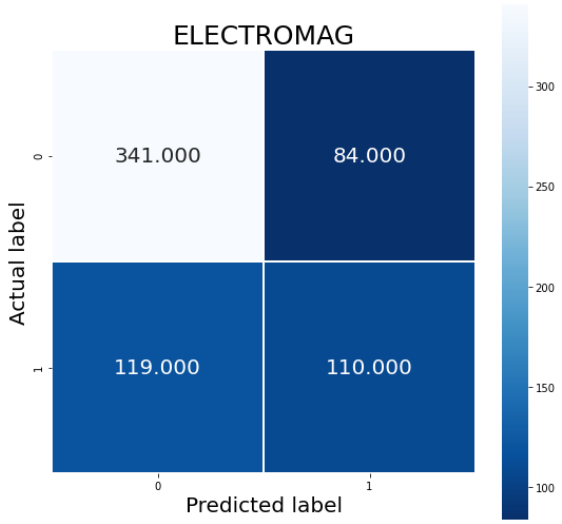


Figura 4.1.3. Matriu de confusió del model de Gradient Boosting amb profunditat 1 i 300 arbres (Dataset A).

Comparant el model de Gradient Boosting de profunditat 1 i 300 arbres amb la Regressió Logística, s'observa que els valors de F1 Score, Recall i Precision són lleugerament millors. A més, amb les matrius de confusió (vegeu *Figura 4.1.1 i Figura 4.1.3*) s'observa que el model de Gradient Boosting obté menys falsos negatius i el nombre de verdaders positius és més elevat. Però no hi ha gaire diferència entre els dos models pel que fa a la qualitat de predicció.

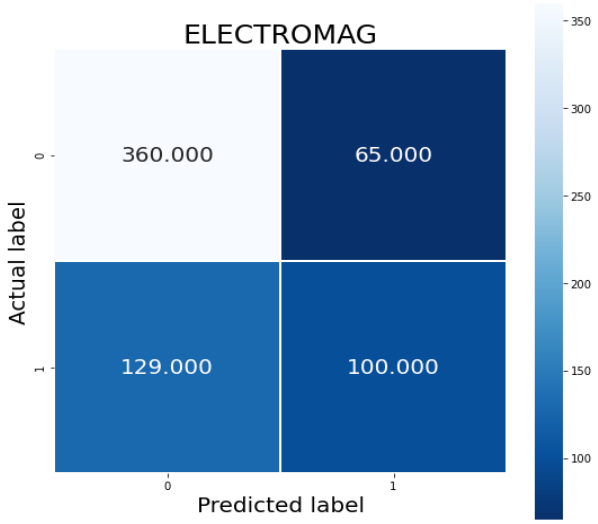


Figura 4.1.4. Matriu de confusió del model de Regressió Logística (Dataset B).

Un cop vist el rendiment dels models utilitzant el Dataset A, es procedirà a fer l'estudi amb el Dataset B. En aquest cas, el model de Regressió Logística té un F1 Score igual a 0.51, una Precision de 0.61 i un Recall de 0.44. A la *Figura 4.1.4* es pot observar la matriu de confusió corresponent. A més, aquest model no presenta *overffiting,* ja que els valors de F1 Score del grup d'entrenament i el d'avaluació difereixen per centèsimes.

Havent vist les característiques del model de Regressió Logística, es passarà a analitzar com es comporten els models de Gradient Boosting en funció dels hiperparàmetres. A la *Figura 4.1.5* es mostra com varia l'F1 Score en funció del nombre d'arbres i la profunditat. Els models que han obtingut els millors valors de les mètriques d'avaluació s'han recollit a la *Taula 4.1.3*.

Com es pot veure a la *Taula 4.1.3,* tots els models seleccionats són pràcticament idèntics pel que fa a la qualitat de les prediccions, ja que els valors obtinguts amb les mètriques d'avaluació són els mateixos. Pel que fa als hiperparàmetres, s'observa que quan augmenta la profunditat dels arbres de decisió, el model necessita menys arbres per obtenir els millors resultats. Per altra banda, es detecta que els models amb profunditat 2 i 3 pateixen un cert *overfitting*.

Taula 4.1.3. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

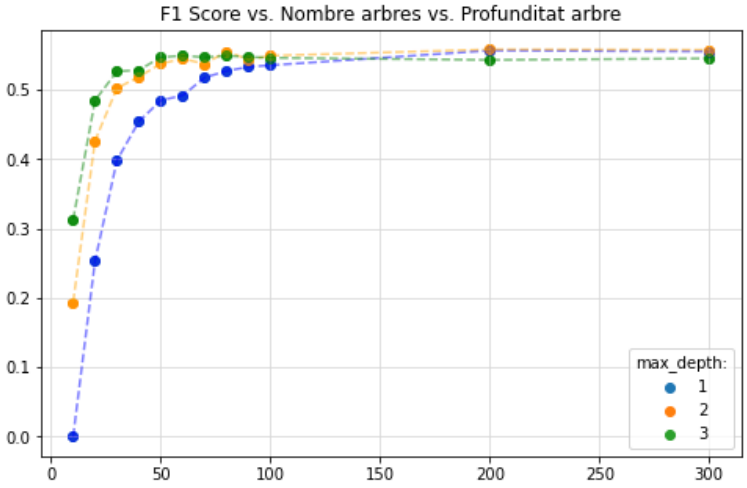
Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	200	0.56	0.57	0.50	0.62
	250	0.56	0.58	0.50	0.62
	275	0.56	0.58	0.50	0.62
2	175	0.58	0.68	0.53	0.63
	225	0.56	0.70	0.52	0.61
	275	0.56	0.72	0.52	0.61
3	55	0.56	0.68	0.50	0.62
	60	0.55	0.69	0.50	0.61
	50	0.55	0.67	0.49	0.61

Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió. Es recomana utilitzar un model amb profunditat 1 i 200 arbres (amb més arbres no milloren les prediccions) o escollir una profunditat 2 i 275 arbres. Les dues opcions proporcionen el mateix nombre de FP, però, en el cas de profunditat 2, el nombre de FN és lleugerament inferior i això es veu reflectit en un augment del nombre de VP predits. En aquest cas, s'hauria de considerar si l'augment de VP utilitzant profunditat 2 és prou significatiu per assumir l'*overfitting* existent. No es recomana utilitzar una profunditat igual a 3 perquè els valors de F1 Score, Recall i Precision són pitjors i, a més, presenta *overfitting*. A la *Taula 4.1.4* es troba un resum dels valors dels hiperparàmetres recomanats.

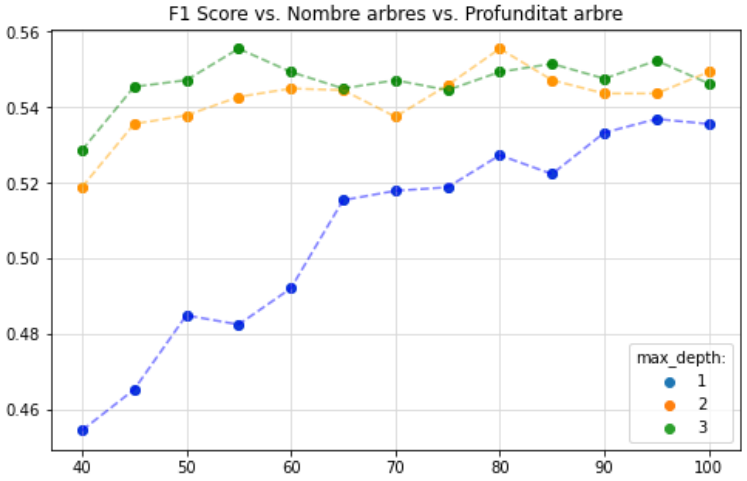
Taula 4.1.4. Combinacions recomanades d'hiperparàmetres (Dataset B).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	200	0.56	0.50	0.62
2	175	0.58	0.53	0.63

(a)



(b)



(c)

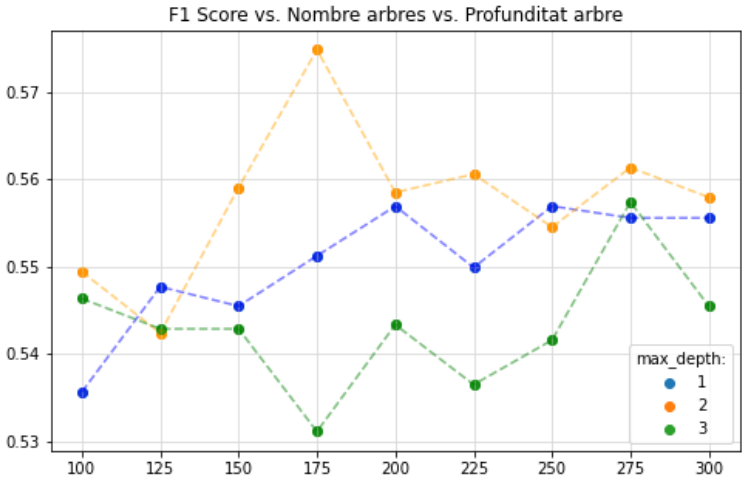


Figura 4.1.5. F1 Score en funció del nombre d'arbres i la profunditat (Dataset B). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 40 i 100. (c) Interval del nombre d'arbres ampliat entre 100 i 300.

Comparant el model de Gradient Boosting de profunditat 1 i 200 arbres amb la Regressió Logística, s'observa que els valors de F1 Score, Recall i Precision són millors. A més, amb les matrius de confusió (vegeu *Figura 7.1.4 i Figura 4.1.6*) s'observa que el model de Gradient Boosting obté bastants menys falsos negatius i el nombre de verdaders positius és més elevat. Així doncs, en aquest cas, es pot dir que el Gradient Boosting és millor que la Regressió Logística.

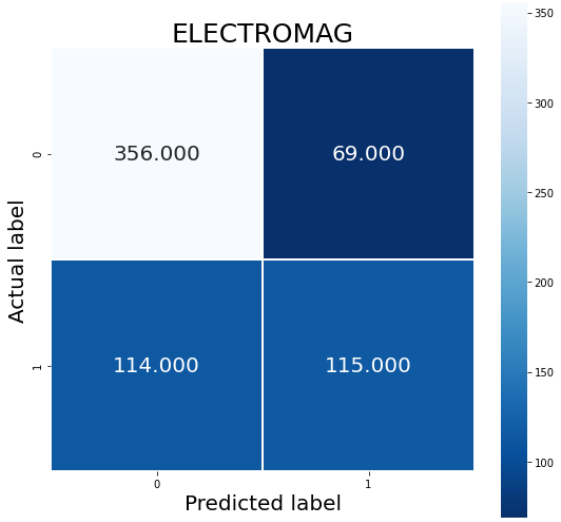


Figura 4.1.6. Matriu de confusió del model de Gradient Boosting amb profunditat 1 i 200 arbres (Dataset B).

Finalment, es realitzarà la comparació dels resultats obtinguts utilitzant el Dataset A i el Dataset B. Comparant els models recomanats, s'observa que utilitzant el Dataset B s'obtenen valors de F1 Score, Recall i Precision més grans. Per tant, amb el Dataset B s'aconsegueixen millors prediccions que amb el Dataset A. Això també es pot comprovar amb les matrius de confusió de la *Figura 4.1.3 i Figura 4.1.6* on es pot observar com varia el nombre de prediccions encertades en funció de quin Dataset s'escull. Per exemple, es pot veure que utilitzant el Dataset B el model prediu 15 FP menys que amb el Dataset A i encerta 5 VP més. Per altra banda, s'observa que amb una profunditat igual a 1 el model és menys complex i prediu millor si s'utilitza el Dataset B, ja que es necessiten 100 arbres menys que amb el Dataset A. Per tant, es podria dir que és millor utilitzar el Dataset B.

4.2. Mètodes Numèrics

A continuació s'estudiarà quins són els millors models per predir els suspesos de l'assignatura de Mètodes Numèrics. Per començar, es comprovarà quina és la proporció d'aprovats i suspesos de l'assignatura, per tal de veure si es disposa de dades equilibrades o no. S'ha trobat que un 88% de les dades corresponen a estudiants que van aprovar la primera convocatòria i un 12%

correspon a estudiants que la van suspendre. En aquest cas, es pot dir que les dades no estan gens equilibrades i costarà molt predir correctament els suspesos.

Primer de tot s'estudiaran els models obtinguts a partir del Dataset A. El model de Regressió Logística, que servirà de referència, obté un F1 Score, una Precision i un Recall igual a 0. Com es pot observar a la *Figura 4.2.1,* aquest model no és capaç de predir cap suspes correctament (VP). El model només ha predit aprovats. Això indica que ha sigut incapaç de superar el biaix generat a causa del desequilibri de les dades.

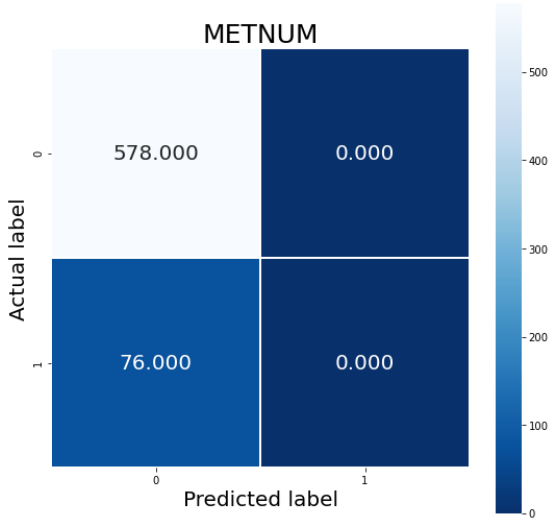


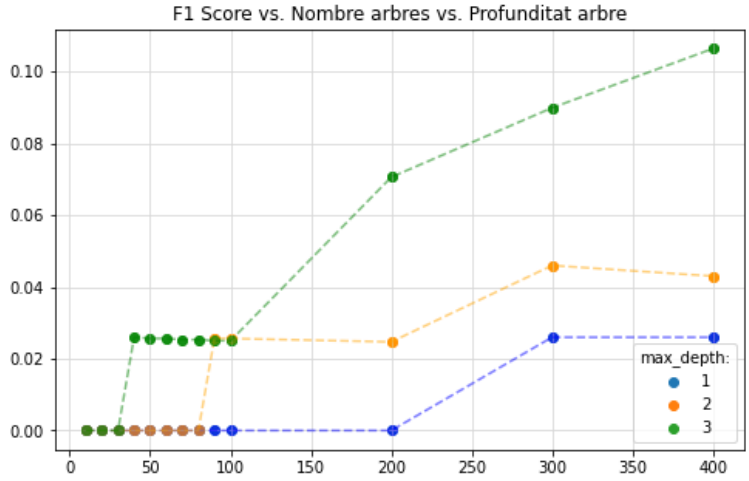
Figura 4.2.1. Matriu de confusió del model de Regressió Logística (Dataset A).

A continuació s'aplicarà el Gradient Boosting Classifier i s'avaluaran les diferents combinacions que s'han fet dels hiperparàmetres. A la *Figura 4.2.2* es mostra com varia l'F1 Score en funció del nombre d'arbres i la profunditat. Els models que han obtingut els millors valors de les mètriques d'avaluació contemplades s'han recollit a la *Taula 4.2.1*.

Taula 4.2.1. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	275 - 400	0.03	0.02 - 0.05	0.01	1.00
2	225	0.05	0.16	0.03	0.33
	250	0.05	0.21	0.03	0.29
3	275	0.11	0.75	0.07	0.36
	350	0.11	0.83	0.07	0.31

(a)



(b)

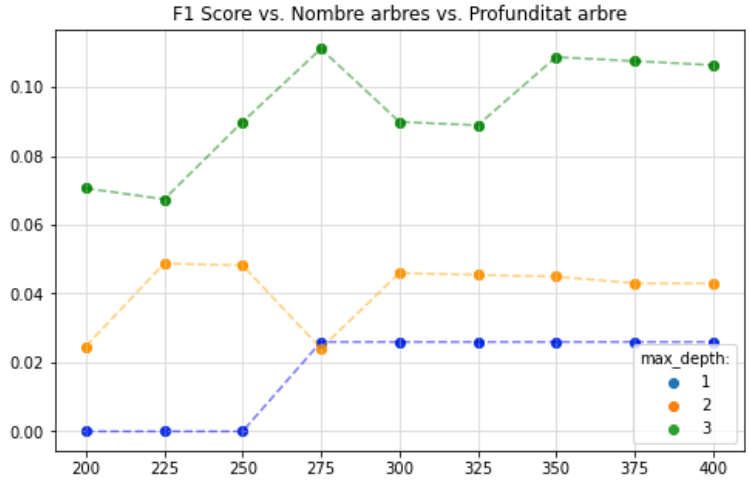


Figura 4.2.2. F1 Score en funció del nombre d'arbres i la profunditat (Dataset A). (a) Interval del nombre d'arbres entre 10 i 400. (b) Interval del nombre d'arbres ampliat entre 200 i 400.

En aquest cas, s'observa que amb profunditat 1 i 2 s'obté un F1 Score aproximadament igual a zero. Amb una profunditat igual a 3 el valor de F1 Score puja fins a 0.1, però segueix sent molt baix. A més, pateix molt *overfitting* i, per tant, el model no és gens fiable. L'aplicació del Gradient Boosting no ha permès millorar els resultats respecte als obtinguts amb la Regressió Logística. En conclusió, es pot deduir que, a causa de la mala qualitat de les dades i del fet que estan desequilibrades, no s'ha pogut trobar un model que aconsegueixi predir els suspesos.

Un cop fet l'estudi amb el Dataset A, es procedirà a analitzar el rendiment dels models utilitzant el Dataset B. En aquest cas, el model de Regressió Logística no presenta *overffiting* i té un F1 Score igual a 0.10, una Precision de 0.29 i un Recall de 0.05. A la *Figura 4.2.3* es pot observar la matriu de confusió corresponent.

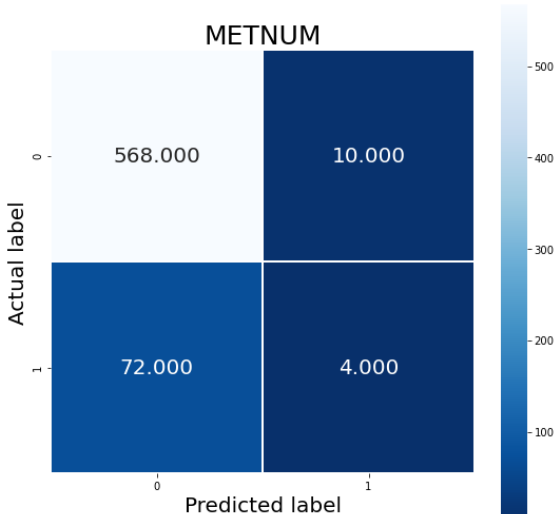


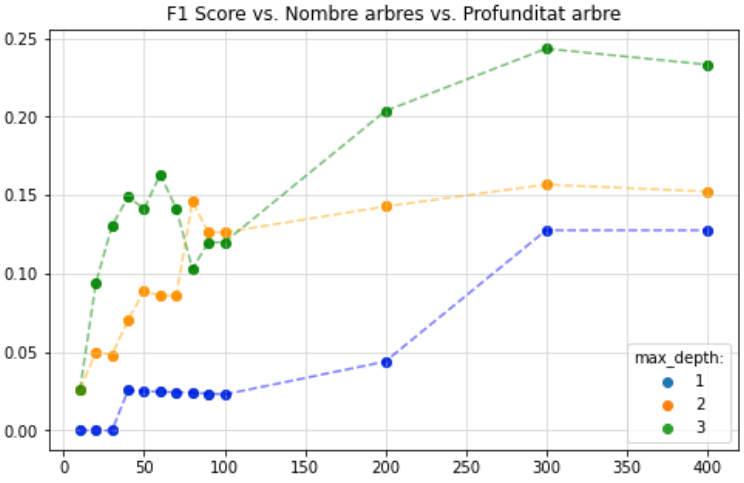
Figura 4.2.3. Matriu de confusió del model de Regressió Logística (Dataset B).

A continuació s'aplicarà el Gradient Boosting Classifier i s'avaluaran les diferents combinacions que s'han fet dels hiperparàmetres 'nombre d'arbres' i 'profunditat'. A la *Figura 4.2.4* es mostra com varia l'F1 Score en funció dels valors dels hiperparàmetres. Els models que han obtingut els millors valors de les mètriques d'avaluació contemplades s'han recollit a la *Taula 4.2.2*. Comparant els valors de F1 Score obtinguts al *train* i al test, es pot veure que tots els models pateixen *overfitting*, sobretot els de profunditat 3. L'existència d'*overfitting* fa que cap dels models seleccionats sigui fiable pel que fa a la qualitat de les prediccions. A més, els valors obtinguts de les mètriques d'avaluació són molt baixos.

Taula 4.2.2. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	300	0.13	0.30	0.08	0.33
	275	0.11	0.30	0.07	0.29
2	125	0.17	0.41	0.11	0.38
	250	0.16	0.55	0.11	0.32
3	275	0.25	0.84	0.18	0.37
	300	0.24	0.87	0.18	0.36

(a)



(b)

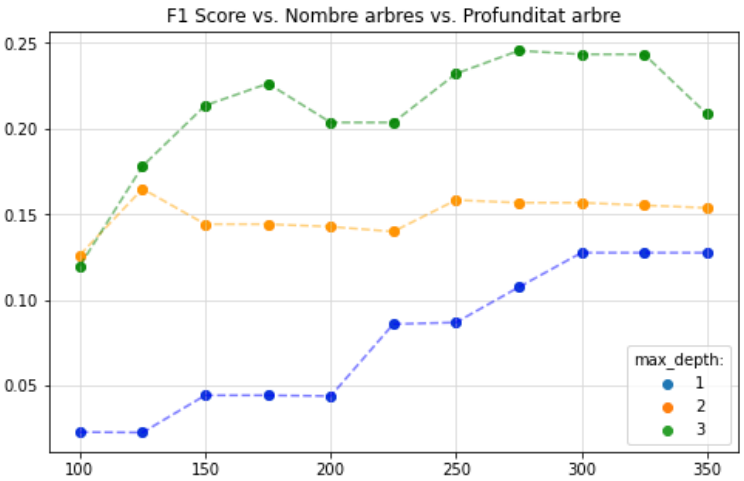


Figura 4.2.4. F1 Score en funció del nombre d'arbres i la profunditat (Dataset B). (a) Interval del nombre d'arbres entre 10 i 400. (b) Interval del nombre d'arbres ampliat entre 100 i 350.

Comparant els models de Gradient Boosting amb el de Regressió Logística, s'observa que els de Gradient Boosting tenen un rendiment de predicció superior, però pateixen *overfitting*. Concretament, utilitzat el Dataset B, els models amb profunditat 2 i 3 han obtingut uns valors de les mètriques d'avaluació bastant superiors als de la Regressió Logística. Es podrien aplicar els models de Gradient Boosting esmentats, però agafant els resultats amb pinces, ja que la presència d'*overfitting* fa que el rendiment estimat d'aquests models no sigui gaire fiable.

Per a l'assignatura de Mètodes Numèrics, tant si s'utilitza el Dataset A com si s'utilitza el Dataset B, no s'ha aconseguit predir gaires suspesos degut al desequilibri de les dades disponibles, que es detecta que és la principal causa que el rendiment dels models contemplats sigui tan baix. També podria ser que per a aquesta assignatura no existissin patrons clars que determinin qui és potencial de suspendre. Tot i així, es pot dir que és millor utilitzar el Dataset B.

4.3. Materials

A continuació s'estudiarà quins són els millors models per predir els suspesos de l'assignatura de Materials. Però, primer, es comprovarà quina és la proporció d'aprovats i suspesos de l'assignatura, per tal de veure si es disposa de dades equilibrades o no. S'ha trobat que un 65% de les dades corresponen a estudiants que van aprovar la primera convocatòria i un 35% corresponen a estudiants que la van suspendre. És possible que la qualitat dels resultats que s'obtindran per a l'assignatura de Materials sigui similar als obtinguts en l'estudi fet de l'assignatura d'Electromagnetisme, ja que en els dos casos la proporció de dades corresponents a estudiants aprovats i suspesos és la mateixa.

Primer de tot s'estudiaran els models obtinguts a partir del Dataset A. El model de Regressió Logística, que servirà de referència, té un valor de F1 Score igual a 0.51, una Precision de 0.58 i un Recall de 0.45. A més, s'ha vist que no hi ha *overfitting*, ja que l'F1 Score del grup d'entrenament i el d'avaluació tenen el mateix valor. A la *Figura 4.3.1* es pot observar la matriu de confusió d'aquest model.

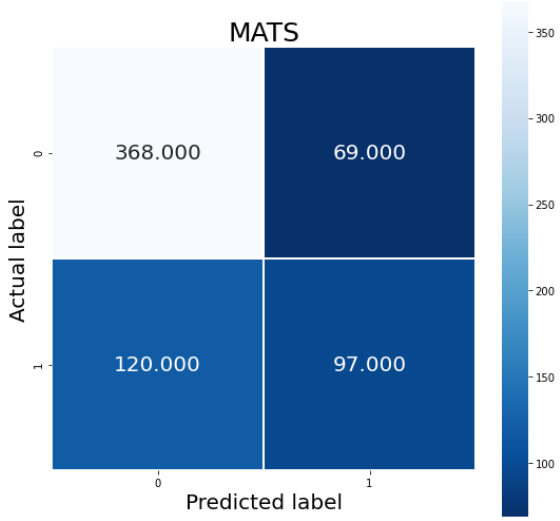
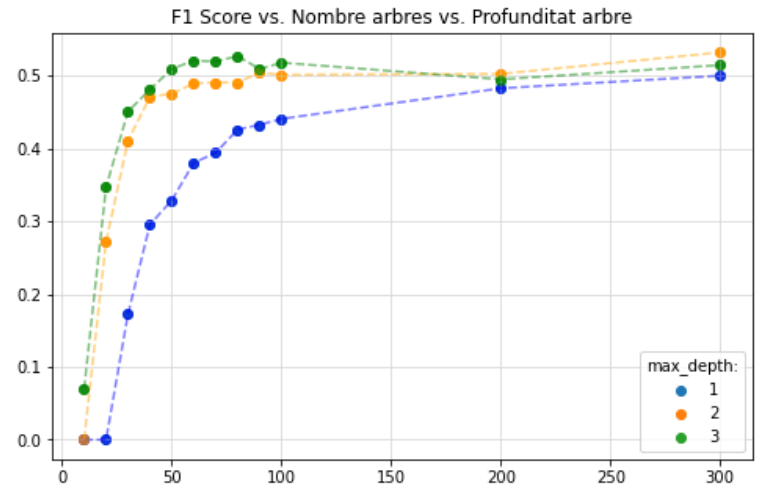


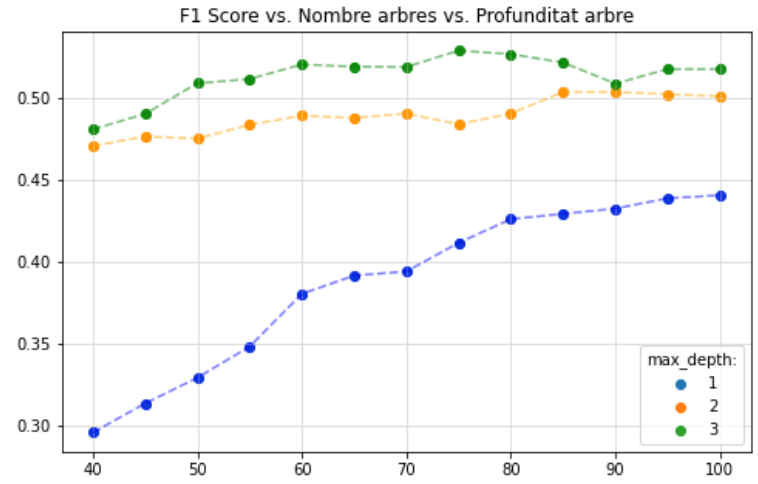
Figura 4.3.1. Matriu de confusió del model de Regressió Logística (Dataset A).

A continuació s'aplicarà el Gradient Boosting Classifier i s'avaluaran les diferents combinacions que s'han fet dels hiperparàmetres 'nombre d'arbres' i 'profunditat'. A la *Figura 4.3.2* es mostra com varia l'F1 Score en funció dels hiperparàmetres.

(a)



(b)



(c)

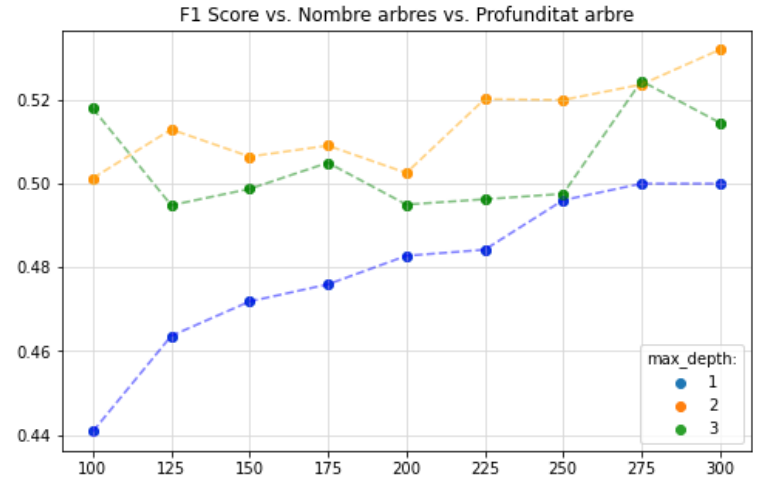


Figura 4.3.2. F1 Score en funció del nombre d'arbres i la profunditat (Dataset A). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 40 i 100. (c) Interval del nombre d'arbres ampliat entre 100 i 300.

Els models que han obtingut els millors valors de les mètriques d'avaluació contemplades s'han recollit a la *Taula 4.3.1.* Comparant els valors de F1 Score obtinguts al *train* i al test, s'observa que els models amb profunditat 1 no pateixen *overfitting* i, en canvi, els models de profunditat 2 i 3 en pateixen una mica.

Taula 4.3.1. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	275	0.50	0.47	0.44	0.58
	250	0.50	0.47	0.43	0.58
	225	0.48	0.47	0.42	0.56
2	300	0.53	0.63	0.50	0.57
	275	0.52	0.62	0.48	0.57
	250	0.52	0.60	0.48	0.57
3	75	0.53	0.62	0.48	0.59
	80	0.53	0.63	0.47	0.60
	85	0.52	0.63	0.47	0.59

Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió. Amb profunditat 1, es recomana utilitzar entre 250 i 275 arbres, ja que és el que obté el menor nombre de falsos positius (FP). Però no obté tants verdaders positius (VP) com altres models. En el cas d'escollir profunditat 2, es recomana utilitzar 300 arbres. Aquest model és una mica més complex respecte al primer perquè augmenta en profunditat i en nombre d'arbres però aconsegueix predir més VP (els FP també pugen en la mateixa proporció) i reduir el nombre de FN. Els models amb profunditat 3 necessiten molts menys arbres que la resta i aconsegueixen iguals o millors resultats. Concretament, si s'escull la profunditat 3 es recomana utilitzar 80 arbres, ja que té els mateixos FP que el model de profunditat 1 recomanat i prediu més VP. A la *Taula 4.3.2* es troba un resum dels valors dels hiperparàmetres recomanats, assumint el petit *overfitting* existent en alguna de les opcions.

Taula 4.3.2. Combinacions recomanades d'hiperparàmetres (Dataset A).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	[250, 275]	0.50	0.43	0.58
2	300	0.53	0.50	0.57
3	80	0.53	0.47	0.60

Comparant el model de Gradient Boosting de profunditat 3 i 80 arbres amb la Regressió Logística, s'observa que els valors de F1 Score, Recall i Precision són lleugerament millors. A més, amb les matrius de confusió (vegeu *Figura 4.3.1 i Figura 4.3.3*) s'observa que amb el model de Gradient Boosting augmenta el nombre de VP en la mateixa quantitat que es redueix el nombre de FN, però es manté la quantitat de FP. Es pot dir que el model de Gradient Boosting és lleugerament millor que el de Regressió Logística tot i que tenen pràcticament la mateixa qualitat de predicció.

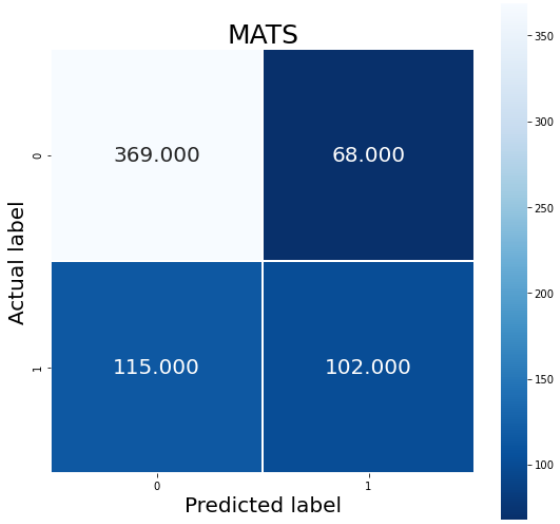


Figura 4.3.3. Matriu de confusió del model de Gradient Boosting amb profunditat 3 i 80 arbres (Dataset A).

Un cop vist el rendiment dels models utilitzant el Dataset A, es procedirà a fer l'estudi amb el Dataset B. En aquest cas, el model de Regressió Logística té un F1 Score igual a 0.50, una Precision de 0.58 i un Recall de 0.45. A més, aquest model no presenta *overffiting,* ja que els valors de F1 Score del grup d'entrenament i el d'avaluació difereixen per centèsimes. A la *Figura 4.3.4* es pot observar la matriu de confusió corresponent.

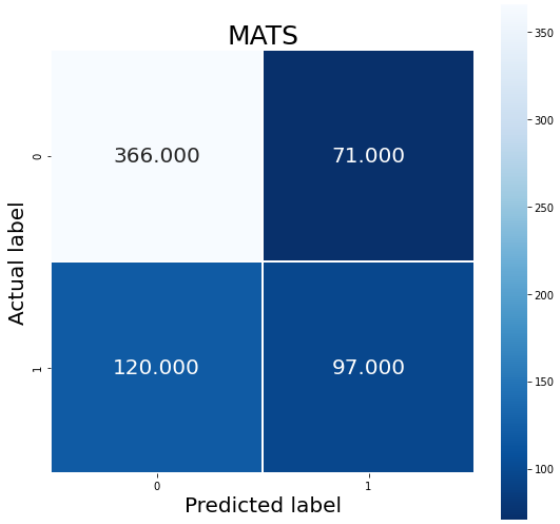


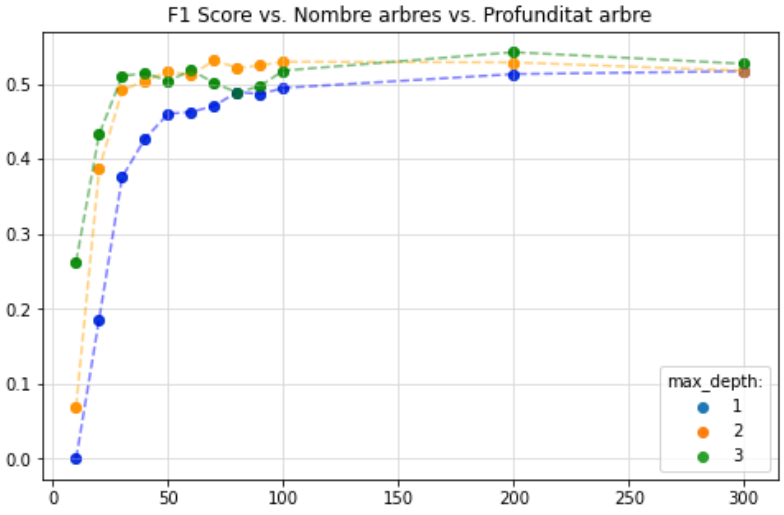
Figura 4.3.4. Matriu de confusió del model de Regressió Logística (Dataset B).

Havent vist les característiques del model de Regressió Logística, es passarà a analitzar com es comporten els models de Gradient Boosting en funció dels hiperparàmetres. A la *Figura 4.3.5* es mostra com varia l'F1 Score en funció del nombre d'arbres i la profunditat. Els models que han obtingut els millors valors de les mètriques d'avaluació s'han recollit a la *Taula 4.3.3*.

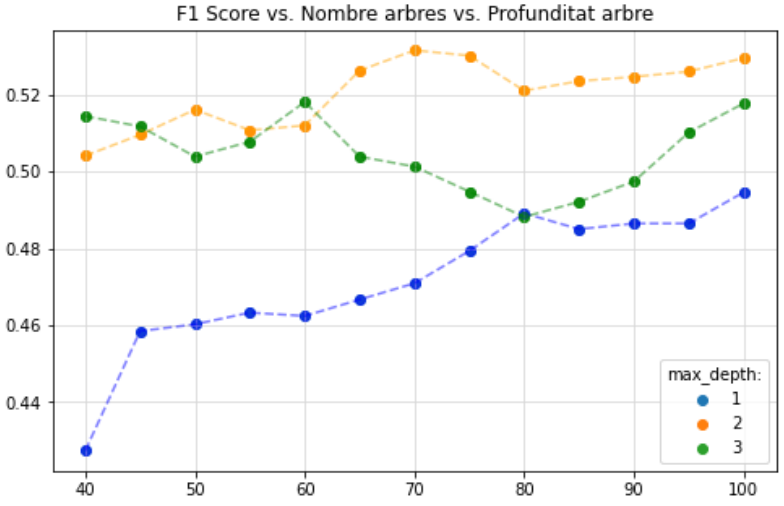
Taula 4.3.3. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	300	0.52	0.53	0.46	0.60
	250	0.52	0.52	0.46	0.59
	200	0.51	0.51	0.45	0.59
2	70	0.53	0.55	0.47	0.62
	95	0.53	0.57	0.47	0.60
	45	0.51	0.52	0.43	0.63
3	200	0.54	0.81	0.50	0.59
	175	0.53	0.78	0.49	0.58
	150	0.53	0.76	0.48	0.59

(a)



(b)



(c)

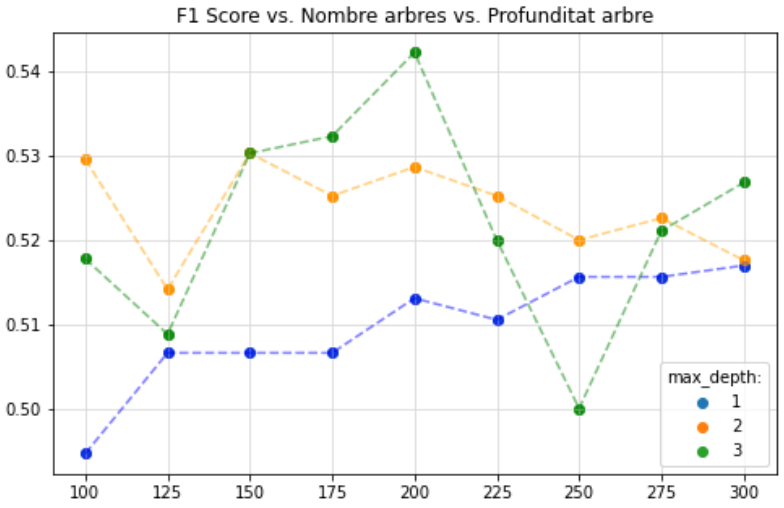


Figura 4.3.5. F1 Score en funció del nombre d'arbres i la profunditat (Dataset B). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 40 i 100. (c) Interval del nombre d'arbres ampliat entre 100 i 300.

Com es pot veure a la *Taula 4.3.3,* tots els models seleccionats són pràcticament idèntics pel que fa a la qualitat de les prediccions, ja que els valors obtinguts de les mètriques d'avaluació són els mateixos. Per altra banda, es detecta que els models amb profunditat 3 pateixen molt *overfitting* i, per tant són poc fiables i no es recomanen.

Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió. En aquest cas, els models amb la profunditat 1 necessiten molts més arbres respecte als models amb profunditat 2 i ni tan sols aconsegueixen millors resultats. Per això, no es recomanen i és preferible centrar-se en els models de profunditat 2. Escollint la profunditat 2, es recomana utilitzar 70 arbres, ja que aquesta combinació té els millors valors de Recall i Precision.

Comparant el model de Gradient Boosting de profunditat 2 i 70 arbres amb la Regressió Logística, s'observa que els valors de F1 Score, Recall i Precision són millors. A més, amb les matrius de confusió (vegeu *Figura 4.3.4 i Figura 4.3.6*) s'observa que amb el model de Gradient Boosting s'aconsegueix reduir el nombre de FN i sobretot de FP. A més, augmenta el nombre de VP. Així doncs, en aquest cas és més recomanable el Gradient Boosting que la Regressió Logística.

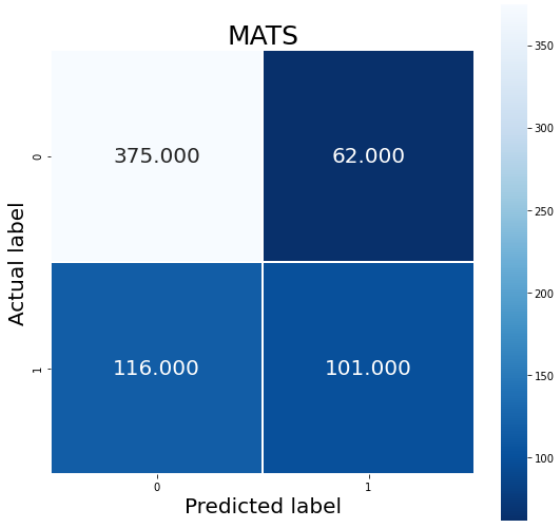


Figura 4.3.6. Matriu de confusió del model de Gradient Boosting amb profunditat 2 i 70 arbres (Dataset B).

Si es comparen els models amb millor rendiment predictiu obtinguts amb el Dataset A i el Dataset B, es pot observar que amb el Dataset B el model de profunditat 2 i 70 arbres prediu millor respecte el model de profunditat 3 i 80 arbres obtingut amb el Dataset A. I, a més de ser un model més senzill, no pateix *overfitting*. Per tant, és millor utilitzar el Dataset B.

4.4. Equacions Diferencials

A continuació s'estudiarà quins són els millors models per predir els suspesos de l'assignatura d'Equacions Diferencials. Per començar, es comprovarà quina és la proporció d'aprovats i suspesos de l'assignatura, per tal de veure si es disposa de dades equilibrades o no. S'ha trobat que no estan equilibrades, ja que que un 81% de les dades corresponen a estudiants que van aprovar la primera convocatòria i un 19% correspon a estudiants que la van suspendre. Aquesta proporció és similar a la de Mètodes Mumèrics i per això pot ser que els resultats dels dos estudis s'assemblin.

Primer de tot s'estudiaran els models obtinguts a partir del Dataset A. El model de Regressió Logística, que servirà de referència, té un F1 Score igual a 0.06, una Precision de 0.57 i un Recall de 0.03. A més, aquest model no presenta *overffiting,* ja que els valors de F1 Score del grup d'entrenament i el d'avaluació difereixen per centèsimes. A la *Figura 4.4.1* es mostra la matriu de confusió corresponent, on es pot observar que el nombre de VP és molt petit.

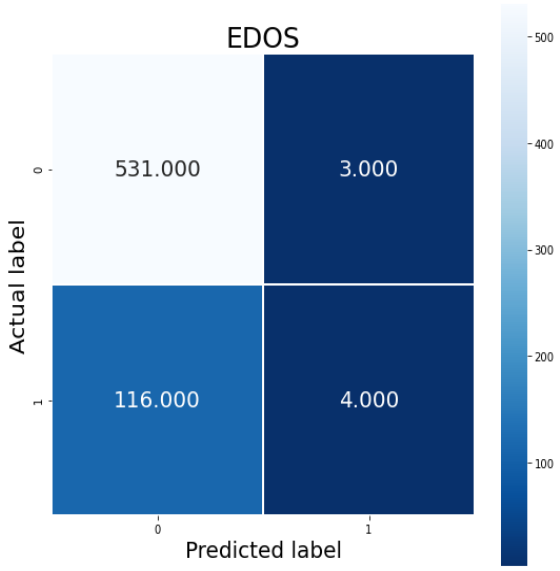
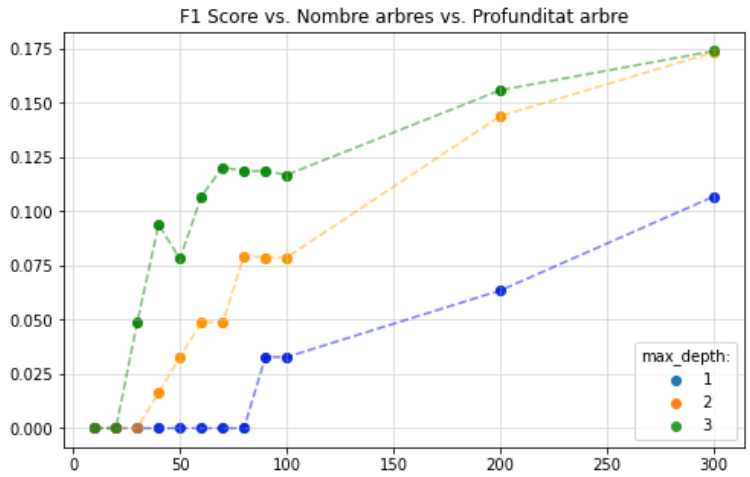


Figura 4.4.1. Matriu de confusió del model de Regressió Logística (Dataset A).

A continuació s'aplicarà el Gradient Boosting Classifier i s'avaluaran les diferents combinacions que s'han fet dels hiperparàmetres 'nombre d'arbres' i 'profunditat'. A la *Figura 4.4.2* es mostra com varia l'F1 Score en funció dels hiperparàmetres. Amb el gràfic *(a)* sembla que tots els models milloren si s'utilitza un nombre elevat d'arbres. Per veure quina tendència tenen si el nombre d'arbres és superior a 300, al gràfic *(b)* s'ha ampliat el rang fins a 400 arbres.

(a)



(b)

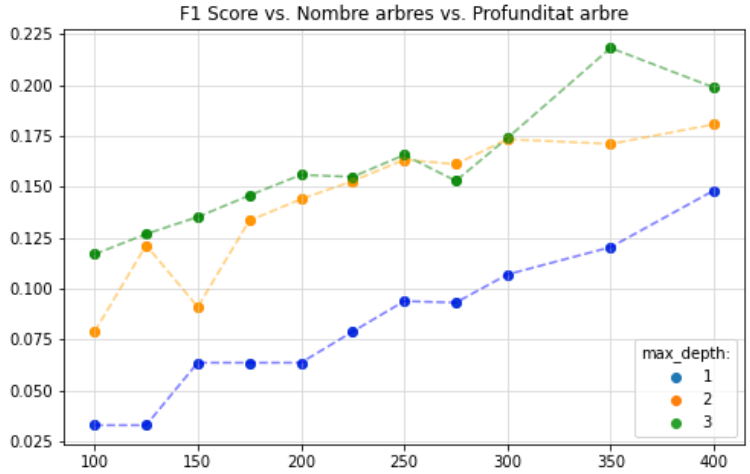


Figura 4.4.2. F1 Score en funció del nombre d'arbres i la profunditat (Dataset A). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 100 i 400.

Havent observat la *Figura 4.4.2,* s'han recollit a la *Taula 4.4.1* les combinacions d'hiperparàmetres que han obtingut els millors valors de F1 Score. Per comprovar l'existència d'*overfitting*, s'han comparat els valors de F1 Score del grup d'entrenament (*train*) amb el d'avaluació (test) i s'ha pogut observar que la diferència entre aquests dos valors augmenta a mesura que augmenta la profunditat. Per tant, com més profunditat tinguin els arbres més *overfitting* patirà el model. S'observa que els models amb profunditat 3 pateixen molt *overfitting* i, per tant, no són fiables i no es recomana utilitzar-los.

Taula 4.4.1. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	400	0.15	0.14	0.08	0.67
	350	0.12	0.15	0.07	0.62
	300	0.11	0.14	0.06	0.64
2	300	0.17	0.36	0.11	0.43
	250	0.16	0.30	0.10	0.44
	200	0.14	0.28	0.08	0.53
3	350	0.22	0.81	0.15	0.40
	300	0.17	0.75	0.12	0.34
	250	0.17	0.69	0.11	0.35

Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió i també comparant els valors de les mètriques d'avaluació. S'observa que els models amb la profunditat 1 tenen la Precision màxima però la resta de mètriques tenen valors pitjors respecte els models amb profunditat 2. Per altra banda, s'ha de tenir en compte que els models amb profunditat 1 són els únics que no pateixen *overfitting*. Si s'escull la profunditat 1, es recomana utilitzar entre 350 i 400 arbres. Amb 350 arbres els resultats empitjoren una mica, però, tenint en compte que redueix la complexitat del model, és una opció favorable. En el cas d'escollir la profunditat 2, assumint l'*overfitting* existent, es recomana utilitzar 250 arbres, ja que és la combinació que prediu més VP. Però té un nombre de FP molt superior al model proposat de profunditat 1. A la *Taula 4.4.2* es troba un resum de les combinacions dels hiperparàmetres recomanades.

Taula 4.4.2. Combinacions recomanades d'hiperparàmetres (Dataset A).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	[350, 400]	0.12	0.07	0.62
2	250	0.16	0.10	0.44

Comparant el model de Gradient Boosting de profunditat 1 i 350 arbres amb la Regressió Logística, s'observa que els valors de les mètriques són millors, sobretot els de F1 Score i Precision. A més, amb les matrius de confusió (vegeu *Figura 4.4.1 i Figura 4.4.3*) s'observa que amb el model de Gradient Boosting es duplica el nombre de VP, reduint el nombre de FN. Es pot dir que el model de Gradient Boosting és lleugerament millor que el de Regressió Logística.

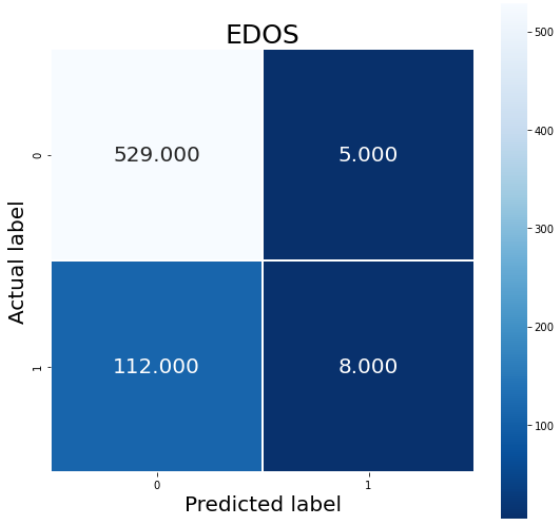


Figura 4.4.3. Matriu de confusió del model de Gradient Boosting amb profunditat 1 i 350 arbres (Dataset A).

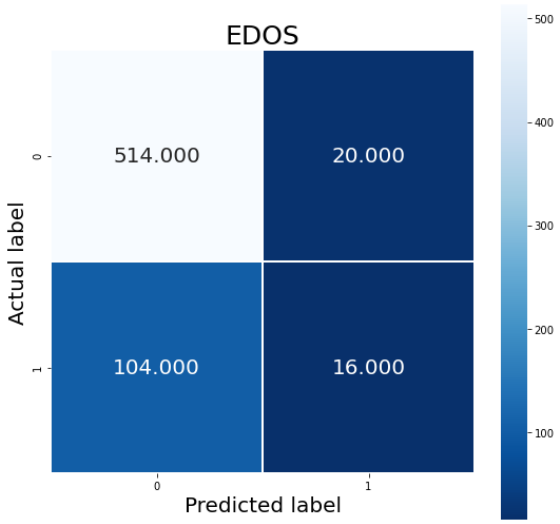
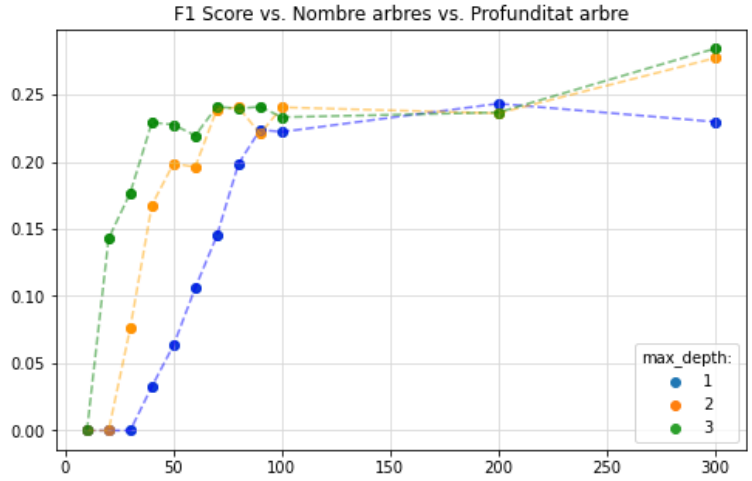


Figura 4.4.4. Matriu de confusió del model de Regressió Logística (Dataset B).

Un cop vist el rendiment dels models utilitzant el Dataset A, es procedirà a fer l'estudi amb el Dataset B. En aquest cas, el model de Regressió Logística té un F1 Score igual a 0.21, una Precision de 0.53 i un Recall de 0.23. A més, aquest model no presenta *overffiting,* ja que els valors de F1 Score del grup d'entrenament i el d'avaluació difereixen per centèsimes. A la *Figura 4.4.4* es pot observar la matriu de confusió corresponent.

A continuació s'analitzarà com es comporten els models de Gradient Boosting en funció dels hiperparàmetres. A la *Figura 4.4.5* es mostra com varia l'F1 Score en funció del nombre d'arbres i la profunditat. Els models que han obtingut els millors valors de les mètriques d'avaluació s'han recollit a la *Taula 4.4.3*.

(a)



(b)

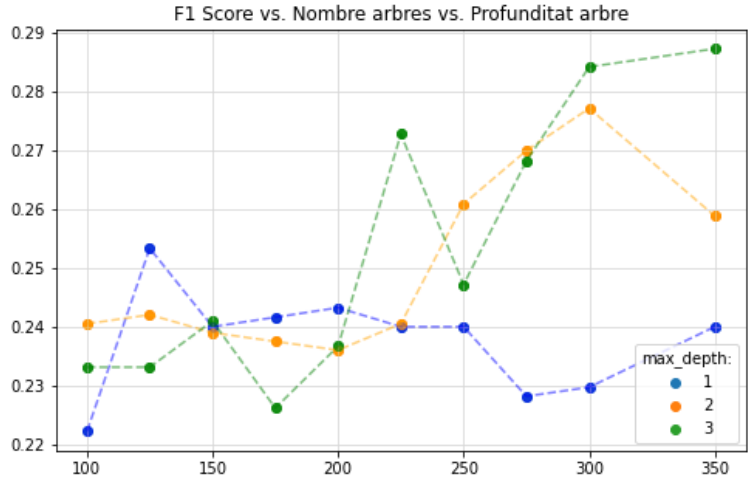


Figura 4.4.5. F1 Score en funció del nombre d'arbres i la profunditat (Dataset B). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 100 i 350.

Taula 4.4.3. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	125	0.25	0.25	0.15	0.63
	200	0.24	0.29	0.15	0.64
	175	0.24	0.27	0.15	0.62
2	300	0.28	0.60	0.19	0.50
	275	0.27	0.58	0.18	0.51
	250	0.26	0.55	0.17	0.51
3	300	0.28	0.87	0.22	0.41
	225	0.27	0.80	0.20	0.43
	275	0.27	0.85	0.20	0.41

Per comprovar l'existència d'*overfitting*, s'ha comparat el valor de F1 Score de *train* amb el de test i s'ha pogut observar que la diferència entre aquests dos valors augmenta a mesura que augmenta la profunditat. Per tant, com més profunditat tinguin els arbres més *overfitting* patirà el model. S'observa que els models amb profunditat 1 no pateixen *overfitting* i els models de profunditat 2 i 3 sí que en pateixen. Concretament no es recomana utilitzar els models amb profunditat 3.

Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió i també comparant els valors de les mètriques d'avaluació. Si interessa minimitzar el nombre de FP es recomana utilitzar la profunditat 1 i 125 arbres, ja que té una Precision molt elevada. Amb la profunditat 2 s'aconsegueixen resultats similars si s'utilitza entre 250 i 275. Però s'ha de tenir en compte que pateix cert *overfitting.* En canvi, el model amb profunditat 1 no pateix *overfitting* i és més senzill. A la *Taula 4.4.4* es troba un resum de les combinacions dels hiperparàmetres recomanades.

Taula 4.4.4. Combinacions recomanades d'hiperparàmetres (Dataset B).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	125	0.25	0.15	0.63
2	[250, 275]	0.26	0.17	0.51

Comparant el model de Gradient Boosting de profunditat 1 i 125 arbres amb la Regressió Logística, s'observa que els valors de les mètriques són millors, sobretot el de Precision. A més, amb les matrius de confusió (vegeu *Figura 4.4.4 i Figura 4.4.6*) s'observa que amb el model de Gradient Boosting es redueix el nombre de FP i de FN. També augmenta una mica el nombre de VP. Així doncs, en aquest cas es pot dir que és millor utilitzar el Gradient Boosting que la Regressió Logística.

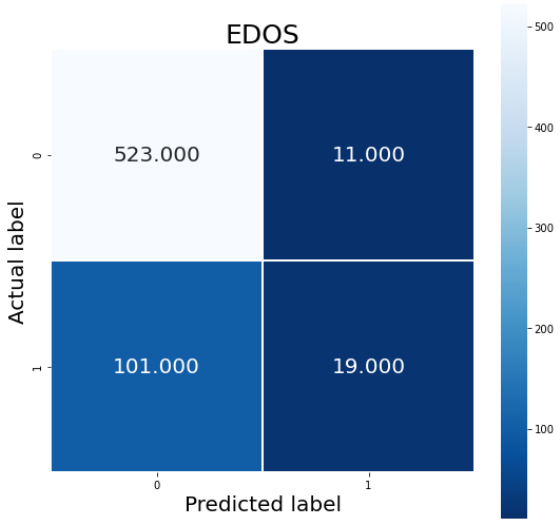


Figura 4.4.6. Matriu de confusió del model de Gradient Boosting amb profunditat 1 i 125 arbres (Dataset B).

Finalment, es realitzarà la comparació dels resultats obtinguts utilitzant el Dataset A i el Dataset B. Concretament es compararan els models recomanats, que tenen el millor rendiment predictiu. Utilitzant el Dataset B i escollint una profunditat 1, s'aconsegueix duplicar el nombre de VP predits respecte al model de profunditat 1 obtingut amb el Dataset A. També baixa molt el nombre de FN però augmenta el nombre de FP (vegeu *Figura 4.4.3 i Figura 4.4.6*). A més el model és més senzill ja que necessita 125 arbres, que és molt menys comparat amb el model del Dataset A, que en necessita 350. Així doncs, es pot dir que utilitzant el Dataset B s'aconsegueix millorar la qualitat de prediccions i, per tant, es recomana utilitzar-lo.

4.5. Informàtica

A continuació s'estudiarà quins són els millors models per predir els suspesos de l'assignatura d'Informàtica. Primer de tot, es comprovarà quina és la proporció d'aprovats i suspesos de l'assignatura, per tal de veure si es disposa de dades equilibrades o no. S'ha trobat que un 78% de les dades corresponen a estudiants que van aprovar a la primera convocatòria i un 22% correspon a estudiants que la van suspendre. Així doncs, s'ha comprovat que les dades no estan

gaire equilibrades i, per tant, serà difícil predir correctament els estudiants potencials de suspendre l'assignatura.

Primer de tot s'estudiaran els models obtinguts a partir del Dataset A. El model de Regressió Logística, que servirà de referència, té un valor de F1 Score igual a 0.19, una Precision de 0.52 i un Recall de 0.12. A la *Figura 4.5.1* es pot observar la matriu de confusió d'aquest model. A més, per comprovar si hi ha *overfitting,* s'han comparat els valors de F1 Score del grup d'entrenament amb el d'avaluació i s'ha trobat que són iguals. Per tant, aquest model no presenta *overffiting*. Totes aquestes característiques es compararan posteriorment amb les obtingudes aplicant el Gradient Boosting.

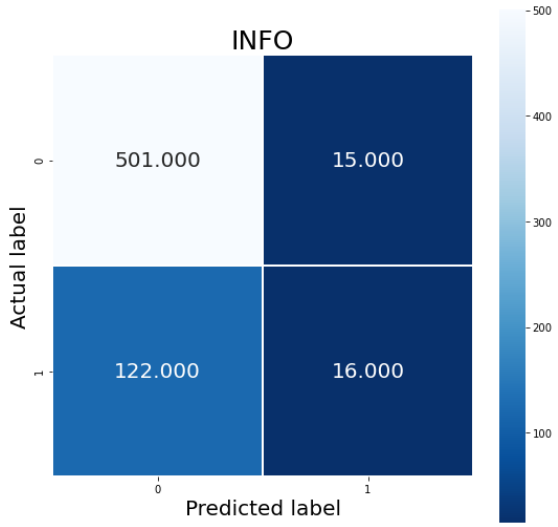
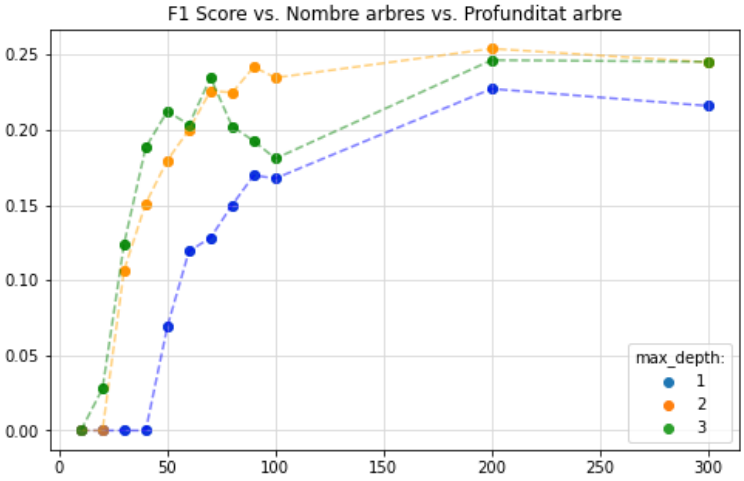


Figura 4.5.1. Matriu de confusió del model de Regressió Logística (Dataset A).

A continuació s'aplicarà el Gradient Boosting Classifier i s'avaluaran les diferents combinacions que s'han fet dels hiperparàmetres 'nombre d'arbres' i 'profunditat'. A la *Figura 4.5.2* es mostra com varia l'F1 Score en funció dels hiperparàmetres.

Havent observat la *Figura 4.5.2,* s'han recollit a la *Taula 4.5.1* les combinacions d'hiperparàmetres que han obtingut els millors valors de F1 Score. Per comprovar l'existència d'*overfitting*, s'han comparat els valors de F1 Score del grup d'entrenament (*train*) amb el d'avaluació (test) i s'ha pogut observar que la diferència entre aquests dos valors augmenta a mesura que augmenta la profunditat. Per tant, com més profunditat tinguin els arbres més *overfitting* patirà el model.

(a)



(b)

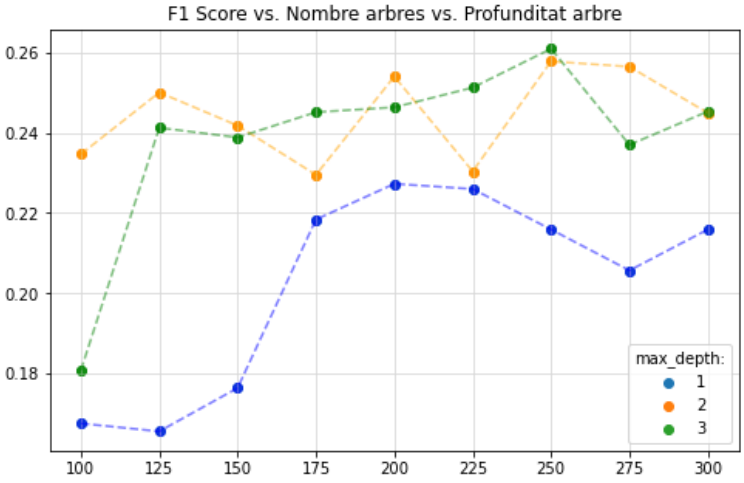


Figura 4.5.2. F1 Score en funció del nombre d'arbres i la profunditat (Dataset A). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 100 i 300.

Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió i també comparant els valors de les mètriques d'avaluació de la *Taula 4.5.1*. Si interessa minimitzar el nombre de FP, es recomanen els models de profunditat 1 amb un nombre d'arbres entre 200 i 300, ja que dins aquest interval les variacions dels valors de Precision i Recall són ínfimes. En el cas de voler el mínim de FN possibles, es recomana utilitzar la profunditat 2 i un nombre d'arbres entre 200 i 250, tenint en compte que hi ha una mica d'*overfitting*. Pel que fa als models de profunditat 3, dues de les opcions proposades pateixen molt *overfitting* i, per tant, es descarten. Amb la profunditat 3 es recomana utilitzar 70 arbres, ja que amb un nombre molt més petit d'arbres aconsegueix els mateixos resultats que els models de profunditat 1. Però s'hauria d'assumir l'*overfitting* existent. A la *Taula 4.5.2* es troba un resum dels models recomanats i les seves propietats.

Taula 4.5.1. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	200	0.23	0.23	0.15	0.53
	225	0.23	0.23	0.14	0.51
	300	0.22	0.25	0.14	0.50
2	250	0.26	0.44	0.18	0.45
	200	0.25	0.41	0.17	0.47
	300	0.25	0.47	0.17	0.41
3	250	0.26	0.75	0.20	0.39
	200	0.25	0.68	0.18	0.38
	70	0.23	0.42	0.15	0.51

Taula 4.5.2. Combinacions recomanades d'hiperparàmetres (Dataset A).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	[200, 300]	0.23	0.14	0.51
2	[200, 250]	0.25	0.17	0.47
3	70	0.23	0.15	0.51

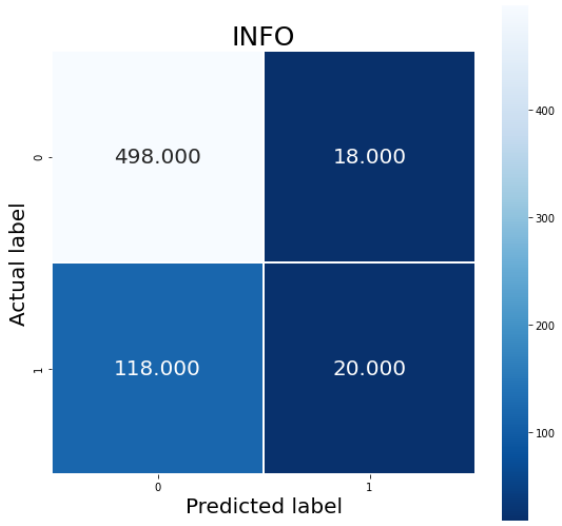


Figura 4.5.3. Matriu de confusió del model de Gradient Boosting amb profunditat 1 i 200 arbres (Dataset A).

Comparant el model de Gradient Boosting de profunditat 1 i 200 arbres amb la Regressió Logística, s'observa que els valors de les mètriques són lleugerament millors. A més, amb les matrius de confusió (vegeu *Figura 4.5.1 i Figura 4.5.3*) s'observa que amb el model de Gradient Boosting augmenta el nombre de VP en la mateixa mesura que es redueix el nombre de FN. Així doncs, es pot dir que utilitzant el Gradient Boosting s'aconsegueixen millors resultats que amb la Regressió Logística.

Un cop vist el rendiment dels models utilitzant el Dataset A, es procedirà a fer l'estudi amb el Dataset B. En aquest cas, el model de Regressió Logística té un F1 Score igual a 0.32, una Precision de 0.53 i un Recall de 0.23. A la *Figura 4.5.4* es pot observar la matriu de confusió corresponent. A més, aquest model no presenta *overffiting,* ja que els valors de F1 Score del grup d'entrenament i el d'avaluació difereixen per centèsimes.

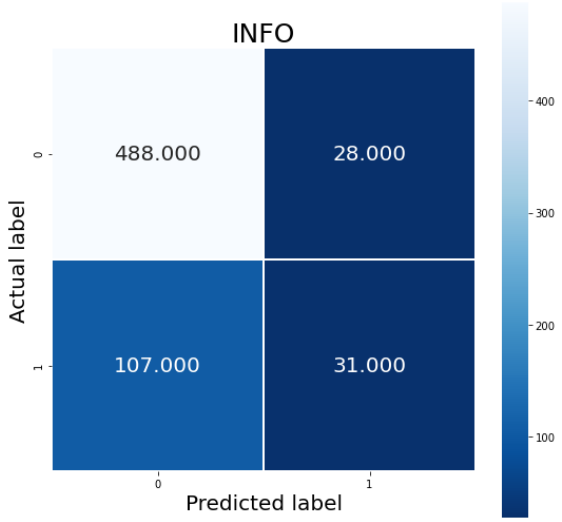


Figura 4.5.4. Matriu de confusió del model de Regressió Logística (Dataset B).

A continuació s'analitzarà com es comporten els models de Gradient Boosting en funció dels hiperparàmetres. A la *Figura 4.5.5* es mostra com varia l'F1 Score en funció del nombre d'arbres i la profunditat. Els models que han obtingut els millors valors de les mètriques d'avaluació s'han recollit a la *Taula 4.5.3.*

Per comprovar l'existència d'*overfitting*, s'ha comparat el valor de F1 Score de *train* amb el de test i s'ha pogut observar que tots els models pateixen *overfitting* amb més o menys mesura. Seguidament, s'ha valorat quins són els millors models d'entre els seleccionats fent una anàlisi detallada a partir de les matrius de confusió i també comparant els valors de les mètriques d'avaluació. S'ha de tenir en compte que tots els models seleccionats tenen valors molt similars de les mètriques d'avaluació i, per tant, en aquest aspecte són pràcticament idèntics.

Taula 4.5.3. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

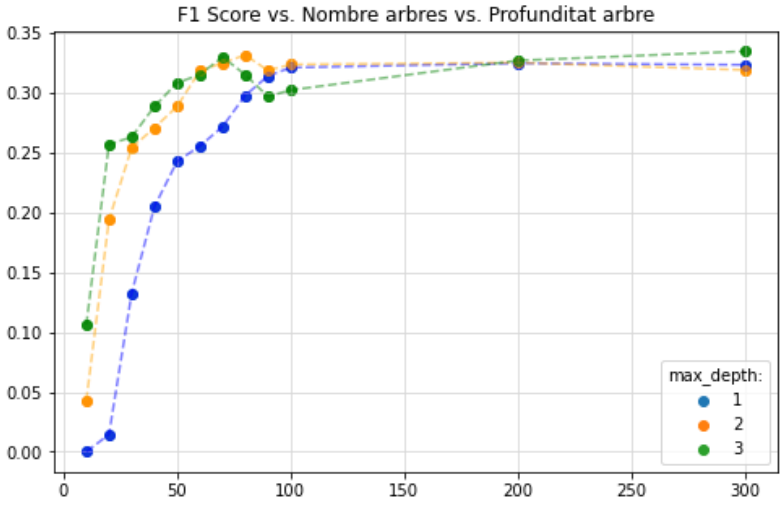
Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	250	0.33	0.45	0.24	0.52
	200	0.33	0.45	0.23	0.54
	100	0.32	0.42	0.22	0.56
2	80	0.33	0.50	0.25	0.51
	70	0.33	0.48	0.24	0.51
	150	0.33	0.56	0.25	0.49
3	250	0.36	0.83	0.28	0.48
	275	0.34	0.86	0.27	0.46
	70	0.33	0.57	0.25	0.50

Escollint una profunditat igual a 1 i assumint el petit *overfitting* existent, es recomana utilitzar entre 100 i 200 arbres. Aquesta combinació d'hiperparàmetres minimitza el nombre de FP predits. En el cas dels models amb profunditat 2, es recomana utilitzar 70 arbres, descartant la resta d'opcions perquè presenten massa *overfitting*. Amb aquesta combinació d'hiperparàmetres s'aconsegueix un model més senzill, ja que el nombre d'arbres necessaris és molt petit. Amb una profunditat igual a 3, els models pateixen massa *overfitting* i, per això, no es recomana utilitzar-la. Sembla que es podria utilitzar una profunditat 3 i 70 arbres, però aquesta combinació no millora els resultats respecte al model de profunditat 2 recomanat i, per tant, no val la pena augmentar la profunditat. A la *Taula 4.5.4* es troba un resum de les combinacions dels hiperparàmetres recomanades.

Taula 4.5.4. Combinacions recomanades d'hiperparàmetres (Dataset B).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	[100, 200]	0.32	0.22	0.56
2	70	0.33	0.24	0.51

(a)



(b)

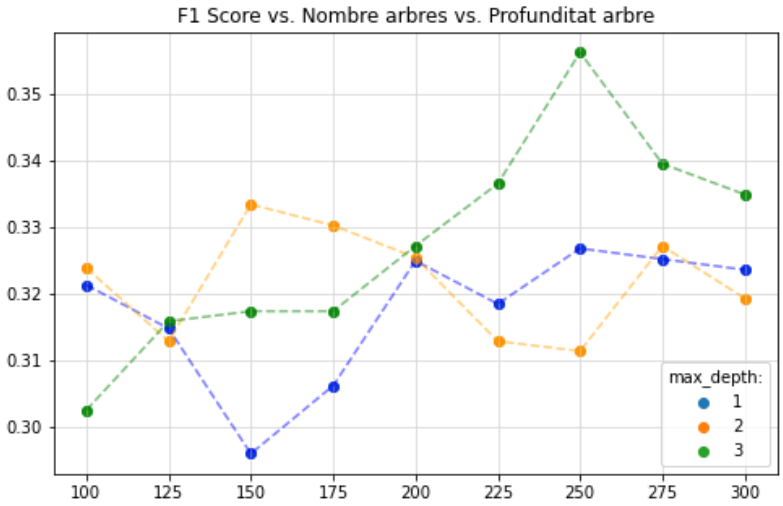


Figura 4.5.5. F1 Score en funció del nombre d'arbres i la profunditat (Dataset B). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 100 i 300.

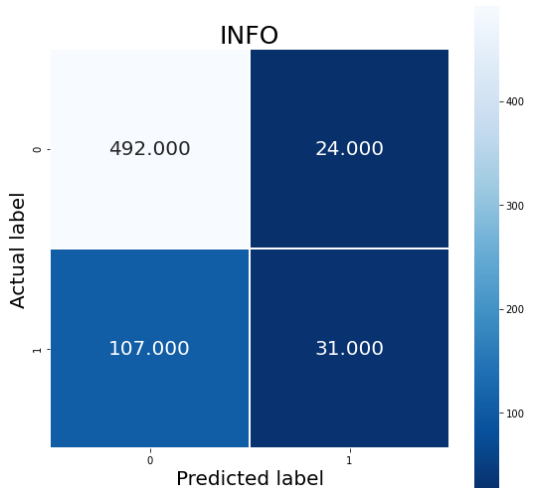


Figura 4.5.6. Matriu de confusió del model de Gradient Boosting amb profunditat 1 i 100 arbres (Dataset B).

Comparant el model de Gradient Boosting de profunditat 1 i 100 arbres amb la Regressió Logística, s'observa que els valors de les mètriques són millors, sobretot el de Precision. A més, amb les matrius de confusió (vegeu *Figura 4.5.4 i Figura 4.5.6*) s'observa que amb el model de Gradient Boosting es redueix el nombre de FP, però els nombres de VP i FN no canvien. Així doncs, es pot dir que el model de Gradient Boosting i el de Regressió Logística tenen una qualitat de predicció similar.

Finalment, es realitzarà la comparació dels resultats obtinguts utilitzant el Dataset A i el Dataset B. Comparant els models recomanats amb profunditat 1, es pot veure que utilitzant el Dataset B s'han necessitat 100 arbres menys, fet que redueix la complexitat del model. A més, aconsegueix reduir el nombre de FN i, en la mateixa mesura, augmenta el nombre de VP (vegeu *Figura 4.5.3 i Figura 4.5.6*). Així doncs, es pot dir que utilitzant el Dataset B s'aconsegueix que amb models més senzills millori la qualitat de prediccions i, per tant, és més recomanable que el Dataset A.

4.6. Mecànica

A continuació s'estudiarà quins són els millors models per predir els suspesos de l'assignatura de Mecànica. Per començar, es comprovarà quina és la proporció d'aprovats i suspesos de l'assignatura, per tal de veure si es disposa de dades equilibrades o no. S'ha trobat que un 53% de les dades corresponen a estudiants que van aprovar la primera convocatòria i un 47% correspon a estudiants que la van suspendre. Aquesta assignatura és amb diferència la que té les dades més equilibrades i cal esperar que s'aconseguiran resultats satisfactoris pel que fa a la qualitat de les prediccions.

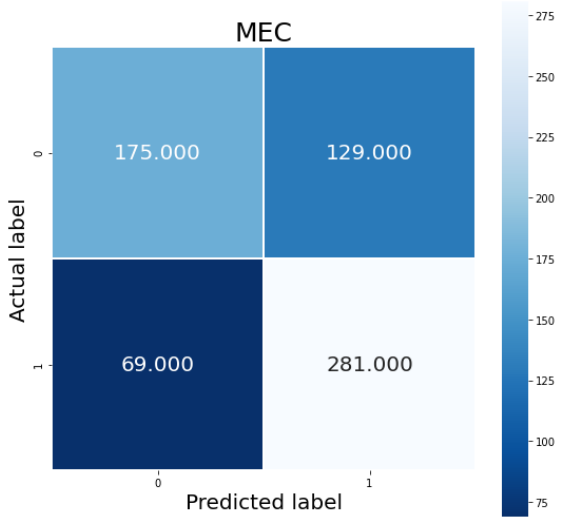


Figura 4.6.1. Matriu de confusió del model de Regressió Logística (Dataset A).

Primer de tot s'estudiaran els models obtinguts a partir del Dataset A. El model de Regressió Logística, que servirà de referència, té un valor de F1 Score igual a 0.74, una Precision de 0.69 i un Recall de 0.80. A la *Figura 4.6.1* es pot observar la matriu de confusió d'aquest model. A més, per comprovar si hi ha *overfitting,* s'han comparat els valors de F1 Score del grup d'entrenament amb el d'avaluació i s'ha trobat que són iguals. Per tant, aquest model no presenta *overffiting*.

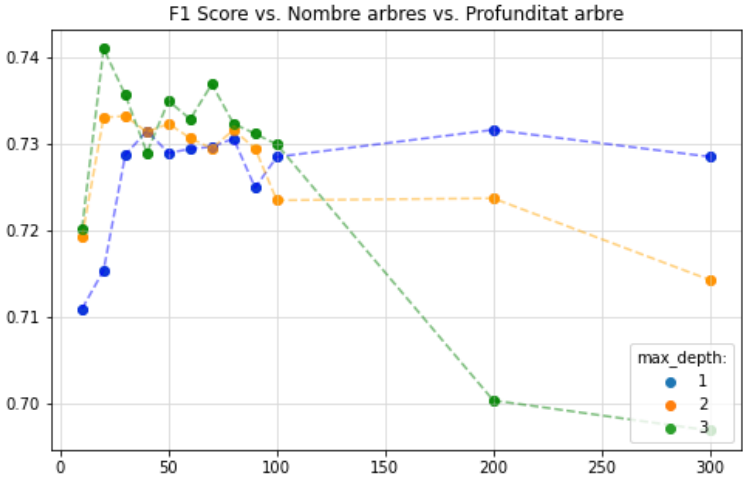
A continuació s'aplicarà el Gradient Boosting Classifier i s'avaluaran les diferents combinacions que s'han fet dels hiperparàmetres 'nombre d'arbres' i 'profunditat'. A la *Figura 4.6.2* es mostra com varia l'F1 Score en funció dels hiperparàmetres. Els models que han obtingut els millors valors de les mètriques d'avaluació s'han recollit a la *Taula 4.6.1*.

Taula 4.6.1. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

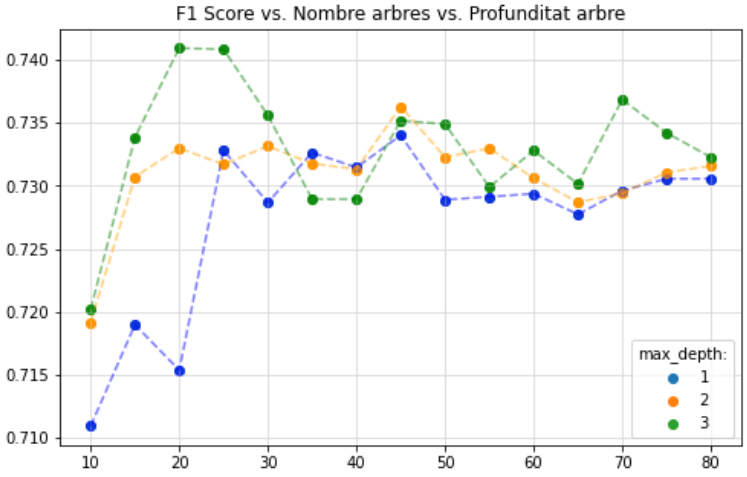
Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	100	0.73	0.72	0.80	0.67
	200	0.73	0.73	0.79	0.68
	250	0.73	0.73	0.79	0.68
2	45	0.74	0.74	0.81	0.68
	55	0.73	0.74	0.80	0.68
	20	0.73	0.73	0.83	0.66
3	20	0.74	0.75	0.82	0.68
	25	0.74	0.76	0.81	0.68
	30	0.74	0.76	0.80	0.68

Comparant els valors de F1 Score de *train* i test de la *Taula 4.6.1* s'ha pogut observar que cap dels models seleccionats pateix *overfitting*. Per poder detectar quins són els millors models d'entre els seleccionats, s'ha fet una anàlisi detallada a partir de les matrius de confusió i també comparant els valors de les mètriques d'avaluació. Escollint una profunditat igual a 1, es recomana utilitzar entre 100 i 200 arbres. Encara que aquesta opció no és preferible, ja que el nombre d'arbres necessari és elevat i no obté millors resultats que la resta de models, que són més simples. Amb una profunditat igual a 2 es recomana utilitzar 45 arbres si es vol minimitzar el nombre de FP. En cas de preferir minimitzar el nombre de FN es recomana utilitzar 20 arbres. Utilitzant una profunditat de 3 i un nombre d'arbres entre 20 i 30 també s'obtenen bons resultats.

(a)



(b)



(c)

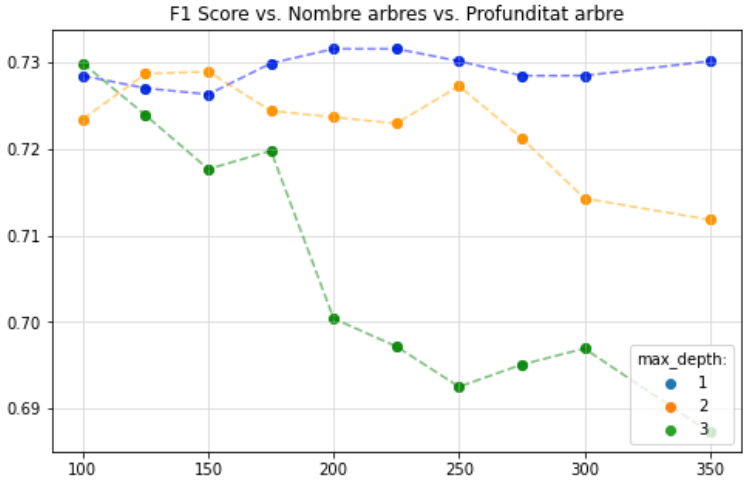


Figura 4.6.2. F1 Score en funció del nombre d'arbres i la profunditat (Dataset A). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 10 i 80. (c) Interval del nombre d'arbres ampliat entre 100 i 350.

Com que totes les combinacions proposades són molt similars pel que fa a la qualitat de predicció, es recomana utilitzar els models de profunditat 2, que són els més simples. A la *Taula 4.6.2* es troba un resum de les combinacions dels hiperparàmetres recomanades.

Taula 4.6.2. Combinacions recomanades d'hiperparàmetres (Dataset A).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
2	45	0.74	0.81	0.68
2	20	0.73	0.83	0.66

Comparant el model de Gradient Boosting de profunditat 2 i 45 arbres amb la Regressió Logística, s'observa que els valors de les mètriques són pràcticament idèntics. Amb les matrius de confusió (vegeu *Figura 4.6.1 i Figura 4.6.3*) s'observa que amb el model de Gradient Boosting el nombre de FN i VP es manté i, en canvi, augmenta el nombre de FP, que no convé. Així doncs, utilitzant el Dataset A el Gradient Boosting no aconsegueix millorar les prediccions respecte la Regressió Logística. Per tant, aplicant la Regressió Logística ja és suficient.

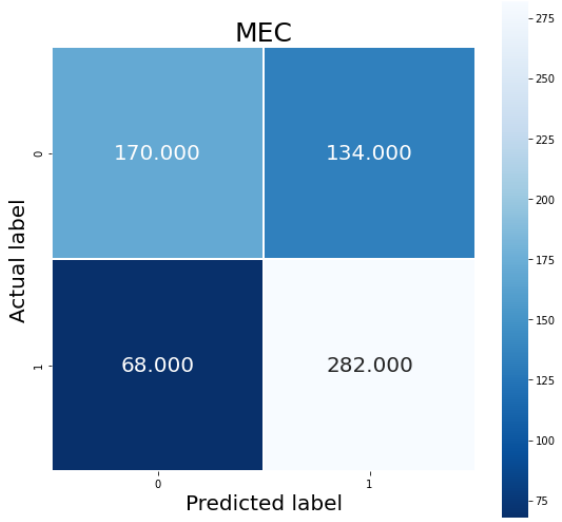


Figura 4.6.3. Matriu de confusió del model de Gradient Boosting amb profunditat 2 i 45 arbres (Dataset A).

Un cop vist el rendiment dels models utilitzant el Dataset A, es procedirà a fer l'estudi amb el Dataset B. En aquest cas, el model de Regressió Logística té un F1 Score igual a 0.73, una Precision de 0.70 i un Recall de 0.76. A la Figura 4.6.4 es pot observar la matriu de confusió corresponent. A més, aquest model no presenta *overffiting*, ja que els valors de F1 Score del grup d'entrenament i el d'avaluació difereixen per centèsimes.

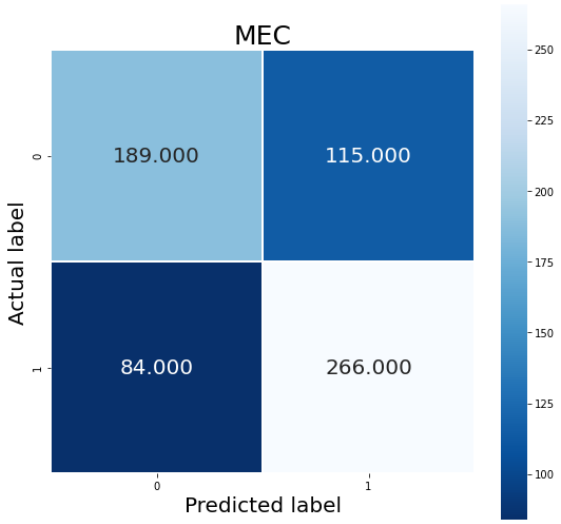
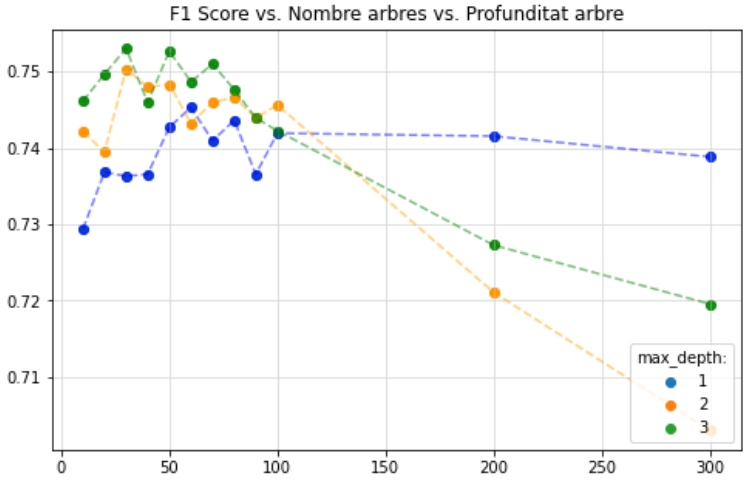


Figura 4.6.4. Matriu de confusió del model de Regressió Logística (Dataset B).

(a)



(b)

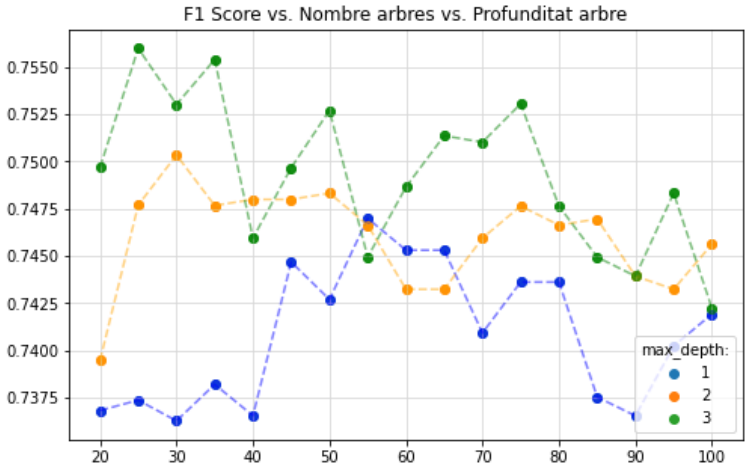


Figura 4.6.5. F1 Score en funció del nombre d'arbres i la profunditat (Dataset B). (a) Interval del nombre d'arbres entre 10 i 300. (b) Interval del nombre d'arbres ampliat entre 20 i 100.

A continuació s'analitzarà com es comporten els models de Gradient Boosting en funció dels hiperparàmetres. A la *Figura 4.6.5* es mostra com varia l'F1 Score en funció del nombre d'arbres i la profunditat. Els models que han obtingut els millors valors de les mètriques d'avaluació s'han recollit a la *Taula 4.6.3.*

Taula 4.6.3. Llistat de les millors combinacions d'hiperparàmetres i dels valors de les mètriques d'avaluació.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	45	0.75	0.73	0.80	0.70
	55	0.75	0.73	0.80	0.70
	65	0.75	0.73	0.79	0.70
2	30	0.75	0.74	0.80	0.70
	50	0.75	0.75	0.79	0.71
	75	0.75	0.76	0.79	0.71
3	25	0.76	0.76	0.81	0.71
	35	0.76	0.76	0.80	0.71
	75	0.75	0.80	0.80	0.71

Per comprovar l'existència d'*overfitting*, s'ha comparat el valor de F1 Score de *train* amb el de test i s'ha pogut observar que tcap dels models pateix *overfitting.* Seguidament, s'ha valorat quins són els millors models d'entre els seleccionats fent una anàlisi detallada a partir de les matrius de confusió i també comparant els valors de les mètriques d'avaluació. S'ha de tenir en compte que tots els models seleccionats tenen valors de les mètriques d'avaluació molt similars i, per tant, en aquest aspecte són pràcticament idèntics.

Si s'escull una profunditat igual a 1, es recomana utilitzar 45 arbres. Aquest model és el més simple i obté els mateixos resultats que el de profunditat 2 i 30 arbres. Amb la profunditat 3 es recomana utilitzar 25 arbres, ja que aquesta combinació aconsegueix els valors màxims de Precision i Recall. A més, és el model que obté el mínim nombre de FN i el màxim de VP. A la *Taula 4.6.4* es troba un resum de les combinacions dels hiperparàmetres recomanades.

Taula 4.6.4. Combinacions recomanades d'hiperparàmetres (Dataset B).

Profunditat	Nombre d'arbres	F1score	Recall	Precision
1	45	0.75	0.80	0.70
3	25	0.76	0.81	0.71

Comparant el model de Gradient Boosting de profunditat 3 i 25 arbres amb la Regressió Logística, s'observa que els valors de les mètriques són superiors, sobretot el Recall. A partir de les matrius de confusió (vegeu *Figura 4.6.4 i Figura 4.6.6*) es pot veure la millora aconseguida amb el model de Gradient Boosting. S'observa que s'han predit 20 verdaders positius (VP) més que amb la Regressió Logística i el nombre de falsos negatius (FN) ha disminuït amb la mateixa proporció. Així doncs, en aquest cas val la pena aplicar el Gradient Boosting, ja que millora la qualitat de les prediccions.

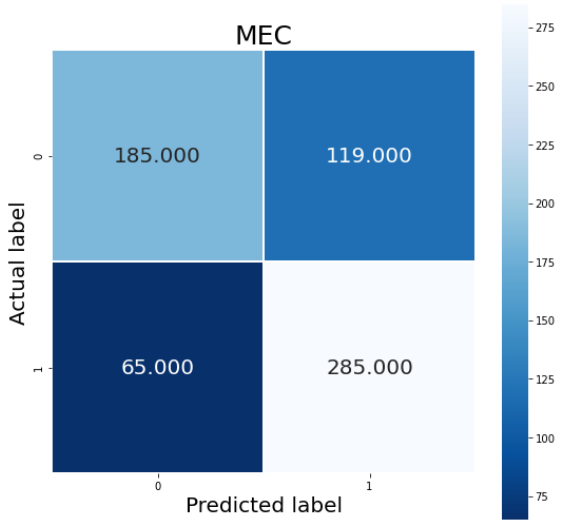


Figura 4.6.6. Matriu de confusió del model de Gradient Boosting amb profunditat 3 i 25 arbres (Dataset B).

Finalment, es realitzarà la comparació dels resultats obtinguts utilitzant el Dataset A i el Dataset B. Concretament es compararan els models amb millor rendiment predictiu. Utilitzant el Dataset B, els models han fet prediccions més encertades. Això es pot veure amb l'augment de VP predits i la reducció de prediccions errònies, sobretot de FP (vegeu *Figura 4.6.3 i Figura 4.6.6*). Encara que amb els dos grups de dades s'han aconseguit molts bons resultats, es pot dir que utilitzant el Dataset B s'aconsegueix millorar la qualitat de prediccions i, per tant, es recomana utilitzar-lo.

4.7. Conclusions globals

El conjunt d'estudis realitzats han permès arribar a la conclusió que les dades disponibles no són prou robustes. En la majoria d'estudis les dades no estaven equilibrades i això ha sigut un inconvenient per als models a l'hora de trobar la relació real que hi ha entre les variables d'entrada i les variables resposta. Tot i així, comparant els rendiments obtinguts amb la Regressió Logística i amb el Gradient Boosting s'ha pogut veure que, per a totes les assignatures, els models de Gradient Boosting han aconseguit un rendiment predictiu superior (vegeu *Taula 4.7.1*). Destaca l'augment del rendiment obtingut en les assignatures d'Electromagnetisme, Materials i Equacions Diferencials. Així doncs, es pot concloure que és molt millor utilitzar el Gradient Boosting, ja que en tots els casos supera els resultats de la Regressió Logística.

Pel que fa als grups de dades utilitzats, s'ha detectat que en la majoria dels casos s'han aconseguit millors prediccions amb el Dataset B que amb el Dataset A. Per això, es pot deduir que les variables predictores del Dataset B, que manquen al Dataset A, aporten informació rellevant i són significatives. En resum, en comparació amb el Dataset A, el Dataset B és més robust i es recomana utilitzar-lo.

Per altra banda, s'ha pogut confirmar que és millor utilitzar valors petits de l'hiperparàmetre 'profunditat'. De fet, els models que han utilitzat la profunditat més gran, amb valor igual a 3, són els que han obtingut pitjors rendiments predictius i també són els que han patit més *overfitting*. Per tant, es conclou que les profunditats d'1 i 2 són les més adequades.

També s'ha observat que en la majoria dels models estudiats hi ha una diferència d'aproximadament un 10% entre l'F1 Score de *train* i de test. Un motiu d'aquesta diferència podria ser que les dades d'entrenament i les d'avaluació estiguessin equilibrades amb una proporció diferent. S'assumeix aquesta variància i, per tant, no es considera que en aquests casos hi hagi *overfitting*.

A la *Taula 4.7.2* es troba el recull de tots els models de Gradient Boosting recomanats de cada assignatura, obtinguts a partir del Dataset B. Com es pot observar, els models que tenen una qualitat de predicció major són els de Mecànica, Electromagnetisme i Materials. En canvi, els models predictius de Mètodes Numèrics, d'Equacions Diferencials i d'Informàtica tenen un rendiment més baix. S'ha detectat que l'assignatura amb les dades més equilibrades (Mecànica) és la que ha obtingut el millor rendiment predictiu. I, al contrari, l'assignatura amb les dades més desequilibrades (Mètodes Numèrics) ha obtingut el pitjor rendiment predictiu. Amb això, es

conclou que els models fan prediccions de més qualitat quan les dades estan més equilibrades.

A més, l'assignatura de Mecànica també és un clar exemple de que, quan les dades estan equilibrades, els models d'*ensemble* necessiten combinar un nombre d'arbres significativament inferior i, per tant, s'aconsegueixen models molt més simples i efectius.

Taula 4.7.1. Augment del rendiment dels models utilitzant Gradient Boosting respecte la Regressió Logística (Dataset B).

Assignatura	RL	GB	Δ
Electromagnetisme	0.51	0.58	14 %
Mètodes Numèrics	0.1	0.17	7%
Materials	0.50	0.56	12%
Equacions Diferencials	0.21	0.26	24%
Informàtica	0.32	0.33	3%
Mecànica	0.73	0.76	4%

Taula 4.7.2. Models de Gradient Boosting recomanats de cada assignatura i els valors de les mètriques d'avaluació corresponents (Dataset B).

Assignatura	Profunditat	Nombre d'arbres	F1score	Recall	Precision
Electromagnetisme	1	200	0.56	0.50	0.62
	2	175	0.58	0.53	0.63
Mètodes Numèrics	2	125	0.17	0.11	0.38
Materials	2	70	0.56	0.50	0.62
Equacions Diferencials	1	125	0.25	0.15	0.63
	2	[250, 275]	0.26	0.17	0.51
Informàtica	1	[100, 200]	0.32	0.22	0.56
	2	70	0.33	0.24	0.51
Mecànica	1	45	0.75	0.80	0.70
	3	25	0.76	0.81	0.71

5. Estudis addicionals

Durant el desenvolupament del treball i a mesura que s'anaven obtenint els diferents resultats dels estudis han anat sorgint una sèrie de qüestions que mereixen un estudi més detallat. A continuació es duran a terme uns estudis addicionals centrats en l'*overfitting* i la qualitat de predicció dels models.

5.1. Correcció de l'overfitting

Com s'ha pogut observar al llarg dels diversos estudis realitzats, hi ha hagut una sèrie de models que a primera vista semblaven bons predictors però quan s'han contemplat més en detall s'ha vist que patien *overfitting* i, per tant, no són fiables. El Gradient Boosting sol ser bastant resistent a aquest fenomen i un nombre elevat d'arbres acostuma a donar bons resultats. Però, s'ha de tenir en compte que, degut a les limitacions d'equipament esmentades anteriorment, els models recomanats s'han obtingut ajustant només dos dels hiperparàmetres del Gradient Boosting: la profunditat (*max_depth*) i el nombre d'arbres (*n_estimators*). Hi ha altres hiperparàmetres que poden evitar l'*overfitting* i no s'han ajustat, com ara el *learning_rate* (definit a l'apartat 3.2).

L'objectiu d'aquest apartat és estudiar l'efecte del *learning_rate* sobre l'*overfitting* i comprovar si es pot reduir. Per això, es provaran diferents valors del *learning_rate* i s'analitzarà el seu impacte. S'estudiaran els models més afectats per l'*overfitting*, que són els de profunditat 3 de les assignatures de Mètodes Numèrics, Equacions Diferencials i Informàtica, tant amb el Dataset A com amb el Dataset B. Aquests models tenen un *learning_rate* de 0.1 i es compararan amb els models obtinguts amb els següents valors de *learning_rate*: 0.001, 0.01 i 0.2. Els valors de 0.001 i 0.01 s'han escollit perquè, com s'explica a l'apartat 3.2, el més habitual és utilitzar valors petits d'aquest hiperparàmetre. A més, per no descartar la possibilitat d'obtenir bons resultats amb valors superiors, s'ha volgut contemplar el valor de 0.2.

A la *Taula 5.1.1* es troben els valors de F1 Score obtinguts en l'entrenament (*train*) i en l'avaluació (test) de cada model. Els models amb el *learning_rate* igual a 0.001 no apareixen a la taula perquè en tots els casos s'han obtingut valors de F1 Score iguals a zero, fet que indica que amb aquest valor tan petit de *learning_rate* els models no aconsegueixen fer cap predicció encertada dels positius ($\mathrm{VP} \approx 0$).

En canvi, amb un *learning_rate* igual a 0.01 i utilitzant el Dataset B, s'ha observat que els models de les assignatures d'Equacions Diferencials i Informàtica han aconseguit reduir l'*overfitting.* Aquests models són fiables però tenen baix rendiment de predicció i, per això, no es recomanen.

Utilitzant el Dataset A, els models d'aquestes dues assignatures no tenen *overfitting* però han obtingut valors de F1 Score gairebé nuls i, per tant, no són capaços de predir cap positiu correctament ($\mathrm{VP} \approx 0$).

Taula 5.1.1. Valors de F1 Score de test i train en funció del learning_rate.

			learning_rate					
			0.01		0.1		0.2	
Assignatura	**Dataset**	**Nombre d'arbres**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**
Mètodes Numèrics	A	275	0.00	0.00	0.11	0.75	0.12	0.95
		300	0.00	0.01	0.10	0.77	0.12	0.98
		350	0.00	0.01	0.11	0.83	0.13	0.99
	B	275	0.05	0.35	0.25	0.84	0.18	1.00
		300	0.07	0.35	0.24	0.87	0.19	1.00
Equacions Diferencials	A	250	0.02	0.01	0.17	0.69	0.21	0.93
		300	0.03	0.04	0.17	0.75	0.22	0.96
		350	0.05	0.06	0.22	0.81	0.24	0.99
	B	225	0.12	0.20	0.27	0.80	0.28	0.94
		275	0.15	0.27	0.27	0.85	0.27	0.98
		300	0.20	0.31	0.28	0.87	0.28	0.98
Informàtica	A	200	0.01	0.03	0.25	0.68	0.33	0.86
		250	0.08	0.10	0.26	0.75	0.33	0.90
	B	250	0.30	0.43	0.36	0.83	0.44	0.96
		275	0.30	0.44	0.34	0.86	0.43	0.97

En el cas dels models amb *learning_rate* igual a 0.2, es pot observar que tots els valors de F1 Score d'entrenament (*train*) són pràcticament igual a 1. Això vol dir que en l'etapa d'entrenament el model aconsegueix predir gairebé tots el negatius i positius correctament ($\mathrm{FP} \approx 0$ i $\mathrm{FN} \approx 0$). En canvi, els valors de F1 Score de l'avaluació (test) són molt baixos. Per tant es pot dir que, respecte als models amb *learning_rate* igual a 0.1, encara ha augmentat més l'*overfitting*. Per tant, no es recomana utilitzar el valor de 0.2. Per acabar de descartar la possibilitat d'utilitzar valors de *learning_rate* superiors a 0.1, s'ha comprovat com serien els models amb un valor de

0.5 i s'ha trobat que totes els F1 Score d'entrenament (*train*) són exactament igual a 1 i, com en el cas anterior, hi ha molt d'*overfitting*. Ara sí, s'ha comprovat que amb valors de *learning_rate* superiors a 0.1 no s'aconsegueix reduir l'*overfitting*.

En resum, degut al desequilibri de les dades disponibles és difícil aconseguir que no hi hagi *overfitting*. S'observa que els models de l'assignatura de Mètodes Numèrics són amb diferència els pitjors predictors. Independentment del valors de *learning_rate*, tots pateixen *overfitting* o presenten un F1 Score igual a zero i no s'ha aconseguit cap millora al respecte. En canvi, si que s'ha aconseguit reduir l'*overfitting* d'alguns dels models corresponents a les assignatures d'Equacions Diferencials i Informàtica.

5.2. Variables significatives

Un cop vist que el Dataset B és el més robust i permet fer millors prediccions, seria interessant esbrinar quines són en concret les variables predictives que aporten més informació al model. En aquest apartat s'ha decidit fer una prova amb la variable 'Nota Accés', que correspon a la nota de la selectivitat (PAU) i el Batxillerat. Es pretén eliminar aquesta variable del Dataset B per veure com varia la qualitat de predicció dels models. Aquest estudi es farà per a totes les assignatures excepte Mètodes Numèrics, ja que es considera que les dades d'aquesta assignatura no són de qualitat i, com s'ha vist no permeten fer prediccions encertades. De manera que no s'arribaria a cap conclusió consistent.

Taula 5.2.1. Valors de les mètriques d'avaluació dels models d'Electromagnetisme.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	200	0.53	0.57	0.47	0.61
	250	0.53	0.58	0.48	0.61
	150	0.53	0.56	0.47	0.61
2	300	0.55	0.72	0.50	0.61
	275	0.54	0.72	0.49	0.60
	150	0.54	0.65	0.49	0.61
3	150	0.55	0.79	0.49	0.61
	250	0.54	0.86	0.50	0.60
	275	0.54	0.87	0.50	0.59

Els millors resultats obtinguts per a l'assignatura d'Electromagnetisme, utilitzant la versió modificada del Dataset B, es troben a la *Taula 5.2.1*. S'observa que, en comparació amb els models obtinguts amb el Dataset B complet (vegeu *Taula 4.1.3*), els models de profunditat 2 i 3 han necessitat un nombre més elevat d'arbres però no han obtingut millors prediccions i pateixen un *overfitting* més pronunciat. El model de profunditat 2 i 150 arbres és l'únic que no pateix *overfitting* i, a més, té un nombre d'arbres molt baix, però també ha obtingut pitjors valors de les mètriques d'avaluació que si s'utilitza el Dataset B complet. Pel que fa als models de profunditat 1, s'observa que tenen un nombre d'arbres similars però han obtingut uns valors de de les mètriques d'avaluació una mica inferiors. En resum, es pot dir que la variable 'Nota Accés' si que és significativa perquè quan s'utilitza s'obtenen millors prediccions.

Seguidament, s'ha fet l'estudi per a l'assignatura de Materials. A la *Taula 5.2.2* s'han recollit els models amb major rendiment predictiu. S'observa que, respecte als resultats obtinguts amb el Dataset B complet (vegeu *Taula 4.3.3*), els models amb profunditat 1 han necessitat menys arbres i tenen pràcticament el mateix rendiment predictiu, només tenen el Recall una mica per sota. Així doncs, en aquest cas, s'han obtingut models més simples al eliminar la variable 'Nota Accés' i s'ha mantingut la qualitat de les prediccions. En canvi, els models amb profunditat 2 han necessitat molts més arbres per aconseguir els mateixos resultats i, per tant, es dedueix que la informació que proporciona la variable 'Nota Accés' permet utilitzar menys arbres. Els models amb profunditat 3 no han variat pel que fa al rendiment de prediccions i pateixen *overfitting* de la mateixa manera que utilitzant el Dataset B complet.

Taula 5.2.2. Valors de les mètriques d'avaluació dels models de Materials.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	175	0.51	0.51	0.44	0.60
	200	0.51	0.52	0.44	0.60
2	175	0.53	0.62	0.48	0.59
	200	0.53	0.64	0.48	0.58
	225	0.53	0.66	0.48	0.58
3	175	0.54	0.79	0.49	0.60
	250	0.54	0.84	0.51	0.57

A continuació es farà l'estudi per a l'assignatura d'Equacions Diferencials. A la *Taula 5.2.3* es poden observar les principals característiques dels millors models obtinguts a partir del Dataset B modificat. Utilitzant una profunditat igual a 1 i el mateix nombre d'arbres, s'han obtingut millors prediccions que quan s'ha utilitzat el Dataset B complet (vegeu *Taula 4.4.3*). Han augmentat els valors de totes les mètriques d'avaluació. En aquest cas, es pot dir que la variable 'Nota Accés' dificulta que el model trobi la relació real entre les variables d'entrada i sortida. Pel que fa als models de profunditat 2 i 3, tots pateixen *overfitting* de la mateixa manera que quan s'ha utilitzat el Dataset B complet. En conclusió, per a l'assignatura es pot dir que és millor no incloure la variable 'Nota Accés' al grup de dades.

Taula 5.2.3. Valors de les mètriques d'avaluació dels models d'Equacions Diferencials.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	175	0.27	0.26	0.17	0.71
	150	0.26	0.25	0.16	0.70
	125	0.25	0.24	0.15	0.78
2	275	0.30	0.57	0.22	0.50
	250	0.28	0.55	0.20	0.49
	225	0.28	0.54	0.20	0.48
3	225	0.33	0.80	0.25	0.49
	250	0.32	0.83	0.24	0.48

Taula 5.2.4. Valors de les mètriques d'avaluació dels models d'Informàtica.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	225	0.34	0.43	0.22	0.53
	200	0.33	0.44	0.24	0.53
2	125	0.34	0.55	0.25	0.52
	150	0.34	0.55	0.25	0.53
3	250	0.37	0.81	0.30	0.50
	275	0.36	0.83	0.29	0.48

Els models amb els millors resultats obtinguts per a l'assignatura d'Informàtica es troben a la *Taula 5.2.4*. Els resultats dels models amb profunditat 1 són els mateixos utilitzant el Dataset B amb i sense la variable 'Nota Accés' (vegeu *Taula 4.5.3*), per tant aquesta no és gaire significativa en aquest cas. Els models de profunditat 2 i 3 han necessitat més arbres i no han aconseguit millors resultats. A més, tenen molt *overfitting.* Així doncs, es pot concloure que per fer prediccions sobre l'assignatura d'Informàtica no és necessari utilitzar les dades relatives a la nota d'Accés.

Taula 5.2.5. Valors de les mètriques d'avaluació dels models de Mecànica.

Profunditat	Nombre d'arbres	F1score (test)	F1score (train)	Recall	Precision
1	80	0.75	0.73	0.80	0.71
	75	0.75	0.73	0.80	0.70
	70	0.75	0.72	0.80	0.70
2	30	0.74	0.74	0.79	0.70
	95	0.74	0.76	0.78	0.71
3	55	0.75	0.78	0.80	0.70
	60	0.75	0.79	0.80	0.70

Finalment, s'estudiaran el models predictius de l'assignatura de Mecànica al utilitzar el Dataset B modificat. A la *Taula 5.2.5* s'han recollit els models amb major rendiment predictiu. Comparant amb la *Taula 4.6.3*, es pot observar que tots els models, independentment de la profunditat utilitzada, han necessitat més arbres per aconseguir els mateixos resultats. En definitiva, a l'eliminar la variable 'Nota Accés' ha costat més trobar la relació entre les variables d'entrada i sortida i per això s'han necessitat models predictius més complexos. Per tant, és millor incloure la variable 'Nota Accés' al Dataset B quan es vulguin fer prediccions sobre l'assignatura de Mecànica.

Havent vist com varien les prediccions de cada assignatura a l'eliminar la variable 'Nota Accés' del Dataset B, es pot dir que en la majoria de casos augmenta la complexitat dels models i majoritàriament empitjora la qualitat de les prediccions. A més, també augmenten els casos d'*overfitting*. En resum, es recomana utilitzar aquesta variable per tal d'aconseguir millors prediccions, excepte en els casos concrets que s'han comentat prèviament.

6. Planificació

En aquest apartat es mostra la planificació que s'ha seguit durant el desenvolupament del projecte, des de l'inici fins a la finalització. El projecte s'ha realitzat en cinc etapes diferents i cada una engloba un conjunt d'activitats en concret. Al Diagrama de Gantt de la *Taula 6.1* es pot observar la planificació completa del projecte, amb les dates d'inici i fi de cada etapa i activitat.

Taula 6.1. Diagrama de Gantt del projecte.

		Setembre-21				Octubre-21				Novembre-21				Desembre-21				Gener-22				Febrer-22			
INICI DEL PROJECTE	Comprensió del problema i definició d'objectius																								
	Recerca sobre Data Science																								
	Instal·lació de les eines de programació																								
	Repàs de Python (llibreria Pandas)																								
COMPRENSIÓ I PREPARACIÓ DE DADES	Comprensió de dades																								
	Neteja i filtratge de dades																								
	Transformació de dades																								
MODELATGE I AVALUACIÓ	Recerca de mètriques i mètodes d'avaluació																								
	Familiarització amb Gradient Boosting																								
	Construcció i selecció dels models òptims																								
	Anàlisi dels resultats																								
	Estudis addicionals																								
CREACIÓ DE LA MEMÒRIA	Redacció de la memòria																								
	Redacció de conclusions																								
PRESENTACIÓ	Preparació de la presentació																								
	Presentació del projecte																								

7. Estudi econòmic

A continuació es farà l'estudi econòmic a partir d'un càlcul aproximat dels costos associats a la realització del projecte. L'estudi s'ha dividit en dos blocs de costos: el cost de treball i el cost d'equips.

El cost de treball correspon a les hores que l'autor del treball ha dedicat a realitzar totes les tasques necessàries. Per tal de quantificar aquestes hores s'ha buscat l'equivalència entre els crèdits ECTS del Treball de Fi de Grau i les hores de treball que impliquen. Així doncs, sabent que 1 crèdit ECTS equival a 25 hores de feina i que el Treball de Fi de Grau és de 12 crèdits ECTS, s'obté que el nombre d'hores dedicades a aquest projecte són 300 hores. S'ha fitxat un sou de 8 € / h, ja que segons la normativa de la UPC és el sou mínim per a un estudiant d'aquesta institució. Amb tot això, ja es pot fer el càlcul del cost de treball, que és de 2.400 €.

Per calcular el cost d'equips es tindran en compte les despeses associades a l'ús de l'ordinador. S'ha negligit el cost de la resta de material utilitzat, com ara fulls i bolígrafs, ja que és gairebé nul. Així doncs, considerant que l'ordinador està valorat en 1.000 € i té una esperança de vida de 6 anys, s'ha calculat el cost d'amortització associat al seu ús durant les 22 setmanes dedicades al projecte, que és de 71 €.

Un cop calculats els diferents costos, s'obté que el desenvolupament del projecte suposa un cost total de 2.471 €.

8. Impacte ambiental

Aquest treball ha tingut un impacte ambiental mínim. Tota la feina s'ha desenvolupat amb l'ordinador i no s'han generat gens de residus, ja que no ha sigut necessari l'ús de papers. Però s'ha de tenir en compte el consum energètic que han suposat l'ús de l'ordinador i la connexió a internet proporcionada per la xarxa sense fil. Tot i així, aquesta despesa energètica es considera menyspreable i, per tant, es pot dir que aquest projecte és sostenible i s'ha realitzat respectant el medi ambient.

Conclusions

Havent finalitzat el desenvolupament del projecte, es pot dir que s'han complert tots els objectius marcats a l'inici d'aquest. Tot seguint la metodologia CRISP-DM, s'ha aconseguit crear models de Gradient Boosting en base a les dades disponibles reorganitzades, fent les operacions de neteja i transformació pertinents. S'ha pogut analitzar el rendiment de les prediccions obtingudes amb el Gradient Boosting i en la majoria de casos s'ha observat que aquest algorisme aconsegueix fer millors prediccions dels suspesos que la Regressió Logística. Tot i així, el rendiment predictiu dels models de Gradient Boosting ha sigut més baix del que s'esperava i s'ha arribat a la conclusió de que això és degut al fet que les dades no estan equilibrades. Per això, es dedueix que, si s'equilibressin les dades disponibles, els models de Gradient Boosting millorarien significativament els resultats.

Per a cada assignatura del tercer quadrimestre, s'han recomanat els millors models obtinguts, que fan les prediccions més encertades sobre els estudiants potencials a suspendre. Tot i així, no queda clar si seria recomanable posar-los en funcionament, ja que pot ser que no tinguin un rendiment predictiu prou elevat. L'única assignatura que podria aplicar els models recomanats és Mecànica, ja que les seves prediccions són molt fiables. El professorat de l'ETSEIB que ensenya l'assignatura en qüestió podria utilitzar aquests models per tal d'estimar quins estudiants necessitaran reforç acadèmic i, així, intentar prevenir que suspenguin.

Per altra banda, hi ha hagut una sèrie de tasques que no s'han pogut dur a terme degut a l'extensió del treball. En un futur, si es continués l'estudi, seria interessant equilibrar les dades, per exemple, creant mostres 'sintètiques' fent Oversampling de la classe minoritària [16]. És a dir, augmentar la quantitat de dades corresponents a la classe 'suspesos'. Una altra opció seria aplicar l'Undersampling, que elimina mostres de la classe majoritària (classe 'aprovats') de manera aleatòria. Però aquest mètode és menys recomanable. Un cop equilibrades les dades es podria tornar a aplicar el Gradient Boosting i veure si els resultats milloren. Per aconseguir un augment en la qualitat de les prediccions es podria fer una recerca més exhaustiva dels hiperparàmetres i trobar-ne els valors òptims. A més, seria interessant acabar de veure quines són les variables més significatives del grup de dades utilitzat, eliminar les que no ho són i pensar en noves variables que es podrien afegir.

Una altra opció a considerar seria aplicar altres models predictius, com ara k-Nearest Neighbors (k-NN) i Support Vector Machine (SVM), per veure si treballen millor amb les dades disponibles i, per tant, augmenta la qualitat de les prediccions. També es podria provar amb la variant del Gradient Boosting anomenada XGBoost.

A nivell personal, es pot dir que aquest projecte ha sigut molt enriquidor i molt complet. Amb aquest cas pràctic s'ha pogut veure el potencial de la ciència de dades i la informació valuosa que es pot arribar a extreure d'una gran quantitat de dades originalment desordenades i 'brutes'. A més, s'han pogut aplicar els coneixements de programació adquirits durant el grau i ampliar-los. S'ha descobert el gran ventall de possibilitats que ofereix el llenguatge Python i l'entorn de programació Spyder, com ara generar gràfics de tota mena i aplicar eines estadístiques. També s'ha guanyat domini en l'ús de la llibreria Pandas, amb la qual s'han manipulat les dades, i s'ha après a generar models predictius amb la llibreria Sklearn.

Agraïments

Primer de tot, vull agrair al Lluís Talavera per proposar aquest projecte. És una temàtica molt interessant i m'ha captivat molt. També vull donar les gràcies per la seva implicació des del primer moment. Tots els seus consells i correccions m'han sigut molt útils i m'han ajudat a desenvolupar el projecte amb èxit. Gràcies al seu guiatge he pogut treure el màxim profit d'aquest estudi.

També vull agrair a la meva família per tot el suport que m'ha donat durant tot el procés, per ser atents quan els explicava els diferents reptes que anava afrontant i preocupar-se de si portava bé el treball.

Finalment vull donar les gràcies al Bernat, per fer-me costat i animar-me quan més ho necessitava. Gràcies per creure en mi.

Bibliografia

Referències bibliogràfiques

[1] MUÑOZ, R. *El origen y evolución de la Ciencia de Datos (Data Science)*, Fundación iS+D. 2 de juliol del 2021. [https://isdfundacion.org/2021/07/02/el-origen-y-evolucion-de-la-ciencia-de-datos-data-science, 21 de setembre de 2021].

[2] Bigdata-analytics. *¿Qué es Data Science?* [https://bigdata-analytics.es/data-science, 21 de setembre del 2021].

[3] SINGULAR. *CRISP-DM: La metodología para poner orden en los proyectos*, 2021. [https://www.sngular.com/es/data-science-crisp-dm-metodologia, 25 de setembre del 2021].

[4] Iberdrola. *Qué es el 'Machine Learning'*. [https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico, 26 de setembre del 2021].

[5] RECUERO DE LOS SANTOS, P. *Tipos de aprendizaje en Machine Learning: supervisado y no supervisado*, Telefónica Tech. 2 de desembre del 2021. [https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje, 26 del setembre de 2021].

[6] AMAT, J. *Gradient Boosting con Python*, Ciencia de datos. Octubre del 2020. [https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html#Métodos-de-ensemble, 3 d'octubre del 2021].

[7] LOPEZ, R. *Boosting en Machine Learning con Python*, Matemáticas, análisis de datos y python. 10 de juny del 2017. [https://relopezbriega.github.io/blog/2017/06/10/boosting-en-machine-learning-con-python, 4 d'octubre del 2021].

[8] AMAT, J. *Árboles de decisión con Python: regresión y clasificación*, Ciencia de datos. Octubre del 2020. [https://www.cienciadedatos.net/documentos/py07_arboles_decision_python.html, 4 d'octubre del 2021].

[9] BROWNIEE, J. *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*, Machine Learning Mastery. 9 de setembre del 2016.

[https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning, 2 de novembre del 2021].

[10] LOPEZ, R. *Machine Learning con Python – Sobreajuste*, Matemáticas, análisis de datos y python. 29 de maig del 2016. [https://relopezbriega.github.io/blog/2016/05/29/machine-learning-con-python-sobreajuste, 2 de novembre del 2021].

[11] Education Wiki. *Algoritmo XGBoost*. [https://es.education-wiki.com/4794400-xgboost-algorithm, 6 de novembre del 2021].

[12] SINGH, N. *Métricas De Evaluación De Modelos En El Aprendizaje Automático*, Datasource. 2 de setembre del 2020. [https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico, 4 de novembre del 2021].

[13] ALLIBHAI, E. *Hold-out vs. Cross-validation in Machine Learning*, 3 de octubre del 2018. [https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f, 10 de novembre del 2021].

[14] RODRÍGUEZ, D. *La regresión logística*, Analytics Lane. 23 de juliol del 2018. [https://www.analyticslane.com/2018/07/23/la-regresion-logistica, 10 de novembre del 2021].

[15] Scikit-learn. *sklearn.ensemble.GradientBoostingClassifier*. [https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html#sklearn.ensemble.GradientBoostingClassifier, 10 de novembre del 2021].

[16] BROWNIEE, J. *Random Oversampling and Undersampling for Imbalanced Classification*, 15 de gener del 2020. [https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification, 11 de gener del 2022].

Bibliografia complementària

FERRERO, R. *Qué son los árboles de decisión y para qué sirven*, Maxima Formacion. Maig del 2020. [https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven, 4 d'octubre del 2021].

ICHI. *Que es XGBoost y como optimizarlo*. [https://ichi.pro/es/que-es-xgboost-y-como-optimizarlo-232831486673332, 6 de novembre del 2021].

KUMAR, A. *Hold-out Method for Training Machine Learning Models*, Data Analytics. 24 de novembre del 2021. [https://vitalflux.com/hold-out-method-for-training-machine-learning-model, 10 de novembre del 2021].

ICHI. *¿Qué es el Gradient Boosting? ¿En qué se diferencia de Ada Boost?* [https://ichi.pro/es/que-es-el-gradient-boosting-en-que-se-diferencia-de-ada-boost-45556808207053, 2 de novembre del 2021].