

Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Estudi comparatiu de tècniques de mostreig aplicades a la predicció de resultats acadèmics

MEMÒRIA

Autor: María Alcalde Olivares
Director: Luis José Talavera Méndez
Convocatòria: Juny 2021



Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona



Resum

El present document tracta sobre l'estudi comparatiu del rendiment de diverses tècniques de mineria de dades per a predir si un estudiant aprovarà o suspendrà una determinada assignatura del tercer quadrimestre del Grau en Enginyeria en Tecnologies Industrials de l'ETSEIB.

La mineria de dades és un procés que permet obtenir informació molt valuosa a partir de grans bases de dades mitjançant la seva exploració en busca de patrons de conducta que siguin útils per poder predir situacions futures, obtenir beneficis o reduir costos i riscos.

Al llarg de tot el projecte s'ha seguit la metodologia CRISP-DM que comprèn totes les etapes necessàries per dur a terme un estudi de mineria de dades de forma adequada i satisfactòria. S'ha fet ús de diverses eines de programació que treballen amb el llenguatge *Python*, com *Jupyter Notebooks* i les llibreries *Pandas* i *Scikit-learn*.

Els algoritmes de predicció emprats han estat la regressió logística i els arbres de decisió. S'ha aplicat diverses tècniques de mostreig com són *RandomOverSampler*, *SMOTE* i *BorderlineSMOTE* amb l'objectiu d'equilibrar la distribució de dades entre les classes aprovat i suspens.

Les principals conclusions que s'han obtingut de l'estudi són que la distribució no equilibrada de la variable resposta impacta negativament en el rendiment dels models de predicció i que, per tant, l'aplicació de diferents tècniques de mostreig permet obtenir un augment significatiu en la predicció de la classe minoritària quan les dades estan desequilibrades.

Sumari

SUMARI	5
1. GLOSSARI	7
2. INTRODUCCIÓ	9
2.1. Objectius del projecte	11
2.2. Eines utilitzades.....	12
2.2.1. <i>Python</i>	12
2.2.2. <i>Jupyter Notebook</i>	13
2.2.3. <i>Pandas</i>	15
2.2.4. <i>Scikit-learn</i>	15
2.3. Abast del projecte	15
3. COMPRENSIÓ DE LES DADES	17
3.1. Dades de la preinscripció	17
3.2. Dades de la fase inicial i no inicial.....	17
4. PREPARACIÓ DE LES DADES	19
4.1. Selecció i neteja de dades	19
4.2. Transformació de les dades	20
5. MODELATGE I VALIDACIÓ	25
5.1. Models predictius.....	25
5.1.1. Regressió logística.....	26
5.1.2. Arbre de decisió	27
5.1.3. K-Nearest neighbors.....	29
5.1.4. Support Vector Machines (SVM)	29
5.2. Mètriques d'avaluació.....	30
5.2.1. <i>Confusion Matrix</i>	30
5.2.2. <i>Accuracy</i>	31
5.2.3. <i>Precision</i>	32
5.2.4. <i>Recall</i>	32
5.2.5. <i>F1-score</i>	32
5.3. Mètodes de validació.....	32
5.3.1. Hold out.....	33
5.3.2. K-Fold Cross Validation.....	34
6. TÈCNiques DE MOSTREIG PEL DESEQUILIBRI DE DADES	35
6.1. Oversampling	37

6.1.1.	Random Oversampling	37
6.1.2.	Synthetic Minority Oversampling Technique (SMOTE)	37
6.1.3.	Borderline-SMOTE	39
6.2.	Undersampling	39
6.2.1.	Random Undersampling	40
6.2.2.	Near Miss Undersampling	40
6.2.3.	Tomek Links for Undersampling	41
7.	ANÀLISI DE RESULTATS	43
7.1.	Regressió logística	43
7.1.1.	Electromagnetisme	44
7.1.2.	Mètodes numèrics	48
7.1.3.	Materials	52
7.1.4.	Equacions diferencials	56
7.1.5.	Informàtica	60
7.1.6.	Mecànica	64
7.1.7.	Resum resultats regressió logística	68
7.2.	Arbres de decisió	72
7.2.1.	Electromagnetisme	72
7.2.2.	Mètodes numèrics	76
7.2.3.	Materials	80
7.2.4.	Equacions diferencials	84
7.2.5.	Informàtica	88
7.2.6.	Mecànica	92
7.2.7.	Resum resultats arbres de decisió	96
7.3.	Comparació entre els models predictius regressió logística i arbres de decisió	100
7.4.	Aplicació pràctica del model	101
8.	IMPACTE AMBIENTAL	103
9.	PLANIFICACIÓ	105
10.	PRESSUPOST	107
11.	CONCLUSIONS	109
12.	TREBALLS FUTURS	111
BIBLIOGRAFIA		113
	Referències bibliogràfiques	113
	Bibliografia complementària	114

1. Glossari

Data mining: conjunt de tècniques i tecnologies que permeten explorar grans bases de dades, de manera automàtica o semiautomàtica, amb l'objectiu de trobar patrons repetitius que expliquin el comportament d'aquestes dades.

DataFrame: estructura de dades en forma de taula amb la que treballa la llibreria *Pandas*.

Dataset: conjunt de dades.

Imbalanced-learn: paquet de *Python* que ofereix una sèrie de tècniques de remostreig que s'utilitzen habitualment en conjunts de dades que mostren un fort desequilibri entre classes.

Machine learning: disciplina del camp de la Intel·ligència Artificial que, a través d'algoritmes, dota als ordinadors de la capacitat d'identificar patrons en dades per fer prediccions.

NaN: De l'anglès: 'Not a Number'. Format que prenen els valors buits a *Pandas*.

Overfitting: es produeix quan el model d'estudi s'ajusta massa a les dades d'entrenament.

Oversampling: tècnica de mostreig que consisteix en duplicar exemples existents o bé sintetitzar-ne de nous a partir dels de la classe minoritària per equilibrar la distribució de dades.

Pandas: llibreria de *Python* especialitzada en el maneig i anàlisi d'estructures de dades.

Python: llenguatge de programació emprat al llarg del treball.

Scikit-learn: llibreria de *Python* que conté els algoritmes de classificació utilitzats per a la predicció dels resultats acadèmics.

Testing: fase de validació dels models construïts. El conjunt de dades que s'utilitza en aquesta etapa rep el mateix nom.

Training: fase de construcció dels models predictius. El conjunt de dades que s'utilitza en aquesta etapa rep el mateix nom.

Underfitting: es produeix quan al model li es impossible identificar o obtenir resultats correctes per no tenir suficients mostres d'entrenament o un entrenament molt pobre.

Undersampling: tècnica de mostreig que consisteix en eliminar exemples del conjunt de dades d'entrenament que pertanyen a la classe majoritària per tal d'equilibrar millor la distribució de la classe.

2. Introducció

Avui en dia l'emmagatzemament massiu de dades és clau en diferents àmbits com poden ser el món empresarial, la medicina, la banca i l'administració entre d'altres. El fet de disposar d'una gran base de dades aporta informació molt valuosa que permetrà obtenir el màxim rendiment de les competències, augmentar els beneficis o bé, entendre les necessitats del client.

Per poder obtenir la màxima informació de tot aquest conjunt de dades, que no es pot processar manualment, és necessària la mineria de dades o *data mining*. Consisteix a aplicar diferents tècniques o tecnologies, automàtiques o semiautomàtiques, que permeten explorar aquestes grans bases de dades. El seu principal objectiu és detectar anomalies, patrons de conducta i correlacions que permeten predir situacions futures, obtenir beneficis o reduir costos i riscos.

Per exemple, en el màrqueting, analitzant relacions entre paràmetres com l'edat dels clients, el gènere o els seus gustos, és possible esbrinar el seu comportament per dirigir campanyes personalitzades de fidelització o captació. D'altra banda, en l'àmbit de la medicina, la mineria de dades permet realitzar diagnòstics més precisos i prescriure tractaments més efectius. També dona la possibilitat de gestionar de forma més eficaç, eficient i econòmica els recursos sanitaris.

Aquesta branca de l'enginyeria computacional combina mètodes de diverses especialitats científiques com són l'estadística i diverses branques de la intel·ligència artificial com pot ser l'aprenentatge automàtic o *machine learning*.

Actualment, existeixen diversos mètodes per poder processar totes aquestes dades i dur a terme un projecte de mineria de dades de forma adequada i satisfactòria. Una de les metodologies més utilitzades [1] i, la que es seguirà durant aquest projecte és la CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [2]. Es tracta d'un procés cíclic que consta de sis fases. A la Figura 1 s'observa tot el cicle de la metodologia CRISP-DM on les fletxes indiquen les relacions més importants i freqüents entre les diferents fases. Tot i això, no es tracta d'un procés rígid, pot haver-hi avanç i retrocés entre fases.

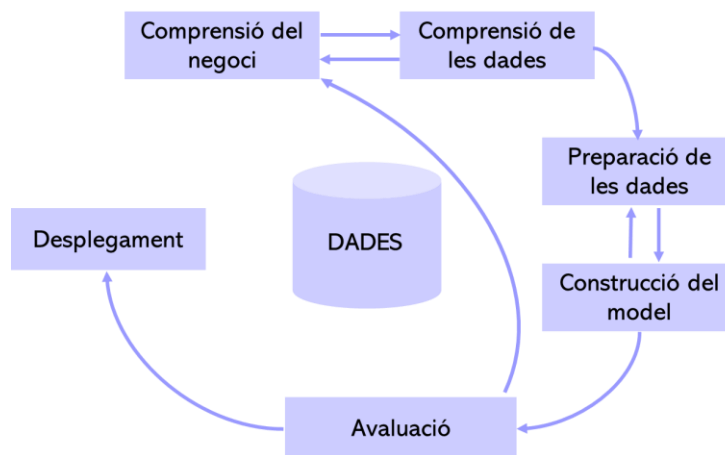


Figura 1: Etapes del cicle de la metodologia CRISP-DM

Comprensió del negoci

En aquesta fase inicial es determinen els objectius del projecte i s'analitza la situació inicial. És essencial per a un estudi de mineria de dades saber per a què serveix aquest estudi. A continuació, s'estableixen els objectius de mineria de dades i, finalment, es desenvolupa un pla de projecte.

Comprensió de les dades

Un cop determinats els objectius i el pla del projecte, es recopilen les dades que es tractaran més endavant. Aleshores, es tracta de familiaritzar-se amb les dades per tal de conèixer la informació que ens donen i poder verificar la seva qualitat.

Preparació de les dades

Durant aquesta etapa es duen a terme totes les activitats necessàries amb l'objectiu d'obtenir les dades finals amb les que es treballarà en l'etapa de modelatge. Algunes d'aquestes activitats són la selecció, neteja i transformació de dades.

Modelatge

L'etapa de modelatge consisteix en seleccionar i aplicar les tècniques de mineria de dades per construir un model adequat per a l'objectiu del projecte. Algunes d'aquestes tècniques són l'associació, la classificació, l'agrupació, les prediccions, els patrons seqüencials i les seqüències de temps similars. Algunes requereixen una forma específica de les dades, per tant, serà molt habitual tornar a la fase de preparació de dades.

Avaluació

En aquesta fase s'avaluen els resultats del model en el context dels objectius establerts durant la primera etapa. Això conduirà a la identificació d'altres necessitats que, sovint, ens faran tornar a fases anteriors del procés.

Desplegament

Aquesta última etapa correspon a la implementació del model final obtingut en qualsevol forma que assoleixi els objectius proposats. Es pot desenvolupar una aplicació, integrar els resultats en un sistema ja existent o bé, simplement generar un informe que ajudi a entendre millor el problema que es tracta i prendre les mesures corresponents. Tot i això, com ja s'ha comentat anteriorment, es tracta d'un procés cíclic, és a dir, aquesta fase pot no ser l'última ja que caldrà controlar l'estudi de mineria de dades per si hi ha canvis amb el temps i és necessari tornar a començar de nou.

Al llarg de la memòria es descriurà tot el procediment seguit per realitzar el projecte de mineria de dades utilitzant la metodologia CRISP-DM. El fet de seguir aquesta metodologia no només agilitza el procés d'obtenció del resultat final sinó que també ens assegura tant la qualitat de les dades utilitzades com la dels resultats obtinguts.

2.1. Objectius del projecte

En aquest projecte s'estudiarà el rendiment de diverses tècniques de mostreig per a predir si un estudiant aprovarà o suspendrà una determinada assignatura del tercer quadrimestre del Grau en Enginyeria en Tecnologies Industrials de l'ETSEIB.

L'objectiu principal del projecte és comparar el rendiment de diverses tècniques de mineria de dades en la predicció de resultats acadèmics. La distribució no equilibrada de la variable resposta impacta negativament en el rendiment d'aquests models de predicció. Per tant, s'estudiarà el funcionament d'aquestes tècniques i s'avaluarà fins a quin punt són eficients per a reduir aquest impacte. Es farà una comparativa entre les diferents tècniques analitzant el seu rendiment en els diferents models predictius en funció d'alguns dels paràmetres de les tècniques de mostreig i dels algoritmes de predicció. Un cop construïts els models és necessari validar-los. S'escollirà quin mètode s'aplicarà als diferents models. És important que tots els models es validin de la mateixa manera per poder obtenir l'eficàcia de cadascun i així comparar-los entre ells.

Tot aquest estudi es farà aplicant de forma rigorosa la metodologia CRISP-DM per al desenvolupament de projectes de mineria de dades. En qualsevol projecte és important seguir una metodologia determinada per poder arribar als millors resultats possibles.

En el cas que ens ocupa l'etapa de desplegament no es durà a terme ja que queda fora dels límits establerts pel projecte. La resta de fases es duran a terme per comprovar la viabilitat dels models amb els diferents mètodes que s'ha aplicat.

Durant l'estudi serà important determinar quins estudiants suspendran les diferents assignatures ja que això permetrà no només que els alumnes puguin preparar-se millor sinó que també els professors puguin enfocar la matèria d'una altra manera i així evitar els suspensos. Tot i això, no es pot deixar de banda la predicció dels aprovats ja que l'error en alguna de les dues variables provocarà una predicció falsa.

D'altra banda, es poden definir també com a objectius complementaris la familiarització amb l'entorn de treball i el coneixement de diverses llibreries de *Python* que seran utilitzades al llarg del projecte. Tot el procés de programació de codi es durà a terme mitjançant el software *Anaconda* que conté totes les eines necessàries per realitzar l'estudi. Per manipular el conjunt de dades s'utilitzarà la llibreria *Pandas* de *Python* la qual ens permetrà fer tot tipus d'operacions i transformacions. A més, la biblioteca *Scikit-learn*, també de *Python*, conté els algorismes de classificació que s'utilitzaran per a la predicció dels resultats acadèmics. Per últim, la programació del codi es farà mitjançant l'aplicació *Jupyter Notebooks* que permet implementar el codi de programació i documentar-ne cada pas. Serà important familiaritzar-se amb tot aquest entorn per poder desenvolupar el projecte de forma satisfactòria.

2.2. Eines utilitzades

2.2.1. *Python*

Python és un llenguatge de programació interpretat d'alt nivell que va ser creat per Guido van Rossum l'any 1991. La seva filosofia de disseny busca llegibilitat en el codi i la seva sintaxi permet als programadors expressar conceptes en menys línies de codi del que seria possible en altres llenguatges com per exemple C. També proveeix estructures per permetre programes més entenedors tant a petita com a gran escala. *Python* suporta diversos paradigmes de programació, presenta un sistema dinàmic i una gestió de la memòria automàtica i té una gran i exhaustiva biblioteca estàndard.

Tot i que *Python* es pot utilitzar per desenvolupar pràcticament qualsevol aplicació, és realment útil a l'hora de treballar en tecnologies com la Intel·ligència Artificial, l'aprenentatge automàtic i l'anàlisi de dades. A més, segons les enquestes del portal *KDnuggets*, és un dels llenguatges més populars al món.

El principal motiu pel qual s'ha escollit el llenguatge *Python* per al projecte és que, com ja s'ha mencionat anteriorment, és un dels llenguatges més utilitzats i, a més, incorpora llibreries útils a l'hora de desenvolupar projectes de mineria de dades com són *Pandas* i *Scikit-learn*. Un altre motiu és que ha estat el llenguatge que s'ha impartit en la docència de les assignatures de programació cursades al llarg del grau, Fonaments d'Informàtica i Informàtica.

2.2.2. Jupyter Notebook

Jupyter Notebook és un entorn informàtic interactiu basat en la web per crear documents de *Jupyter Notebook*, uns documents JSON que segueixen un esquema versionat i que contenen una llista ordenada de cel·les d'entrada/sortida que poden contenir codi, text, matemàtiques i gràfics, generalment acabats amb l'extensió ".ipynb". A més, permeten exportar tot el treball realitzat a fitxers PDF, HTML o ".py".

Existeixen altres opcions per editar el codi de programació, com per exemple *Spyder*, que és un IDE (Integrated Developed Environment), un entorn més adient per un entorn més ordenat en mòduls i funcions. No obstant això, s'ha escollit treballar amb *Jupyter Notebooks* perquè són més útils en l'exploració interactiva de dades i permeten documentar cada pas a més d'incloure gràfics i poder-se visualitzar en un navegador. D'aquesta manera, es poden compartir entre membres d'un equip per seguir els passos duts a terme a l'anàlisi. Que no siguin tan adequats a nivell d'organització de codi no serà tant important en aquest treball doncs l'objectiu no és desenvolupar una aplicació, sinó que és un projecte de caire experimental. A la Figura 2 es mostra com s'organitza el codi amb *Jupyter Notebooks*.

Preparació de les dades

In [1]: `import pandas as pd`

In [2]: `#Dades fase inicial
DadesFI = pd.read_excel('qfaseini19.xlsx')
#Dades fase no inicial
DadesFNI = pd.read_excel('qfasenoini19.xlsx')
#Dades preinscripció
DadesPers = pd.read_excel('dpersnombrespreins19esc.xlsx')`

Dades fase inicial

In [3]: `#Ordenem per expedient, curs i quadrimestre
DadesFI = DadesFI.sort_values(by=['CODI_EXPEDIENT', 'CURS', 'QUAD'])
#Només agafem titulació GETI
DadesFI = DadesFI[DadesFI['CODI_PROGRAMA']==752]
#Treiem convalidacions
DadesFI = DadesFI[DadesFI['GRUP_CLASSE']!='CONV']
#Treiem valor anòmal quadrimestre 0
DadesFI = DadesFI[DadesFI['QUAD']!=0]
#Conservem només els expedients amb assignatures de la fase inicial
FI = set(['240011', '240012', '240013', '240014', '240015', '240021', '240022', '240023', '240024', '240025'])
DadesFI['CODI_UPC_UD'] = DadesFI['CODI_UPC_UD'].apply(lambda x:str(x))
DadesFI = DadesFI[DadesFI['CODI_UPC_UD'].isin(FI)]
#Conservem només els que superen la fase inicial
num_expedient = set(DadesFNI['CODI_EXPEDIENT'])
DadesFI['SUPERA'] = DadesFI['CODI_EXPEDIENT'].isin(num_expedient)
DadesFI = DadesFI[DadesFI['SUPERA'] == True]
DadesFI`

Out[3]:

	CODI_PROGRAMA	CODI_EXPEDIENT	CODI_UPC_UD	CREDITS	CURS	QUAD	SUPERA	NOTA_PROF	NOTA_NUM_AVAL	NOTA_NUM_DEF
5332	752	226410	240011	6.0	2010	1	S	6.6	6.6	6.6
13113	752	226410	240014	6.0	2010	1	N	0.0	0.0	0.0
27548	752	226410	240015	6.0	2010	1	S	8.0	8.0	8.0
27641	752	226410	240012	6.0	2010	1	S	7.6	7.6	7.6
49309	752	226410	240013	6.0	2010	1	S	6.3	6.3	6.3
...
22172	752	365231	240023	6.0	2018	2	S	5.7	5.7	5.7
36721	752	365231	240014	6.0	2018	2	S	5.7	5.7	5.7
36722	752	365231	240025	7.5	2018	2	S	5.7	5.7	5.7
44048	752	365231	240013	6.0	2018	2	S	5.7	5.7	5.7
51234	752	365231	240021	6.0	2018	2	S	5.7	5.7	5.7

38668 rows × 12 columns

In [4]: `#Comprovem que tots els quadrimestres tinguin un nombre de matriculats raonable
TaulaCurs = pd.pivot_table(DadesFI, index='CODI_EXPEDIENT', columns='CURS', values='QUAD')
TaulaCurs`

Out[4]:

	CURS	2010	2011	2012	2013	2014	2015	2016	2017	2018
CODI_EXPEDIENT										
226410	1.545455	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
226431	1.500000	1.25	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
226455	1.500000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
226464	1.545455	1.00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
226467	1.545455	1.60	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...
349941	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0
350142	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0	NaN
363944	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0
365210	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0
365231	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	2.0

3174 rows × 9 columns

Figura 2: Entorn de treball Jupyter Notebook

2.2.3. *Pandas*

En Computació i Ciència de dades, *Pandas* és una llibreria de software per al llenguatge de programació *Python* que permet la manipulació i anàlisi de dades oferint estructures de dades i operacions per manipular taules numèriques i sèries temporals. Proporciona eines que permeten llegir i escriure dades en diversos formats com CSV, Microsoft Excel, bases SQL i format HDF, seleccionar i filtrar de manera senzilla taules de dades en funció de posició, valor o etiquetes, fusionar i unir dades o transformar dades aplicant funcions, entre d'altres.

En *Pandas* existeixen tres tipus d'estructura de dades en funció de les dimensions amb les que es vulgui treballar. La més comuna és el *DataFrame*, una estructura de dades amb dues dimensions en la qual es poden emmagatzemar dades de diferents tipus com caràcter, nombres enters o nombres reals, entre d'altres, en columnes.

La llibreria *Pandas* serà especialment útil en l'etapa de preparació de les dades ja que ens permetrà netejar, manipular i transformar les dades de partida fins aconseguir el conjunt definitiu per a l'etapa de modelatge.

2.2.4. *Scikit-learn*

Scikit-learn és una llibreria per aprenentatge automàtic de software lliure per el llenguatge de programació *Python*. Inclou diversos algorismes de classificació, regressió i anàlisi de grups. Un dels seus avantatges és que ofereix una gran varietat de mòduls i algorismes que faciliten l'aprenentatge i el treball dels científics de dades en les primeres fases del seu desenvolupament.

Aquesta biblioteca serà especialment útil en l'etapa de modelatge i validació ja que ens permetrà aplicar els diferents algorismes de predicció per a la creació dels models estadístics.

2.3. Abast del projecte

Com ja s'ha mencionat anteriorment, per desenvolupar l'estudi es seguirà la metodologia CRISP-DM. Com no es tracta d'una metodologia rígida, les seves fases han estat adaptades al projecte.

- **Comprensió del problema:** en qualsevol projecte és molt important conèixer a fons el problema que es vol tractar. Per tant, en aquesta fase es definiran els objectius i s'analitzarà la situació actual.
- **Comprensió de les dades:** l'objectiu d'aquesta fase serà familiaritzar-se amb el conjunt de dades amb el que es treballarà.

- **Preparació de les dades:** en aquesta fase es depuraran i transformaran les dades per tal d'aconseguir que aquestes tinguin el format necessari per l'etapa de modelatge. Aquest procediment es durà a terme mitjançant un codi de programació.
- **Modelatge i validació:** un cop obtingut el conjunt final de dades es procedirà a la construcció de diversos models aplicant diferents algoritmes de predicció i tècniques de mostreig. Finalment, es validaran els resultats.
- **Avaluació:** es tracta de l'última etapa d'aquest projecte. En aquesta fase es procedirà amb l'anàlisi dels resultats obtinguts.

Cal recordar que en aquest treball l'etapa de desplegament no es durà a terme ja que queda fora dels límits establerts pel projecte. La resta de fases es duran a terme per comprovar la viabilitat dels models amb els diferents mètodes que s'ha aplicat.

3. Comprensió de les dades

La comprensió de les dades és una fase fonamental de la mineria de dades, ja que ens permet familiaritzar-nos amb la informació i determinar com tractar-la en la següent fase.

En aquest cas, disposem de tres arxius *Excel*: “*dpersnomespreins19esc.xlsx*”, “*qfaseini19*” i “*qfasenoini19*”. Dins aquests arxius trobem les dades acadèmiques dels alumnes des de l’any 2010 fins el 2018.

3.1. Dades de la preinscripció

El primer arxiu conté les dades de preinscripció on cada fila és un alumne i cada columna conté la següent informació.

- CODI_EXPEDIENT: número d’expedient de l’alumne.
- SEXE: indica si és home (H) o dona (D).
- CP_FAMILIAR: codi postal del lloc de residència de l’alumne.
- ANY_ACCES: any en que l’alumne va accedir a la universitat.
- TIPUS_ACCES: tipus d’accés a la universitat (pren 1 com a únic valor).
- NOTA_ACCES: nota obtinguda a les proves d’accés a la universitat.
- CENTRE_SECUNDARIA: nom del centre on ha cursat els estudis de secundària.
- CP_CENTRE_SEC: codi postal del centre on ha cursat els estudis de secundària.

A la Taula 1 trobem una petita mostra de les dades d’aquest primer arxiu.

CODI_EXPEDIENT	SEXE	CP_FAMILIAR	ANY_ACCES	TIPUS_ACCES	NOTA_ACCES	CENTRE_SECUNDARIA	CP_CENTRE_SEC
355957	H	8820	2018	1	10,75	JAUME BARMES	8907
354396	D	8173	2018	1	12,216	SANT IGNASI- SARRIÀ	8017
353722	D	8980	2018	1	12,666	SANT PAU	8034
353747	H	8012	2018	1	10,65	LA FARGA	8195
353798	D	8006	2018	1	11,954	SAGRAT COR-SARRIÀ	8034

Taula 1: Mostra de les dades del fitxer “*dpersnomespreins19.xlsx*”

3.2. Dades de la fase inicial i no inicial

D’altra banda, dins els arxius “*qfaseini.xlsx*” i “*qfasenoini.xlsx*” trobem les dades acadèmiques referents a les qualificacions de la fase inicial i la fase no inicial respectivament. A la Taula 2 trobem una petita mostra dels dos arxius anteriors. Les dades es disposen en columnes amb la següent informació.

- CODI_PROGRAMA: codi identificador de la titulació que s'està cursant.
- CODI_EXPEDIENT: número d'expedient de l'alumne.
- CODI_UPC_UD: codi identificador de l'assignatura.
- CREDITS: nombre de crèdits ECTS que corresponen a l'assignatura.
- CURS: any en que es matricula l'assignatura.
- QUAD: quadrimestre en que es matricula l'assignatura (1 pel quadrimestre de tardor i 2 pel quadrimestre de primavera).
- SUPERA: indica si l'alumne aprova (S) o no aprova (N) l'assignatura.
- NOTA_PROF: nota final de l'assignatura proposada pel professor.
- NOTA_NUM_AVAL: nota final de l'assignatura per l'avaluació curricular.
- NOTA_NUM_DEF: nota final de l'assignatura definitiva després de l'avaluació curricular.
- GRUP_CLASSE: grup de classe on s'ha matriculat l'alumne (10, 20, 30, etc.). També pot prendre el valor CONV si es convalida l'assignatura.

CODI_PROGRAMA	CODI_EXPEDIENT	CODI_UPC_UD	CREDITS	CURS	QUAD
752	228884	240024	4,5	2011	1
752	226467	240024	4,5	2011	1
752	228476	240024	4,5	2011	1
752	229037	240024	4,5	2011	1
752	226717	240024	4,5	2011	1

SUPERA	NOTA_PROF	NOTA_NUM_AVAL	NOTA_NUM_DEF	GRUP_CLASSE
S	6,7	6,7	6,7	10
N	4,7	4,7	4,7	10
N	2,5	2,5	2,5	10
N	4,3	4,3	4,3	10
S	6,1	6,1	6,1	20

Taula 2: Mostra de les dades dels fitxers "qfaseini.xlsx" i "qfasenoini.xlsx"

4. Preparació de les dades

La fase de preparació de les dades té com objectiu obtenir les dades finals amb que es treballarà a l'etapa de modelatge. Per tant, un cop analitzades procedirem a seleccionar-les, netejar-les i transformar-les.

Inicialment, tenim totes les dades en fitxers d'*Excel* que haurem de convertir a *DataFrame* per poder treballar amb *Pandas*. L'objectiu és obtenir una única taula on cada fila correspongui a un alumne i les columnes mostrin les notes obtingudes de les diferents assignatures.

A continuació, es descriurà el procediment seguit per passar dels fitxers de dades inicials al conjunt definitiu.

4.1. Selecció i neteja de dades

El propòsit d'aquesta etapa és depurar per obtenir les dades necessàries per l'etapa de modelatge.

Per començar, només ens interessen els estudiants que estiguin cursant el grau d'enginyeria en tecnologies industrials. Aquesta informació la trobem a la columna *CODI_PROGRAMA* i la titulació GETI correspon al codi 752. Per tant, eliminarem tots els expedients que no es corresponen amb aquests estudis.

Existeixen alumnes que cursen la mateixa assignatura o semblant en un altre centre i decideixen convalidar-la. Aquests estudiants es poden detectar fàcilment perquè a la variable *GRUP_CLASSE* apareixen com a 'Conv'. Aquests casos no es poden comparar amb els que cursen la matèria a l'escola ja que la metodologia i l'avaluació són diferents. Per tant, procedirem a eliminar aquests casos de convalidació.

D'altra banda, observem que a la columna *QUAD* apareixen valors iguals a 0. Es tracta de valors anòmals ja que el quadrimestre de tardor es correspon amb el número 1 i el de primavera amb el número 2. En conseqüència, les dades corresponents a aquest quadrimestre quedaran fora perquè no ens proporcionen informació necessària pel nostre estudi.

Com l'objectiu del projecte és predir les qualificacions del tercer quadrimestre a partir de les notes de les assignatures de la fase inicial, seleccionarem només les matèries corresponents a aquests tres quadrimestres. D'aquesta manera, s'eliminaran totes les files que mostrin a la variable *CODI_UPC_UD* un codi d'assignatura que no necessitem.

Finalment, per acabar amb la depuració de dades, seleccionarem aquells estudiants amb la fase inicial superada. Per poder realitzar l'estudi és necessari que l'alumne hagi cursat alguna assignatura del tercer quadrimestre. Per tant, comparant els números d'expedient dels dos fitxers de dades, eliminarem de les dades de la fase inicial tots aquells estudiants que no apareguin al conjunt de dades de la fase no inicial.

4.2. Transformació de les dades

Un cop depurades les dades cal transformar-les amb l'objectiu d'obtenir un *DataFrame* definitiu que sigui còmode per treballar durant l'etapa de modelatge. En aquest cas construirem una única taula on trobarem un estudiant per fila i a les columnes la següent informació.

- Nota d'accés a la universitat
- Última nota definitiva de cada assignatura de la fase inicial
- Mitjana de notes de totes les convocatòries de cada assignatura de la fase inicial
- Nombre de convocatòries de cada assignatura de la fase inicial
- Primera nota de cada assignatura del tercer quadrimestre

El fet d'incorporar la mitjana de les notes de totes les convocatòries permet obtenir informació addicional sobre les notes dels estudiants que es presenten a varies convocatòries.

Pivotar les dades

Les dades inicials estan configurades de forma que cada fila correspon a una convocatòria d'un estudiant per una assignatura. Aquest format no és gaire útil per treballar, per tant, utilitzarem el pivotatge per aconseguir l'estructura desitjada. El procés de pivotatge ens permet convertir els valors de la variable *CODI_UPC_UD*, és a dir, les diferents assignatures, en els títols de columna, i els valors de la variable *CODI_EXPEDIENT* en índex de fila.

Per poder realitzar aquest tipus d'agrupacions disposem de la funció *pivot_table* de *pandas*. A més, ens permet realitzar diferents operacions matemàtiques sobre les dades de les columnes com per exemple sumatoris, mitjanes, màxims i mínims entre d'altres. Els paràmetres de la funció *pivot_table* que seran d'interès pel nostre estudi són els següents:

- *data*: *DataFrame* que volem pivotar.
- *index*: columna del *DataFrame* que es farà servir com a índex.
- *columns*: columna del *DataFrame* els valors de la qual s'usaran com títols de columna.
- *values*: columna del *DataFrame* que conté els valors de la nova taula.
- *aggfunc*: funció matemàtica que es vol aplicar al paràmetre *values*.

Com volem obtenir una sola taula que contingui tota la informació corresponent a cada matèria, construirem tres taules diferents que més endavant unirem. La primera taula conté la qualificació màxima obtinguda, per tant, al paràmetre *aggfunc* s'introdueix la funció *max*. D'altra banda, la segona taula ens mostra la mitjana aritmètica de les qualificacions de totes les convocatòries de cada assignatura. En aquest cas es fa ús de la funció *mean*. Per últim, la darrera taula ens proporciona el número de convocatòries que han estat necessàries per superar l'assignatura fent ús de la funció *count*.

A la Figura 3 es mostra un esquema del funcionament del pivotatge de dades.

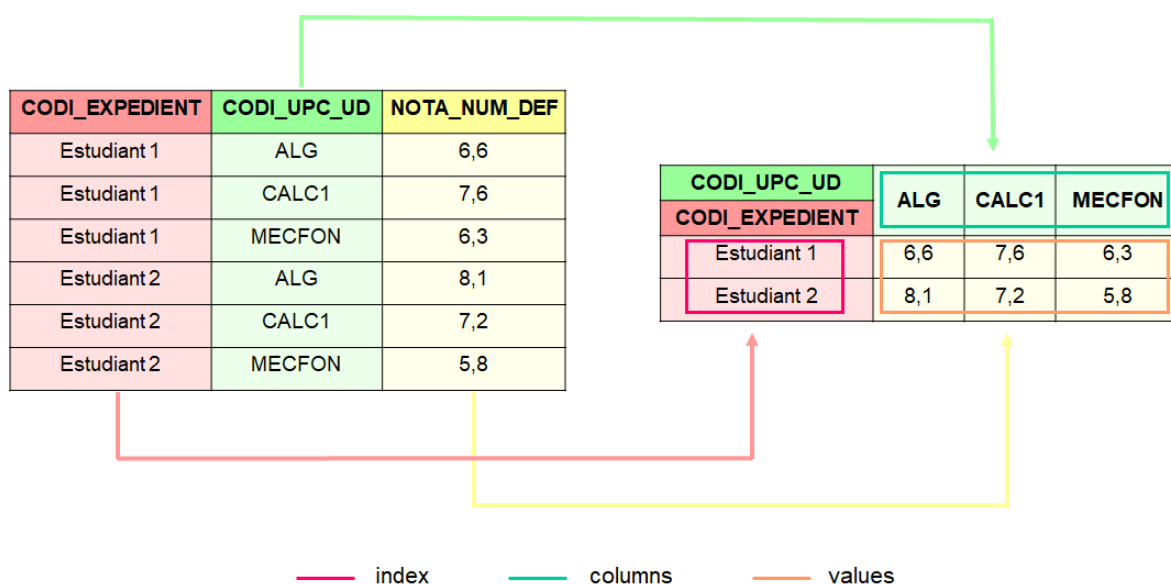


Figura 3: Esquema de pivotatge de dades

Canviar nomenclatura assignatures

Un cop construïda la taula amb les dades pivotades canviarem els noms de les columnes de les diferents assignatures ja que aquestes venen identificades amb el codi que els ha assignat l'escola. Com que treballar amb aquests codis és poc intuïtiu, farem servir la funció *rename* de *Pandas* per anomenar cada columna amb l'abreviatura de la matèria corresponent. A la Taula 3 trobem les equivalències entre el codi assignat per l'ETSEIB, el nom de l'assignatura i l'abreviatura assignada per treballar al *DataFrame*.

Cal destacar que per indicar la mitjana aritmètica s'utilitzarà una 'M' davant del nom de l'assignatura i pel nombre de convocatòries l'abreviatura 'Conv'.

CODI	ASSIGNATURA	ABREVIATURA
240011	ÀLGEBRA	ALG
240012	CÀLCUL 1	CALC1
240013	MECÀNICA FONAMENTAL	MECFON
240014	QUÍMICA 1	QUIM1
240015	FONAMENTS D'INFORMÀTICA	FONINFO
240021	GEOMETRIA	GEO
240022	CÀLCUL 2	CALC2
240023	TERMODINÀMICA FONAMENTAL	TERMOFON
240024	QUÍMICA 2	QUIM2
240031	ELECTROMAGNETISME	ELECTRO
240032	MÈTODES NUMÈRICS	METNUM
240033	MATERIALS	MAT
240131	EQUACIONS DIFERENCIALS	EQDIF
240132	INFORMÀTICA	INFO
240133	MECÀNICA	MEC

Taula 3: Equivalència codi-nom-abreviatura de les assignatures

Unir les taules

Per obtenir la taula definitiva amb la que treballarem durant l'etapa de modelatge, hem d'unir les tres taules de dades pivotades construïdes anteriorment i afegir una columna amb les notes d'accés provinents del fitxer de dades de preinscripció. La funció *merge* de *Pandas* és la que ens permet fer aquesta operació.

Eliminació dels valors NaN

Després de pivotar dades és molt probable que apareguin cel·les buides i, per tant, cal estudiar com tractar aquests valors buits que apareixen com a *NaN* al *DataFrame*. Aquests valors es poden tractar bàsicament de dues maneres. La primera és eliminar-los directament. La segona consisteix en determinar un valor fictici mitjançant un algoritme o criteri, com podria ser la mitjana de totes les qualificacions. En aquest cas, mitjançant la funció *dropna* de *Pandas*, s'ha decidit eliminar les files que contenen aquests valors nuls ja que assignar un valor fictici podria afectar a l'estudi.

Ordenar columnes

Seguidament, ordenarem totes les columnes de manera que primer aparegui la nota d'accés, seguida de les assignatures de la fase inicial i acabant amb les de la fase no inicial. Cal recordar que per a cada matèria de la fase inicial s'indica la qualificació màxima, la mitjana de totes les convocatòries i el nombre de convocatòries, en aquest ordre.

Transformació notes tercer quadrimestre

Per acabar, es transformaran les notes del tercer quadrimestre. L'objectiu de l'estudi es determinar si l'alumne aprovarà o suspèn una determinada assignatura i no predir la nota exacta. Per aquest motiu, classificarem les notes del tercer quadrimestre en aquestes dues categories indicant amb el valor 1 l'aprobat i el valor 0 pel suspès.

Un cop realitzades totes les activitats mencionades anteriorment, hem acabat amb la fase de preparació de dades i estem llestos per començar amb l'etapa de modelatge. A la Taula 4 es pot observar una mostra del *DataFrame* definitiu amb el que treballarem durant les etapes posteriors de l'estudi.

CODI_EXPEDIENT	NOTA_ACCES	ALG	M_ALG	Conv_ALG	...	ELECTRO	METNUM	MAT	EQDIF	INFO	MEC
226410	12,507	6,6	6,6	1		1	1	1	1	1	1
226431	11,796	6,2	5,1	2		1	1	1	1	1	0
226455	10,642	8,1	8,1	1		1	1	1	1	1	1
226464	10,916	5	4,5	2		1	1	1	0	1	1
226467	11,686	7,1	4,867	3		0	0	0	0	0	0

Taula 4: Mostra del *DataFrame* definitiu

5. Modelatge i validació

Un cop hem acabat amb la comprensió i preparació de les dades comencen les etapes de modelatge i validació. En primer lloc, es construirà el model per després validar-lo i estimar el rendiment dels resultats que ens proporciona. Si s'obté un bon rendiment dels resultats del model es continuarà amb el projecte. En cas contrari, es tornarà enrere per modificar alguns paràmetres del model o bé provar-ne un de nou. Per aquest motiu, cal dur a terme aquestes dues fases de forma conjunta.

5.1. Models predictius

Un model predictiu és un model matemàtic o d'altre tipus que utilitza dades existents per predir resultats futurs mitjançant l'anàlisi de patrons. El seu objectiu és avaluar la probabilitat de que una unitat similar en una mostra diferent exhibeixi un comportament específic. Els models predictius sovint executen càlculs durant les transaccions en curs de forma que aportí coneixement a l'hora de prendre una decisió. Gràcies als avanços de l'enginyeria en l'anàlisi de grans volums de dades aquests models són capaços de simular el comportament humà davant estímuls o situacions específiques.

En aquest cas es vol predir si un alumne suspendrà una determinada assignatura en funció de les seves qualificacions en les matèries de la fase inicial. Per poder predir si l'estudiant aprovarà o no, el model buscarà relacions entre aquestes notes i els patrons que ha creat a partir de les dades anteriors.

Existeixen dos tipus de models predictius. Per una banda, els models de classificació permeten predir la pertinència a una classe. D'altra banda, els models de regressió permeten predir un valor. En el nostre cas d'estudi farem ús dels models de classificació ja que l'objectiu és determinar si l'alumne aprovarà o suspendrà. Dins d'aquesta categoria trobem diversos models com poden ser la regressió logística, els arbres de decisió, k-veïns més propers o màquines de vectors de suport entre d'altres. Finalment s'ha optat per utilitzar la regressió logística, un model senzill i fàcil d'executar, i els arbres de decisió, un model molt diferent a l'anterior i un dels més utilitzats tal i com podem veure a la Figura 4. El fet d'escollir dos models molt diferents ens permetrà fer una comparació molt interessant entre ambdós resultats.

A continuació, es farà una breu explicació de tots els models mencionats anteriorment aprofundint més en els dos escollits per a l'estudi, la regressió logística i els arbres de decisió.

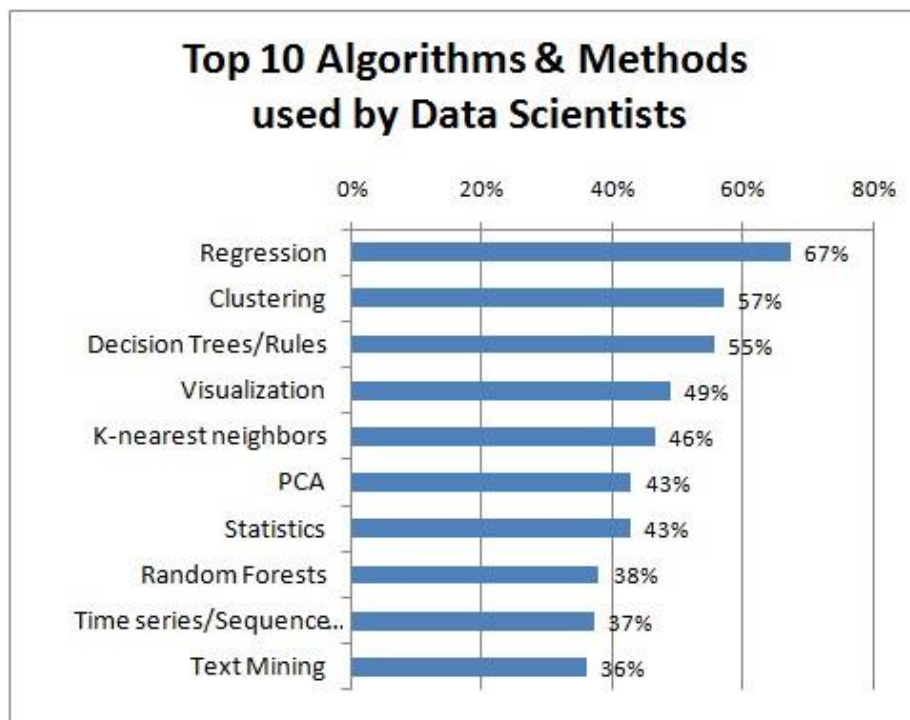


Figura 4: Top 10 d'algoritmes i mètodes utilitzats pels enginyers de dades [3]

5.1.1. Regressió logística

La regressió logística permet predir el resultat d'una variable qualitativa en funció de variables quantitatives. És un dels més comuns i és molt útil per modelar la probabilitat quan la variable categòrica només té dos possibles opcions, en el nostre cas aprovat o suspens. [4]

L'equació de la regressió logística és molt similar a la de la regressió lineal, però aplicant el que s'anomena funció logística. La funció logística (Eq. 1) és una corba en forma de S que pot prendre qualsevol valor real i assignar-lo a un valor entre 0 i 1.

$$f(t) = \frac{1}{1 + e^{-t}} \quad (\text{Eq. 1})$$

En l'equació de la regressió logística (Eq. 2) els valors d'entrada x_i es combinen linealment utilitzant pesos de coeficients β_i per predir un valor de sortida p . El valor de sortida s'interpreta com una probabilitat que ens permetrà determinar el valor de la variable categòrica. Tal i com s'observa a la Figura 5 si la probabilitat és major del 50%, la variable resposta pren valor 1 i, per tant, la predicció és d'aprovat. D'altra banda, si és menor se li assignarà el valor 0 el qual significa suspès.

$$p(x_1, \dots, x_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}} \quad (\text{Eq. 2})$$

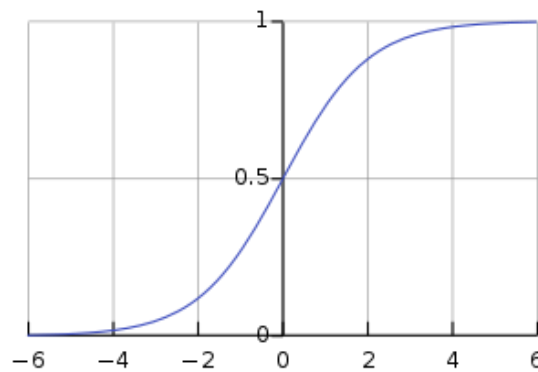


Figura 5: Gràfica de la regressió logística [4]

Per programar el model de regressió logística en *Python* és fa ús de la classe `sklearn.linear_model.LogisticRegression()`. Un dels paràmetres més importants de la regressió logística és el paràmetre *C* o inversa de la força de regularització que ha de ser un real positiu. La regularització és un mètode per evitar el sobre ajust (comentarem aquest concepte més endavant amb detall), mitjançant la penalització de models. La regularització funciona afegint la penalització associada als valors de coeficient a l'error de la hipòtesi. Per tant, un model precís amb valors de coeficient elevats es penalitzaria més i un model menys precís amb valors més petits es penalitzaria menys. Es farà diverses proves per determinar el valor òptim de *C* que ens proporciona un millor rendiment.

5.1.2. Arbre de decisió

L'arbre de decisió és un model amb estructura d'arbre que permet classificar exemples i ve a representar un conjunt de regles. Per construir-lo es parteix d'un únic node que es va ramificant amb els possibles resultats. Aquests, donaran lloc a altres nodes que també es ramificaran, i així successivament formant l'estructura d'arbre. Els nodes de l'arbre contenen les característiques del conjunt de dades mentre que les branques o ramificacions representen les regles de decisió. Dins dels nodes en trobem de dos tipus:

- Node arrel: node a partir del qual es desenvolupa l'arbre.
- Node intern: node utilitzat per prendre una decisió i que té diverses branques. Està situat entre el node arrel i el node terminal o fulla.
- Node fulla: node terminal que representa el resultat de les decisions i que no conté més branques.

A la Figura 6 es simula un exemple d'arbre de decisió per predir el nombre d'aprovats i suspesos que hi haurà a l'assignatura d'Electromagnetisme.

En aquest cas, es parteix de la mitjana de l'assignatura de Termodinàmica Fonamental que és el node arrel establint la condició: si la qualificació és inferior o igual a 6,025 es ramificarà cap a l'esquerra i, en cas contrari, es prendrà la ramificació de la dreta. Aquest procediment s'anirà repetint consecutivament amb la resta de nodes que proporcionaran noves condicions a partir de les decisions preses segons el criteri de la impuresa de Gini que s'explica més endavant.

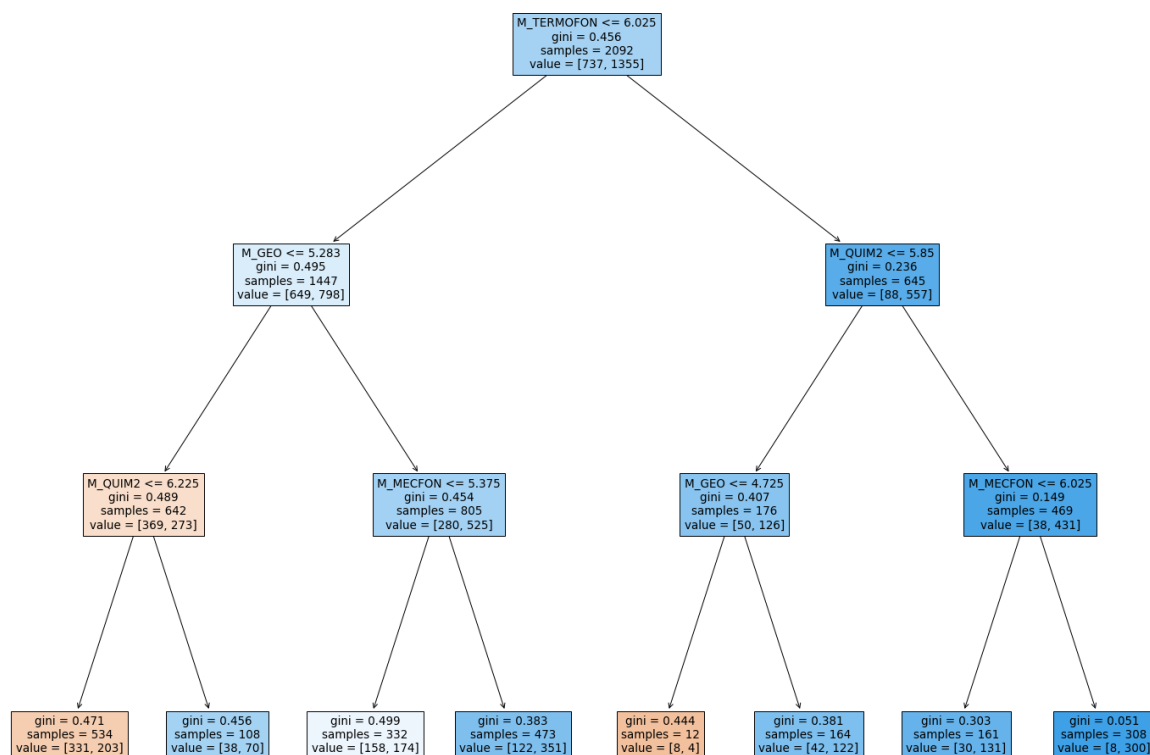


Figura 6: Exemple arbre de decisió

Per programar el model d'arbres de decisió a *Python* s'utilitza la classe `sklearn.tree.DecisionTreeClassifier()`. Aquesta funció consta de diversos paràmetres dels quals es poden destacar els següents:

- **criterion:** funció que utilitza per mesurar la qualitat d'una divisió. Els criteris admesos són “*gini*” per a la impuresa de Gini i “*entropy*” per al guany d'informació. En aquest cas s'utilitzarà el criteri que ve per defecte “*gini*”. L'índex de Gini (Eq. 3) quantifica la variància total en el conjunt de les K classes del node m , és a dir, mesura la puresa del node. Quan p_{mk} és proper a 0 o a 1 (el node conté majoritàriament observacions d'una sola classe), el terme $p_{mk}(1 - p_{mk})$ és molt petit. Com a conseqüència, quant més gran sigui la puresa del node, menor serà el valor d'índex Gini.

$$G_m = \sum_{k=1}^K p_{mk} \cdot (1 - p_{mk}) \quad (\text{Eq. 3})$$

- **max_depth**: defineix la profunditat màxima de l'arbre, és a dir, fins a quin nivell de nodes es ramificarà l'arbre des del node arrel. Aquest paràmetre per defecte pren el valor "None", deixarà que l'arbre es vagi ramificant fins que tots els nodes fulla siguin purs. Si deixem aquest paràmetre per defecte és molt probable que retorni un arbre de decisió sobre ajustat. Per evitar aquest problema es construiran diversos arbres amb diferents profunditats i així poder escollir la que ens proporcioni millors resultats.

5.1.3. K-Nearest neighbors

El mètode *K-Nearest Neighbors* (o K veïns propers en català) busca en les observacions més properes a la que s'està intentant predir i classifica el punt d'interès basat en la majoria de dades que el rodegen (Veure Figura 7). És a dir, calcula la distància entre el nou punt i cadascun dels ja existents i ordena aquestes distàncies de menor a major per poder seleccionar el grup al que pertany.

És un algoritme d'aprenentatge supervisat, a partir d'un conjunt de dades inicial el seu objectiu serà classificar correctament totes les noves instàncies. A diferència d'altres algoritmes d'aprenentatge supervisat, el *K-Nearest Neighbors* no genera un model fruit de l'aprenentatge amb dades d'entrenament, sinó que emmagatzema tots els registres possibles per poder classificar més endavant les dades de test. A aquest tipus d'algoritmes se'ls anomena *lazy learning methods* (mètodes d'aprenentatge mandrós en català).

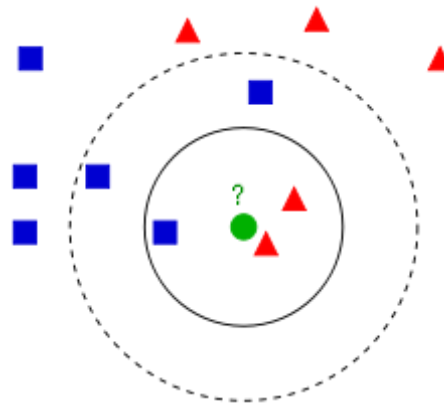


Figura 7: Exemple algoritme K-Nearest Neighbors [5]

5.1.4. Support Vector Machines (SVM)

El mètode *Support Vector Machines* (o màquines de vectors de suport en català) és un algoritme d'aprenentatge supervisat que té com objectiu trobar un hiperplà que separi de la millor manera possible dues classes diferents de punts de dades. Això implica trobar l'hiperplà amb el marge més ampli entre les dues classes. El marge es defineix com l'amplada màxima de la regió paral·lela a l'hiperplà que no té punts de dades interiors. (Veure Figura 8)

L'algoritme només pot trobar aquest hiperplà en problemes que permeten separació lineal. No obstant, *Support Vector Machines* pertany a una classe d'algoritme d'aprenentatge automàtic denominats mètodes kernel, on es pot utilitzar una funció kernel per transformar les característiques. Aquestes funcions assignen les dades a un espai dimensional diferent, normalment superior, amb l'objectiu de que resulti més fàcil separar les classes després d'aquesta transformació, simplificant potencialment els límits de decisió complexos no lineals per fer-los lineals en l'espai dimensional de característiques superiors assignat.

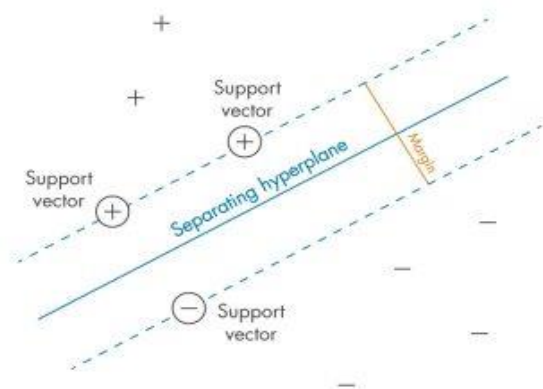


Figura 8: Exemple algoritme Support Vector Machines SVM [6]

5.2. Mètriques d'avaluació

L'avaluació del model d'aprenentatge automàtic és essencial en qualsevol projecte. Per tant, és necessari definir les mètriques que s'utilitzaran per a mesurar la qualitat del model. Aquest projecte es centra en la predicció de suspesos. No obstant, en cap cas es pot deixar de banda la predicció dels aprovats ja que això ens portaria a una predicció falsa. Cal tenir en compte que un model pot funcionar bé amb una mesura d'una mètrica d'avaluació però fer-ho malament amb una de diferent. Per aquest motiu, és molt important utilitzar diverses mètriques per avaluar el rendiment del model i així garantir que el model funcioni de forma correcta i òptima. Tot seguit, s'explicaran les mètriques *Confusion Matrix*, *Accuracy*, *Precision*, *Recall* i *F1-score*, que són les que s'utilitzaran al llarg de l'estudi.

5.2.1. Confusion Matrix

La matriu de confusió o *Confusion Matrix* (Figura) és una representació matricial dels resultats de les prediccions de qualsevol prova binària que s'utilitza sovint per descriure el rendiment del model sobre un conjunt de dades de prova els valors reals de la qual es coneixen. La matriu de confusió ens permet detectar si el sistema està confonent les classes de classificació. Cada columna representa les instàncies en la classe predita i cada fila les instàncies en la classe real.

		Valor de la predicció	
		Negatiu	Positiu
Valor real	Negatiu	TN True Negative	FP False Positive
	Positiu	FN False Negative	TP True Positive

Figura 9: Confusion Matrix

Cada predicció es pot classificar en quatre resultats en funció de la seva coincidència amb el valor real:

- True Negative (TN): el valor predit com a negatiu coincideix amb el real que també és negatiu.
- False Positive (FP): el valor predit com a positiu és incorrecte ja que el valor real és negatiu.
- False Negative (FN): el valor predit com a negatiu és incorrecte ja que el valor real és positiu.
- True Positive (TP): el valor predit com a positiu coincideix amb el real que també és positiu.

Cal destacar que en el nostre cas els negatius corresponen als suspesos i els positius als aprovats.

5.2.2. Accuracy

L'exactitud o *Accuracy* (Eq. 4) indica la relació entre el número de prediccions correctes i el número total de mostres d'entrada, és a dir, mesura el percentatge de casos que el model ha classificat correctament. Es calcula mitjançant la següent equació:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (\text{Eq. 4})$$

Cal destacar que la mètrica exactitud o *accuracy* té certes limitacions. No funciona correctament quan les classes estan desequilibrades, és a dir, quan una classe conté molts elements en comparació amb l'altra. En aquests casos serà millor utilitzar altres mètriques que veurem a continuació.

5.2.3. Precision

La precisió o *Precision* (Eq. 5) és una mètrica específica de la classe que indica la capacitat del model de classificar només els punts de dades rellevants. Aquesta mètrica es defineix com el nombre de positius predits correctament entre el nombre total de positius predits.

$$Precision = \frac{TP}{TP + FP} \quad (\text{Eq. 5})$$

5.2.4. Recall

L'exhaustivitat o *Recall* (Eq. 6) determina la capacitat del model per trobar tots els casos rellevants dins d'un conjunt de dades. Aquesta mètrica es defineix com el nombre de positius veritaders entre el nombre de positius veritaders i negatius falsos, és a dir, el nombre de positius predits correctament entre el nombre total real de positius.

$$Recall = \frac{TP}{TP + FN} \quad (\text{Eq. 6})$$

5.2.5. F1-score

És possible que en alguna situació sapiguem que volem maximitzar la precisió o l'exhaustivitat a costa de reduir l'altre. No obstant, hi ha casos en que volem trobar una combinació òptima entre les dues mètriques ja que, normalment, a major precisió menor exhaustivitat i viceversa. La puntuació *F1* o *F1-Score* (Eq. 7) ens serà molt útil en aquests casos perquè és una combinació d'ambdues mètriques. Es calcula fent la mitjana harmònica entre la precisió i l'exhaustivitat.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (\text{Eq. 7})$$

5.3. Mètodes de validació

La finalitat d'un model és predir la variable resposta en observacions futures o observacions que el model no ha vist abans. L'error mostrat per defecte després d'executar un model amb tot el conjunt de dades el que comet en predir observacions ja vistes i no és una estimació realista de com es comporta el model davant noves dades. És aquí on apareixen els conceptes *overfitting* i *underfitting*.

El *overfitting* succeeix quan el model només s'ajusta a aprendre els casos particulars que li ensenya i, per tant, és incapaç de reconèixer noves dades d'entrada. Quan sobre entrenem el model i es produeix *overfitting*, l'algoritme estarà considerant com a vàlides només les dades idèntiques a les del nostre conjunt d'entrenament incloent els seus defectes i sent incapaç de distingir les entrades bones com a fiables si es surten una mica dels rangs preestablerts. D'altra banda, es produeix *underfitting* quan el model és molt simplista i no pot capturar els matisos, particularitats i complexitats en les dades, fet que ens provocarà prediccions nefastes ja que el model serà incapaç de reconèixer noves dades. Per no caure en cap d'aquests dos problemes caldrà trobar un punt mig en l'aprenentatge del model.

Per aconseguir una estimació més carter a es recorre al que s'anomenen estratègies de validació. Els mètodes de validació són estratègies que permeten estimar la capacitat predictiva dels models quan s'apliquen a noves observacions. La idea en la que es basen és dividir el conjunt de dades en dos subconjunts el d'entrenament i el de prova. El conjunt d'entrenament serveix per entrenar el model mentre que el de prova permet avaluar-lo. A continuació s'explicarà en que consisteixen alguns d'aquests mètodes de validació com son el *Hold-out* i *K-Fold Cross Validation*.

5.3.1. Hold out

El mètode *hold-out* consisteix en dividir el conjunt de dades amb les quals es treballa en dos subconjunts, el d'entrenament o *train* i el de prova o *test*. És habitual assignar el 80% de les dades al conjunt d'entrenament i el 20% restant al conjunt de proves.

El conjunt de dades d'entrenament és examinat per l'algoritme d'aprenentatge per determinar, o aprendre, les combinacions òptimes de variables que generaran un bon model predictiu. L'objectiu és aconseguir un model entrenat que pugui fer prediccions precises de dades desconegudes. Per altra banda, el conjunt de dades de proves serveix per a què el model faci prediccions i així poder estimar-ne el rendiment de forma fiable. Per reduir el risc de problemes com el sobre ajust o *overfitting* les dades del conjunt de proves no poden utilitzar-se per entrenar el model.

El principal inconvenient d'aquest mètode és el fet que suposa que les dades dels dos subconjunts són del mateix tipus, és a dir, tenen les mateixes propietats exactes. Com que es tracta d'una simple divisió aleatòria, aquesta suposició pot no ser certa. No obstant, per conjunts de dades grans és un mètode molt útil.

5.3.2. K-Fold Cross Validation

El mètode *K-Fold Cross Validation*, al igual que el *hold-out*, també divideix les dades en dos subconjunts, el d'entrenament i el de prova. La diferència és que el conjunt total de dades es divideix aleatòriament en k grups. Un dels grups s'utilitza com a conjunt de prova i la resta com a conjunt d'entrenament. El model s'entrena amb les dades d'entrenament i es puntua el rendiment amb les de prova. Seguidament, es repeteix el procediment fins que cada grup únic s'hagi utilitzat com a conjunt de prova. A la Figura 10 es pot observar el funcionament d'aquest mètode de validació.

És un mètode molt utilitzat ja que dona al model l'oportunitat d'entrenar-se en múltiples divisions de prova. Aquest fet proporciona una millor indicació del rendiment del model. Ara bé, cal mencionar que la validació creuada utilitza múltiples divisions de prova, per tant, necessita més poder computacional i més temps d'execució que el mètode *hold-out*.

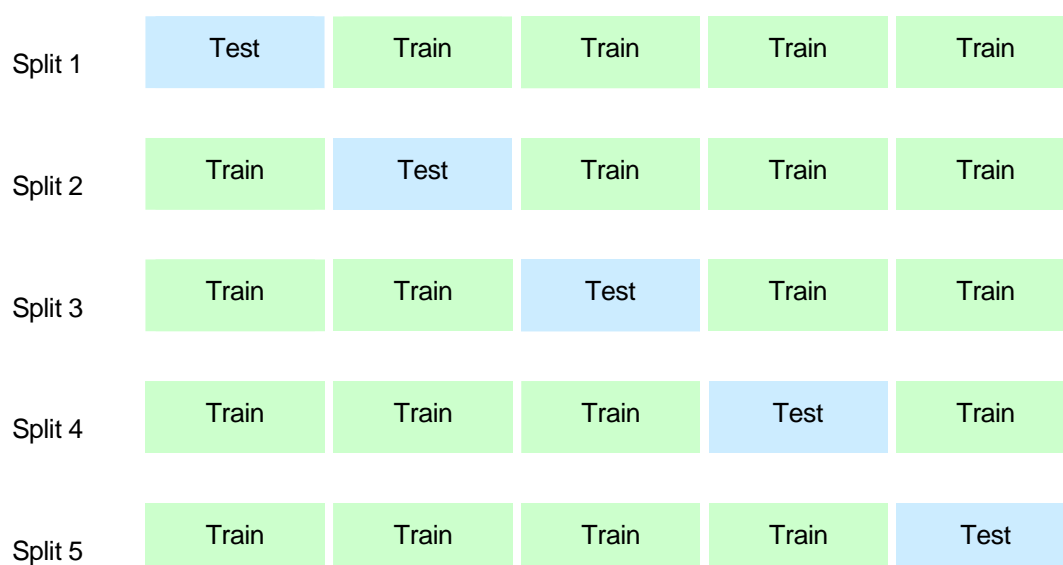


Figura 10: Iteracions K-Fold Cross Validation

6. Tècniques de mostreig pel desequilibri de dades

En els problemes de classificació és molt habitual trobar conjunts de dades desequilibrades on les mostres d'una classe superen significativament les de l'altra. La classe minoritària sol representar el concepte més important d'aprendre i és difícil identificar-lo. És el cas de les nostres dades, on trobem que hi ha una gran quantitat d'aprovats enfront als suspesos en la majoria d'assignatures del tercer quadrimestre tal i com podem veure a la Figura 11.

Les tècniques d'aprenentatge automàtic sovint fracassen o ofereixen un rendiment enganyosament optimista en conjunts de dades amb una distribució de classes desequilibrada. La raó és que molts algorismes estan dissenyats per operar amb dades amb un nombre equilibrat d'observacions per a cada classe i, quan aquest no és el cas, els algorismes interpreten que molts pocs exemples no són importants i es poden ignorar per aconseguir un bon rendiment.

El mostreig de dades proporciona una col·lecció de tècniques que transformen el conjunt de dades d'entrenament per tal d'equilibrar la distribució de la classe. Un cop equilibrades, els algorismes d'aprenentatge automàtic es poden entrenar directament sobre el conjunt de dades transformat sense cap modificació permetent abordar el repte de la classificació desequilibrada.

La solució més popular per una per un problema de classificació desequilibrada és aplicar tècniques de mostreig per canviar la composició del conjunt de dades d'entrenament. Bàsicament, en lloc de tractar el model amb el desequilibri, podem intentar equilibrar les freqüències de classe eliminant així la qüestió del desequilibri que afecta a la formació de models. El mostreig només es realitza al conjunt de dades d'entrenament i no al conjunt de prova ja que la intenció és eliminar el biaix de classe de l'ajust del model, però avaluant el model resultant en dades reals i representatives del domini del problema objectiu. Principalment, existeixen dos tipus de mostreig de dades que són *Oversampling* i *Undersampling*.

Per programar l'aplicació dels mètodes *Oversampling* i *Undersampling* es farà ús de la llibreria *imbalanced-learn* de *Python* que ofereix varietat de tècniques de mostreig que s'utilitzen habitualment en conjunts de dades desequilibrades.

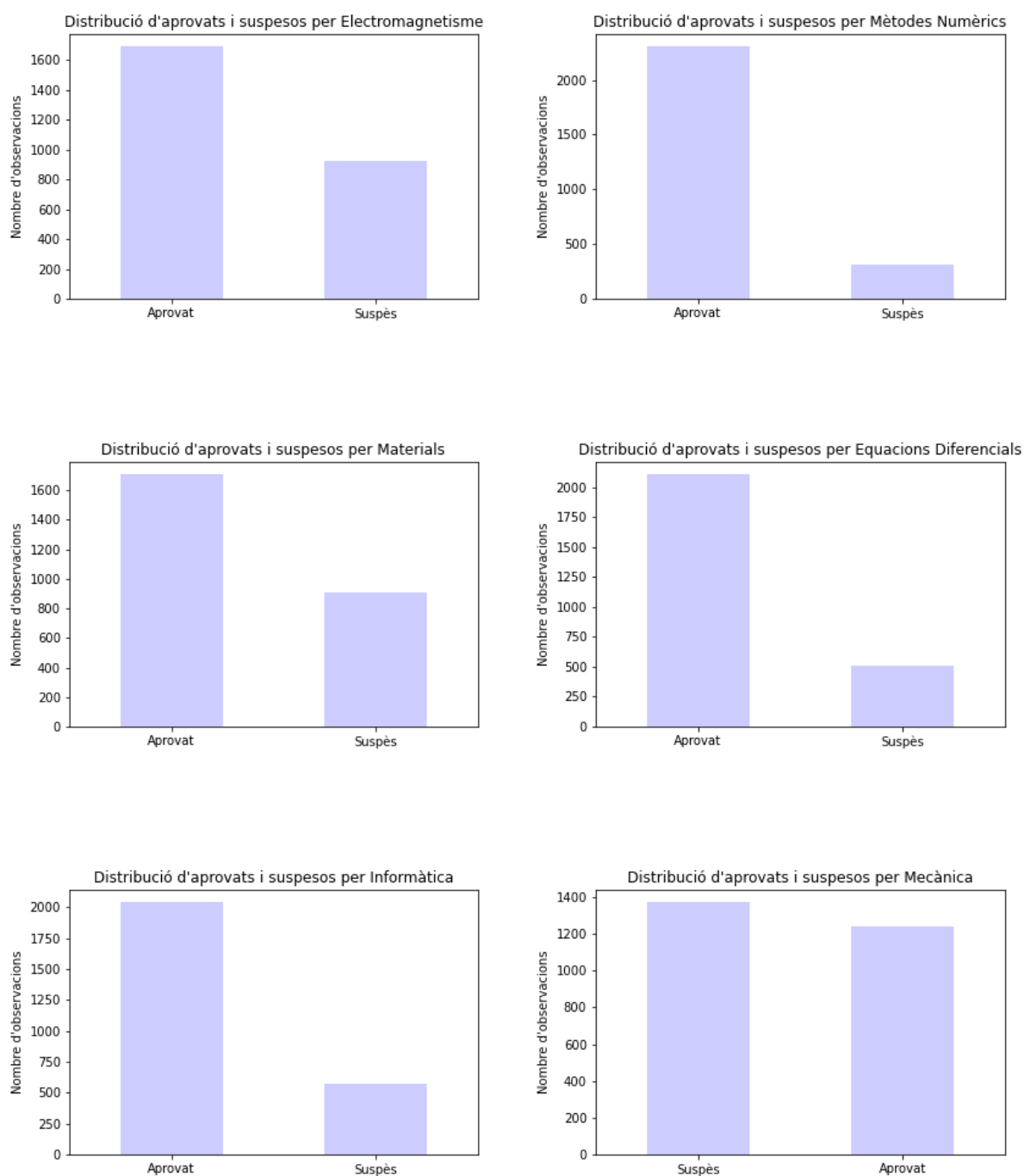


Figura 9: Distribució d'aprovat i suspesos de cada assignatura del tercer quadrimestre

6.1. Oversampling

Oversampling és una tècnica de sobre mostreig que consisteix en duplicar exemples existents o bé sintetitzar-ne de nous a partir de mostres de la classe minoritària per poder equilibrar la distribució de dades. A la Figura 12 es mostra el seu funcionament.

Existeixen diversos mètodes de *Oversampling*. En aquest projecte s'implementaran tres dels més utilitzats com són *Random Oversampling*, *Synthetic Minority Oversampling Technique (SMOTE)* i *Borderline-SMOTE*.

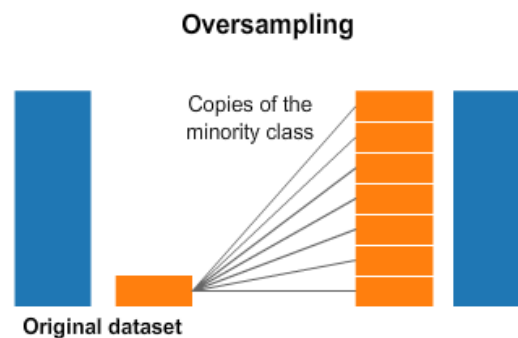


Figura 10: Funcionament Oversampling [7]

6.1.1. Random Oversampling

El mètode *Random Oversampling* implica duplicar aleatòriament mostres de la classe minoritària i afegir-los al conjunt de dades d'entrenament.

En alguns casos, el sobre mostreig aleatori pot provocar que els algoritmes afectats s'adaptin a la classe minoritària ja que fa còpies exactes dels seus exemples, cosa que provoca un sobre ajust. L'efecte pot ser un millor rendiment en el conjunt de dades d'entrenament, però pitjor en el conjunt de prova.

El sobre mostreig aleatori es pot implementar mitjançant la classe *RandomOverSampler()* indicant al paràmetre *sampling_strategy* la classe on volem afegir mostres. Un altre paràmetre important és *shrinkage*, que afegeix una petita pertorbació a les mostres creades.

6.1.2. Synthetic Minority Oversampling Technique (SMOTE)

Com ja s'ha mencionat anteriorment, una manera de resoldre el problema de les dades desequilibrades és duplicar exemples de la classe minoritària. Això pot equilibrar la distribució de la classe, però no proporciona cap informació addicional al model. Una millora en la duplicació d'exemples és sintetitzar-ne de nous a partir de les mostres ja existents.

El mètode més utilitzat per sintetitzar nous exemples s'anomena *Synthetic Minority Oversampling Technique* o *SMOTE* i funciona seleccionant exemples propers a l'espai de característiques, dibuixant una línia entre els exemples de l'espai de característiques i dibuixant una nova mostra en un punt al llarg d'aquesta línia. Concretament, primer es tria un exemple aleatori de la classe minoritària i llavors es troben k dels veïns més propers. Aleshores es tria un veí seleccionat aleatòriament i es crea un exemple sintètic en un punt seleccionat aleatòriament entre els dos exemples de l'espai de característiques. A la Figura 13 es pot observar el funcionament d'aquesta tècnica de mostreig.

El mètode SMOTE es pot implementar mitjançant la classe *SMOTE()* indicant al paràmetre *sampling_strategy* la classe on volem afegir mostres i al paràmetre *k_neighbors* el nombre de veïns propers per construir els exemples sintètics.

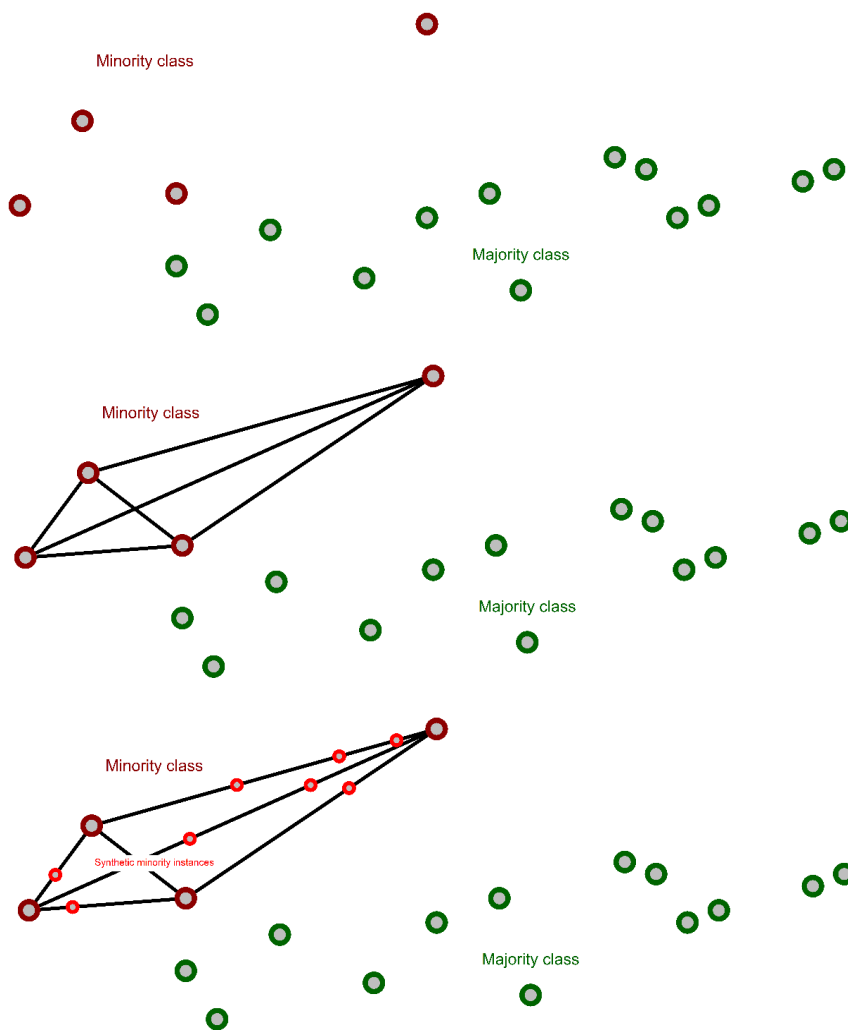


Figura 11: Funcionament tècnica SMOTE [8]

6.1.3. Borderline-SMOTE

El mètode *Borderline-SMOTE* és una extensió popular del mètode *SMOTE* i consisteix a seleccionar aquelles instàncies de la classe minoritària que estan mal classificades. El nombre de veïns de cada instància minoritària s'utilitza per dividir les instàncies minoritàries en 3 grups; segur, perill o soroll. Aleshores, el model fa el sobre mostreig només en aquells casos difícils, els de perill, proporcionant més resolució només quan sigui necessari.

Els exemples del límit i els propers són més aptes per ser classificats erròniament que els que estan lluny del límit i, per tant, són més importants per a la classificació. Aquests exemples mal classificats probablement siguin ambigus i es trobin en una regió de la vora o la frontera del límit de decisió on la pertinença a la classe es pugui superposar. Per tant, en lloc de generar nous exemples sintètics per a la classe minoritària a cegues, el mètode *Borderline-SMOTE* només crea exemples sintètics al llarg del límit de decisió entre les dues classes.

A la Figura 14 podem veure una comparació entre els mètodes *SMOTE* i *Borderline-SMOTE* que mostra com es creen els nous exemples. Els punts de color lila formen part de la classe minoritària mentre que els grocs són de la classe majoritària. Amb el mètode *SMOTE* els exemples es creen construint línies entre nodes veïns aleatòriament. En canvi, amb el *Borderline-SMOTE* es fa el mateix però entre els nodes veïns molt a prop del límit respecte a la classe majoritària.

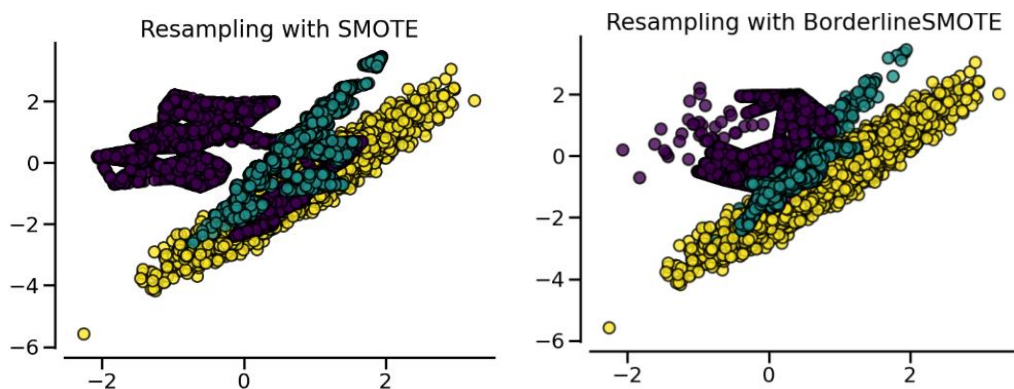


Figura 12: Comparació SMOTE i Borderline-SMOTE [9]

6.2. Undersampling

Undersampling és una tècnica de submostreig que consisteix en eliminar exemples del conjunt de dades d'entrenament que pertanyen a la classe majoritària per tal d'equilibrar millor la distribució de la classe. Podem veure com funciona a la Figura 15. Pot ser un enfocament més adequat per a aquells conjunts de dades en què, tot i que hi hagi un desequilibri de dades, hi ha un nombre suficient d'exemples de la classe minoritària.

En el cas que ens ocupa, algunes de les assignatures del tercer quadrimestre contenen molt pocs exemples de la classe minoritària, per tant, només s'aplicaran tècniques de sobre mostreig. No obstant, s'exposarà en que consisteixen algunes tècniques de submostreig com son *RandomUndersampling*, *Near Miss Undersampling* i *Tomek Links for Undersampling*.

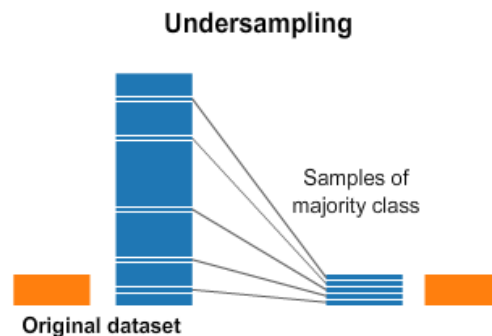


Figura 13: Funcionament Undersampling [7]

6.2.1. Random Undersampling

El mètode *Random Undersampling* implica seleccionar aleatòriament exemples de la classe majoritària per eliminar del conjunt de dades d'entrenament.

Una limitació del submostreig aleatori és que es suprimeixen exemples de la classe majoritària que poden ser útils, importants o, fins i tot, crítics per ajustar una decisió sòlida. Com els exemples es suprimeixen a l'atzar, no hi ha manera de detectar o conservar exemples "bons" o amb més informació de la classe majoritària.

6.2.2. Near Miss Undersampling

La tècnica *Near Miss Undersampling* usa una col·lecció de mètodes de submostreig que seleccionen exemples basats en la distància de les mostres de classes majoritàries a les de classes minoritàries. En aquest cas, la distància es determina en l'espai de característiques mitjançant la distància euclidiana o similar.

Hi ha tres versions de la tècnica, anomenades *NearMiss-1*, *NearMiss-2* i *NearMiss-3*.

- *NearMiss-1*: selecciona exemples de la classe majoritària que tenen la distància mitjana més petita a les tres mostres més properes de la classe minoritària.
- *NearMiss-2*: selecciona exemples de la classe majoritària que tenen la distància mitjana més petita a les tres mostres més allunyades de la classe minoritària.
- *NearMiss-3*: selecciona un nombre determinat d'exemples de classes majoritàries per a cada mostra de la classe minoritària més propera.

6.2.3. Tomek Links for Undersampling

El mètode *Tomek Links* és una de les modificacions de la tècnica de submostreig *Condensed Nearing Neighbors (CNN)*. A diferència del mètode *CNN* que només selecciona a l'atzar les mostres amb els seus k veïns més propers de la classe majoritària, *Tomek Links* troba parells d'exemples, un de cada classe, que tenen la menor distància euclidiana entre sí en l'espai característic. Aquests parells entre classes es coneixen generalment com "Tomek Links" i són valuosos ja que defineixen el límit de classe. Finalment, s'acaba eliminant l'element majoritari de l'enllaç, que proporciona un límit de decisió millor per a un classificador. A la Figura 16 podem veure com funciona aquest mètode.

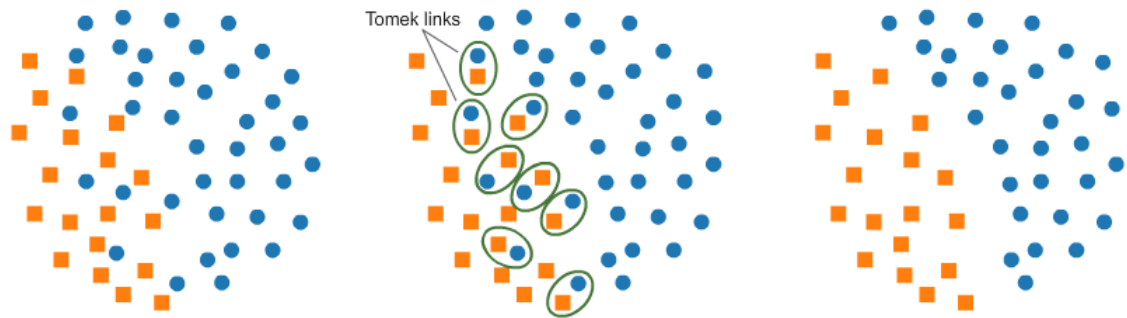


Figura 14: Funcionament mètode Tomek Links [7]

7. Anàlisi de resultats

Un cop vistos diversos models predictius i tècniques de mostreig, és el moment de construir els diferents models i analitzar els resultats que ens proporcionen. S'aplicaran diferents tècniques de mostreig amb l'objectiu d'equilibrar el conjunt de dades i comparar com varia el rendiment en la predicció dels resultats acadèmics amb cadascuna d'aquestes tècniques.

Per poder realitzar aquest estudi s'ha escollit dos models predictius com són la regressió logística i els arbres de decisió. Per validar-los s'ha utilitzat el mètode *Holdout* on es divideix el conjunt de dades en dos subconjunts, el *training set* i el *testing set*. Aquesta divisió de dades es farà de forma que el conjunt d'entrenament contingui un 80% de les dades i el conjunt de test estigui format pel 20% restant. La distribució de dades es farà de manera aleatòria. S'ha fixat el paràmetre *random_state* per aconseguir que la divisió aleatòria de *train* i *test* sigui la mateixa en tots els experiments i, per tant, es puguin comparar entre ells. Encara que la divisió sigui de forma aleatòria, s'ha preservat el percentatge d'aprovat i suspesos en ambdós subconjunts per mantenir la semblança amb el conjunt original i fer així una millor predicció. Per poder mantenir aquesta distribució d'aprovat i suspesos s'ha usat el paràmetre *stratify*.

Un cop tenim les dades dividides en els conjunts de *training* i *testing* és hora d'aplicar les diferents tècniques de mostreig que volem comparar. En aquest cas s'ha escollit només tècniques d'*oversampling* ja que el conjunt de dades no és molt gran i, per tant, amb tècniques d'*undersampling* ens quedaria molt poca quantitat de dades. Es compararan els resultats assolits amb *RandomOverSampler*, *SMOTE* i *BorderlineSMOTE* entre ells i amb els obtinguts sense aplicar tècniques de mostreig. D'aquesta manera es podrà veure la diferència entre conjunts de dades desequilibrades i quan s'aplica la tècnica de mostreig per equilibrar les dues classes. Cal recordar que les tècniques de mostreig només s'apliquen al conjunt de *training*.

Per poder avaluar el rendiment de la predicció dels resultats acadèmics s'utilitzarà principalment la mètrica *F1*. Tanmateix, també es farà ús de les mètriques *Precision* i *Recall* com a suport de l'anterior. És necessari mencionar que els models predictius i les tècniques de mostreig utilitzats tenen paràmetres que poden fer variar els resultats. Per aquest motiu s'ha analitzat com varia *F1* de la classe suspesos quan modifiquem el valor d'aquests paràmetres.

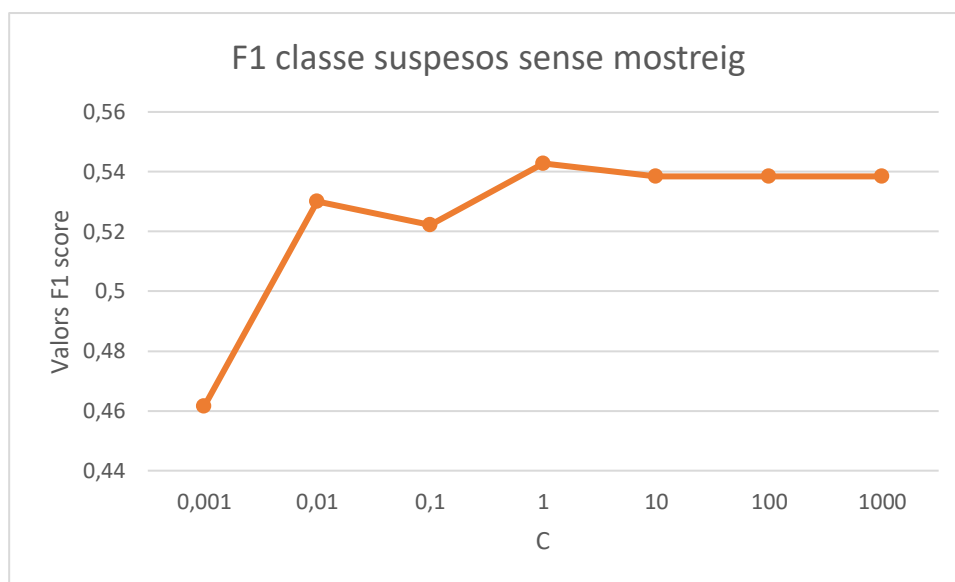
7.1. Regressió logística

En aquest apartat és durà a terme l'anàlisi amb el model predictiu regressió logística. Es mostraran els resultats obtinguts sense tècniques de mostreig i amb les tècniques *RandomOverSampler*, *SMOTE* i *BorderlineSMOTE*.

Per a cadascun d'ells es mostrarà gràficament com varia la mètrica $F1$ de la classe suspesos en funció dels paràmetres C de la regressió logística, *shrinkage* de *RandomOverSampler* i *k-neighbors* de *SMOTE* i *BorderlineSMOTE*. Un cop vist com afecten els paràmetres a la predicció de suspesos, es seleccionarà la configuració que ens proporciona un millor rendiment i es mostraran les mètriques *Precision*, *Recall* i $F1$ per ambdues classes.

7.1.1. Electromagnetisme

Sense tècniques de mostreig

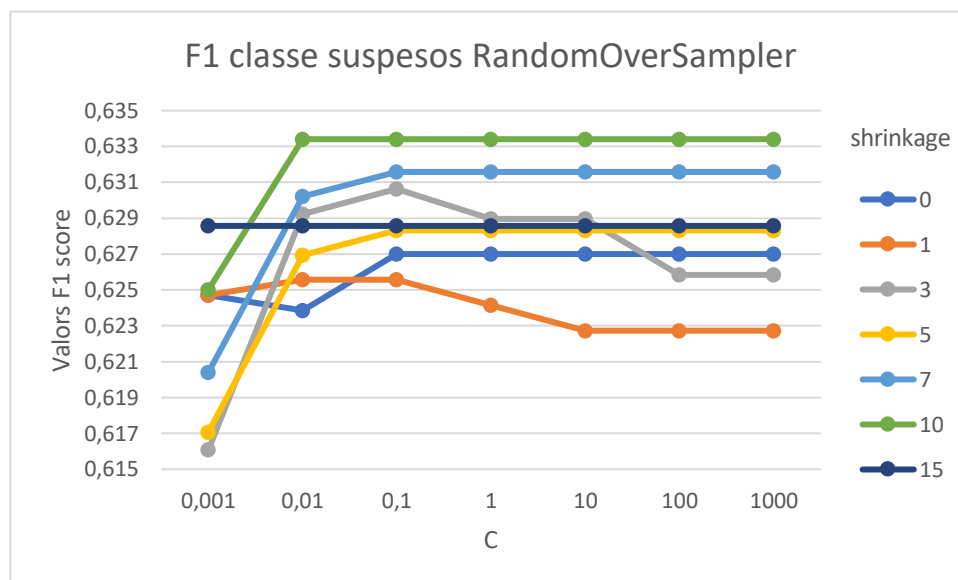


Gràfic 1: Valors $F1$ score per a la classe suspesos per Electromagnetisme sense tècniques de mostreig

Tal i com podem veure al Gràfic 1, els valors de $F1$ generalment augmenten quan augmentem el paràmetre C de la regressió logística. No obstant, a partir de C igual a 1 es produeix un estancament. El màxim valor de $F1$ s'obté quan C és igual a 1. A la Taula 5 es mostren les mètriques *Precision*, *Recall* i $F1$ tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	$F1$
Aprovats	0,75	0,82	0,78
Suspesos	0,60	0,50	0,54

Taula 5: Mètriques per Electromagnetisme sense tècniques de mostreig

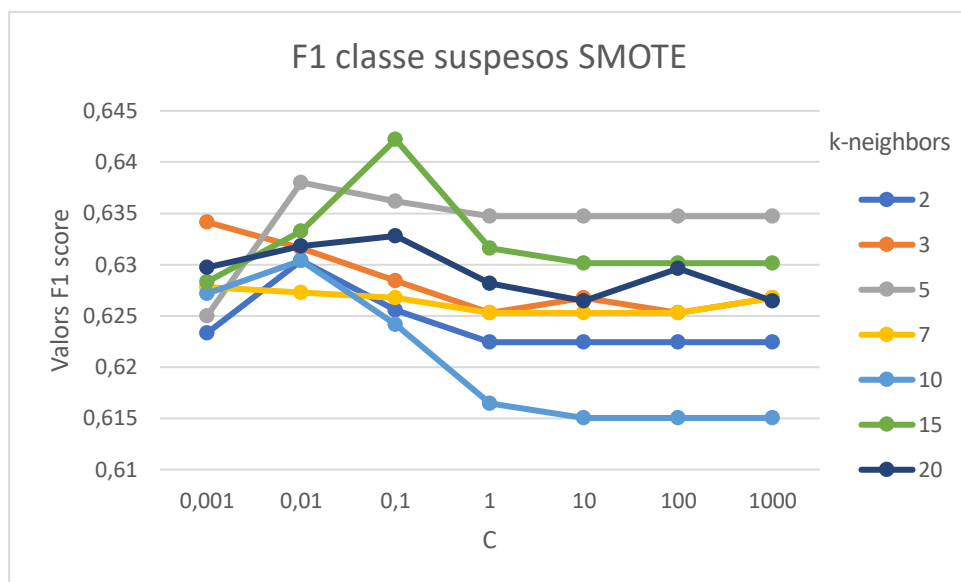
RandomOverSampler

Gràfic 2: Valors *F1 score* per a la classe suspesos per Electromagnetisme amb *RandomOverSampler*

El Gràfic 2 ens mostra que per als diferents valors de shrinkage, el paràmetre de la tècnica de mostreig *RandomOverSampler*, s'obté una tendència molt similar i és que els valors de *F1* augmenten considerablement quan *C* és igual a 0,01 i a partir d'aquí es mantenen constants. En aquest cas trobem que s'obté el màxim valor de *F1* per a la classe suspesos quan el paràmetre shrinkage pren el valor 10. A la Taula 6 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,84	0,62	0,71
Suspesos	0,53	0,79	0,63

Taula 6: Mètriques per Electromagnetisme amb *RandomOverSampler*

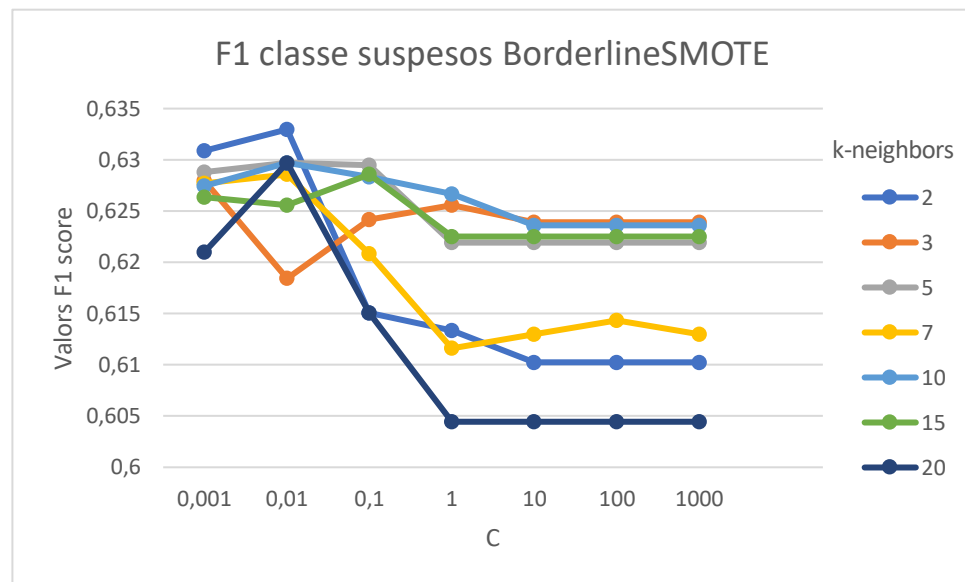
SMOTE

Gràfic 3: Valors F1 score per a la classe suspesos per Electromagnetisme amb SMOTE

En el Gràfic 3 podem observar com els valors de *F1* segueixen un patró semblant per al diferent nombre de veïns de la tècnica de mostreig *SMOTE*. Es veuen diferents pics en els valors de *C* de 0,01 i 0,1 obtenint el valor de *F1* més elevat amb *C* igual a 0,1 i 15 veïns. No obstant, s'obté un resultat pràcticament igual amb un model més senzill, 5 veïns i una *C* de 0,01. A la Taula 7 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,84	0,66	0,74
Suspesos	0,55	0,76	0,64

Taula 7: Mètriques per Electromagnetisme amb SMOTE

BorderlineSMOTE

Gràfic 4: Valors *F1 score* per a la classe suspesos per Electromagnetisme amb BorderlineSMOTE

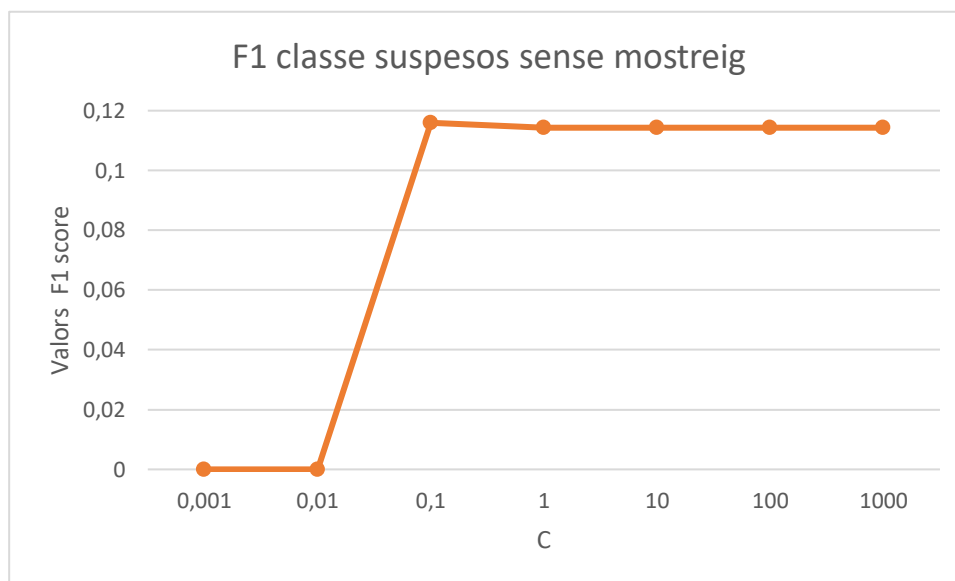
Tal i com podem veure al Gràfic 4, els valors de *F1* disminueixen a mida que augmenta *C*. Aquesta tendència es veu per a gairebé tots els valors de nombre de veïns per a la tècnica de mostreig *BorderlineSMOTE*. El millor valor de *F1* s'obté per a 2 veïns i *C* igual a 0,01. A la Taula 8 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,84	0,63	0,72
Suspesos	0,53	0,78	0,63

Taula 8: Mètriques per Electromagnetisme amb BorderlineSMOTE

7.1.2. Mètodes numèrics

Sense tècniques de mostreig

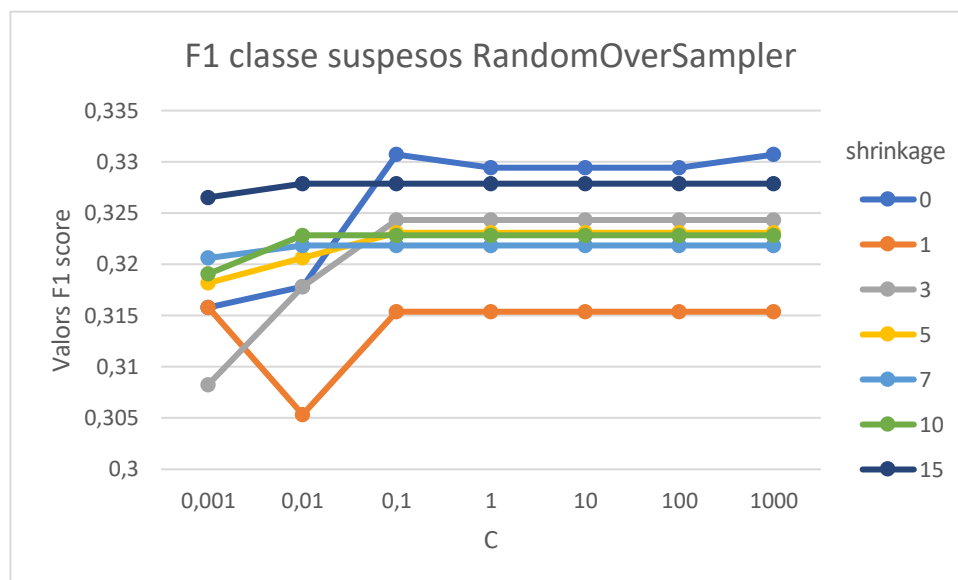


Gràfic 5: Valors *F1* score per a la classe suspesos per Mètodes Numèrics sense tècniques de mostreig

Els resultats del *F1* per a l'assignatura de mètodes numèrics sense tècniques de mostreig per a la classe suspesos són molt baixos en comparació amb la resta. Aquest fet és degut a que aquesta assignatura té un gran nombre d'aprovatats en comparació amb el de suspesos, és a dir, és una de les més desequilibrades. Per als valors de *C* 0,001 i 0,01 el model no és capaç de predir cap suspès correctament. Quan *C* augmenta el model prediu correctament algun suspens però tal i com podem veure al Gràfic 5 els valors de *F1* són molt baixos. El valor més elevat de *F1* correspon a *C* igual a 0,1. A la Taula 9 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovatats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,89	0,99	0,94
Suspesos	0,57	0,06	0,12

Taula 9: Mètriques per Mètodes Numèrics sense tècniques de mostreig

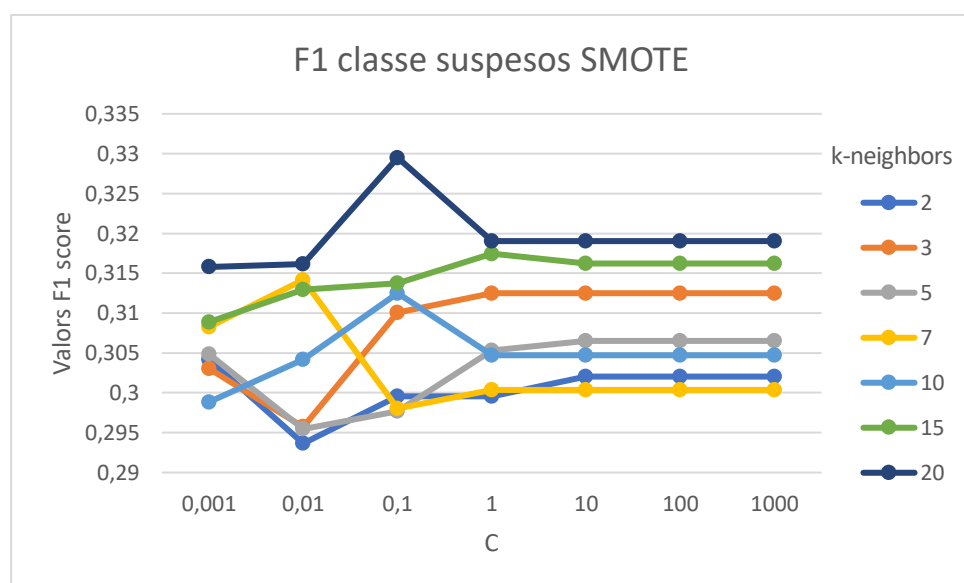
RandomOverSampler

Gràfic 6: Valors F1 score per a la classe suspesos per Mètodes Numèrics amb RandomOverSampler

En el Gràfic 6 s'observa una tendència constant per a tots els valors de shrinkage on el valor de $F1$ es manté gairebé constant a excepció del 0 i el 3 on $F1$ augmenta i a partir de C igual a 0,1 es manté constant i de 1 on el valor de $F1$ cau en C igual a 0,01 i després torna a recuperar. El valor més elevat de $F1$ s'aconsegueix amb shrinkage igual a 0 i C igual a 0,1 i 1000. A la Taula 10 es mostren les mètriques *Precision*, *Recall* i $F1$ tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	$F1$
Aprovats	0,94	0,68	0,79
Suspesos	0,22	0,68	0,33

Taula 10: Mètriques per Mètodes Numèrics amb RandomOverSampler

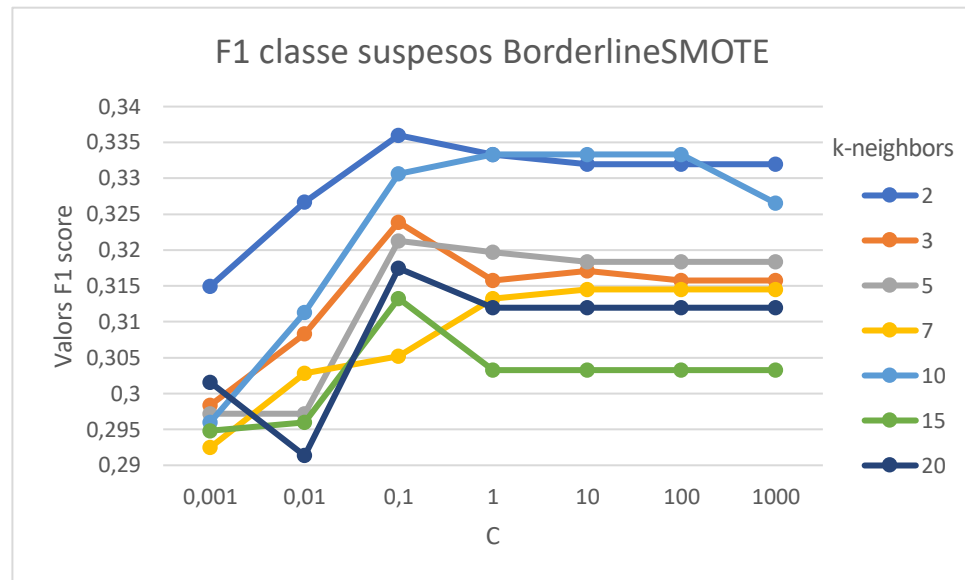
SMOTE

Gràfic 7: Valors F1 score per a la classe suspesos per Mètodes Numèrics amb SMOTE

En aquest cas, tal i com veiem al Gràfic 7, els valors de *F1* es mantenen relativament constants per als diferents valors de *C*. Cal destacar que el nombre de veïns que té valors de *F1* més elevats és 20 especialment quan *C* val 0,1. A la Taula 11 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,94	0,66	0,78
Suspesos	0,22	0,69	0,33

Taula 11: Mètriques per Mètodes Numèrics amb SMOTE

BorderlineSMOTE

Gràfic 8: Valors F1 score per a la classe suspesos per Mètodes Numèrics amb BorderlineSMOTE

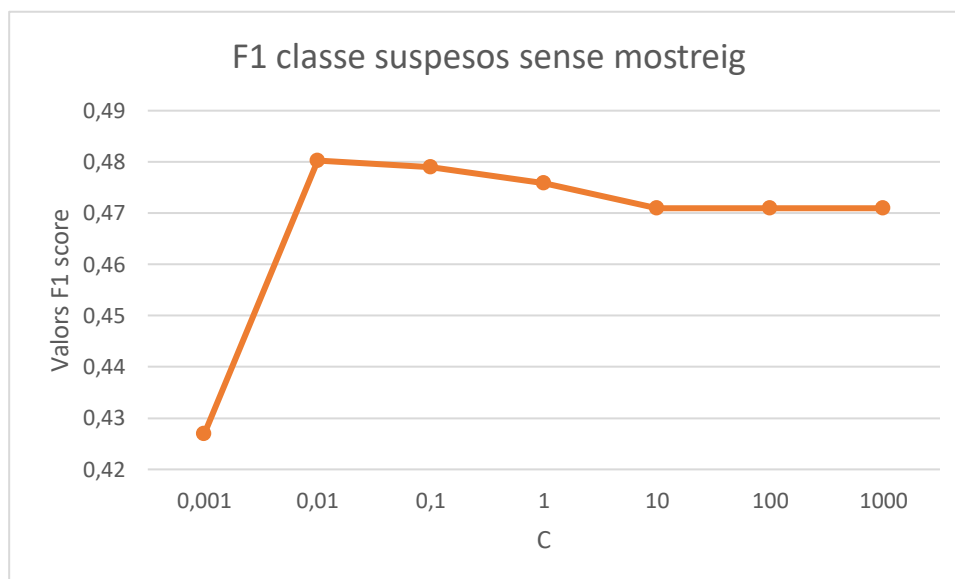
En el Gràfic 8 es torna a veure una mateixa tendència per als diferents valors de nombre de veïns. Destaquen els pics quan C val 0,1 i la posterior estabilització a partir de C igual a 1. El millor valor de *F1* s'obté amb 2 veïns i C igual a 0,1. A la Taula 12 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,94	0,68	0,79
Suspesos	0,22	0,68	0,34

Taula 12: Mètriques per Mètodes Numèrics amb BorderlineSMOTE

7.1.3. Materials

Sense tècniques de mostreig

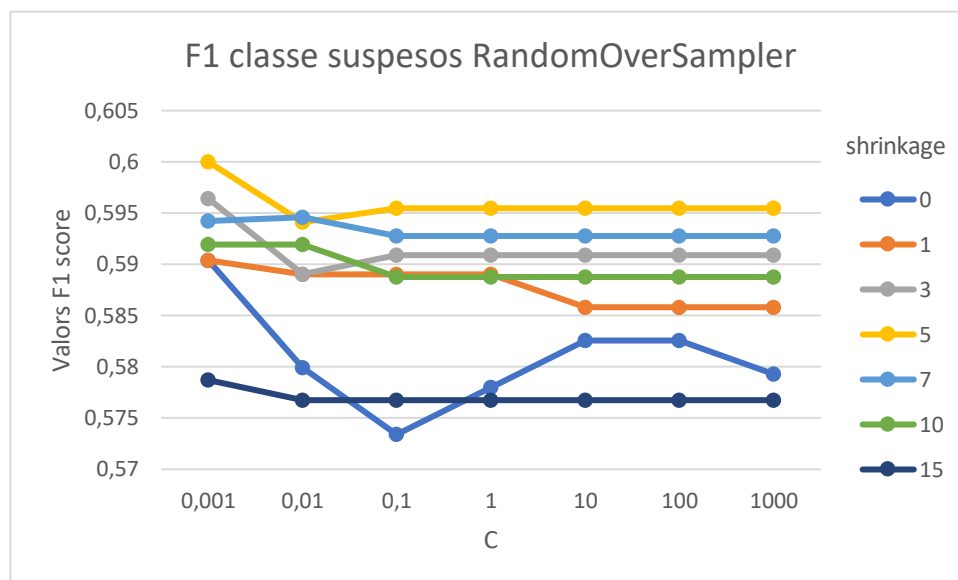


Gràfic 9: Valors *F1* score per a la classe suspesos per Materials sense tècniques de mostreig

En el cas de l'assignatura de materials sense tècniques de mostreig podem observar un augment molt brusc entre 0,001 i 0,01 i després un petit descens fins que els valors de *F1* acaben mantenint-se constants. El Gràfic 9 mostra que el valor més elevat de *F1* s'obté per a *C* igual 0,01. A la Taula 13 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,73	0,86	0,79
Suspesos	0,60	0,40	0,48

Taula 13: Mètriques per Materials sense tècniques de mostreig

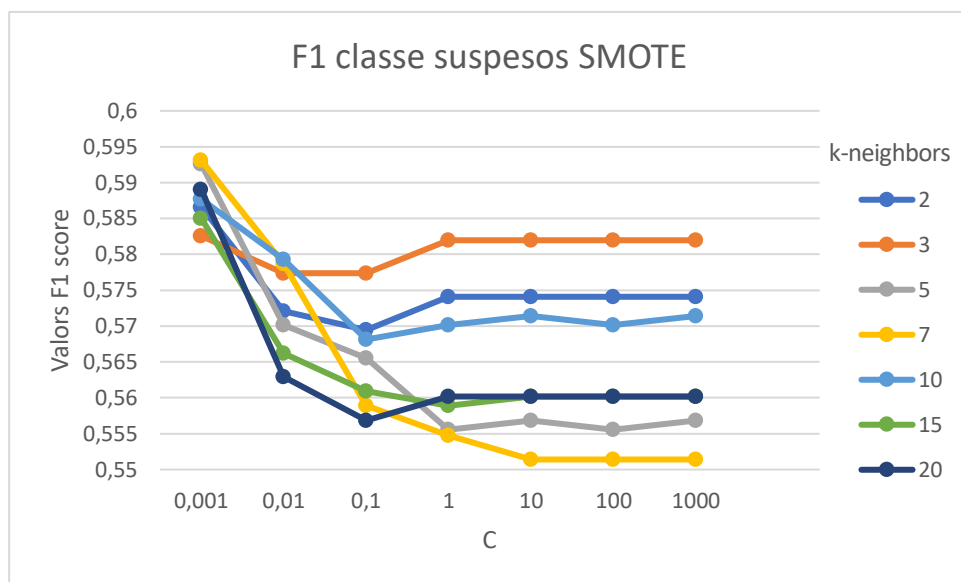
RandomOverSampler

Gràfic 10: Valors F1 score per a la classe suspesos per Materials amb RandomOverSampler

El gràfic 10 ens mostra els resultats amb la tècnica *RandomOverSampler*. S'observa un patró constant per als diferents valors de shrinkage on els valors de *F1* es mantenen relativament constants a excepció del 0 que varia. Cal destacar que s'obtenen millors resultats amb un shrinkage intermig com pot ser 5 i pitjors amb els extrems 0 i 15. El valor de *F1* més elevat s'obté amb shrinkage 5 i C 0,001. A la Taula 14 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,82	0,61	0,70
Suspesos	0,50	0,74	0,6

Taula 14: Mètriques per Materials amb RandomOverSampler

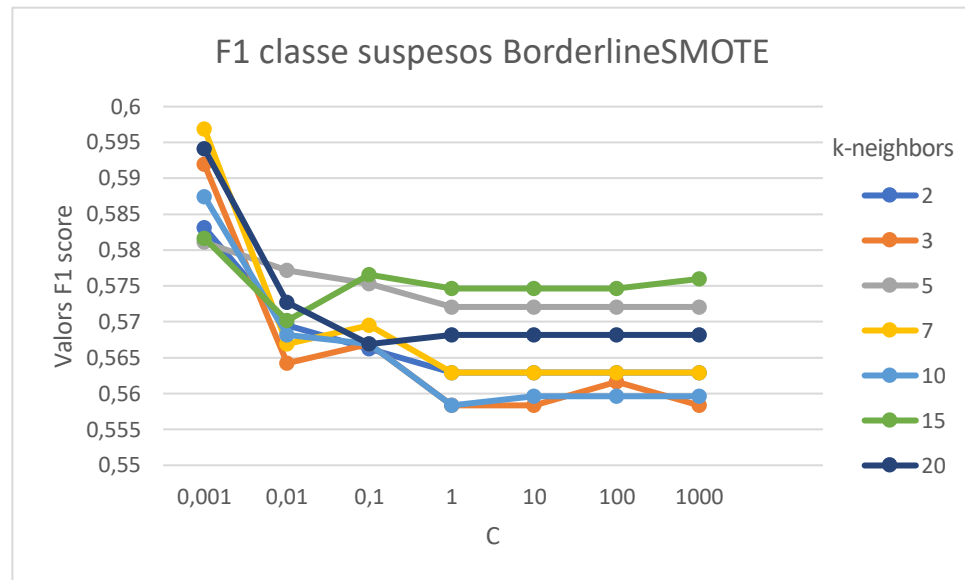
SMOTE

Gràfic 11: Valors F1 score per a la classe suspesos per Materials amb SMOTE

Tal i com podem veure al Gràfic 11 que mostra els resultats amb la tècnica *SMOTE* els valors de *F1* disminueixen en augmentar *C* fins a estabilitzar-se a partir de *C* igual a 10. Aquest patró es segueix per al diferent nombre de veïns. El millor resultat s'obté amb 5 veïns i una *C* de 0,001. A la Taula 15 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,80	0,64	0,71
Suspesos	0,51	0,70	0,59

Taula 15: Mètriques per Materials amb SMOTE

BorderlineSMOTE

Gràfic 12: Valors F1 score per a la classe suspesos per Materials amb BorderlineSMOTE

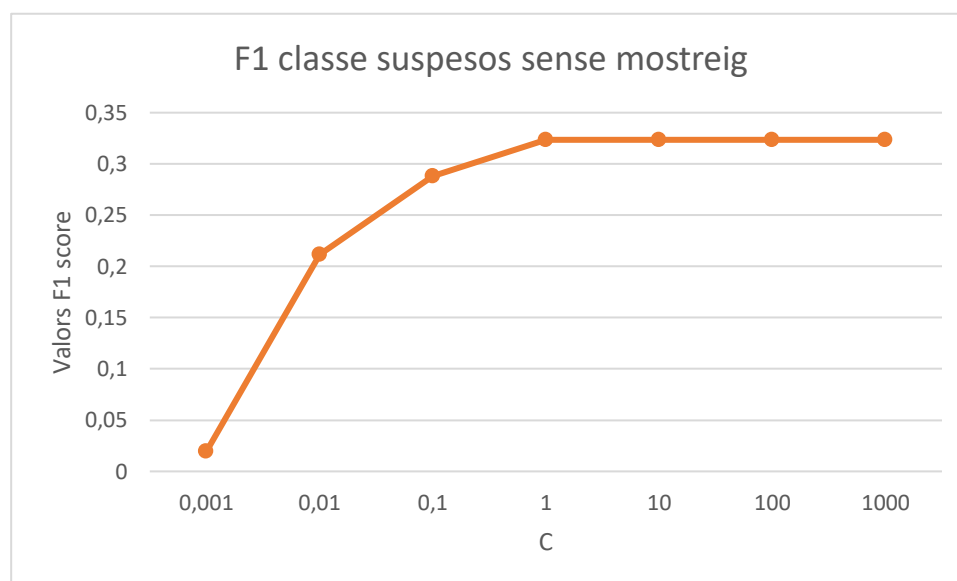
En el Gràfic 12 s'observen els valors de *F1* amb la tècnica *BorderlineSMOTE*. Es veu la mateixa tendència que amb el *SMOTE* on els valors van disminuint a mida que augmenta *C* però, en aquest cas, s'estabilitzen en *C* igual a 1. El millor resultat s'obté amb 7 veïns i una *C* de 0,001 però amb 3 veïns s'aconsegueix un pràcticament igual sent aquest un model més senzill. A la Taula 16 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,81	0,61	0,70
Suspesos	0,50	0,73	0,59

Taula 16: Mètriques per Materials amb BorderlineSMOTE

7.1.4. Equacions diferencials

Sense tècniques de mostreig

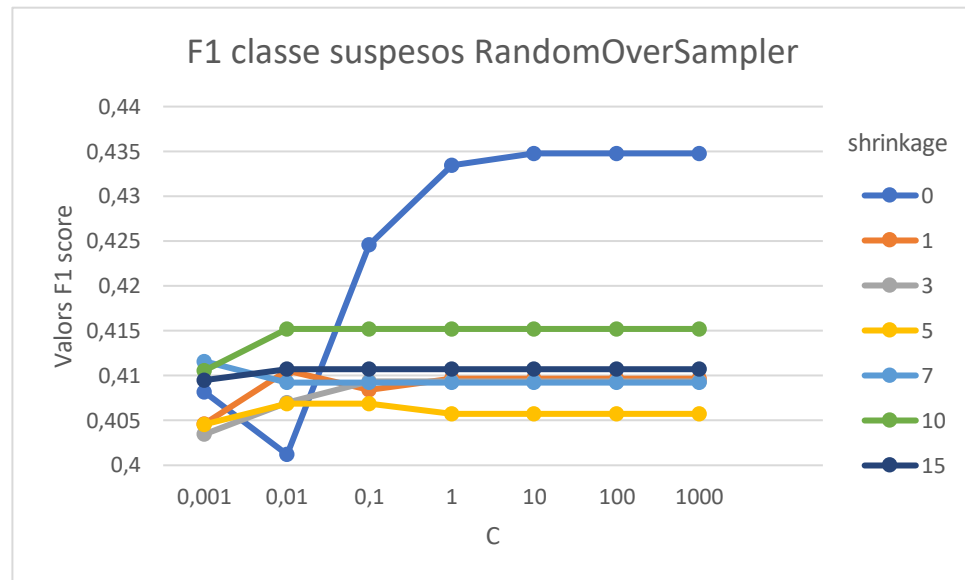


Gràfic 13: Valors *F1* score per a la classe suspesos per Equacions Diferencials sense tècniques de mostreig

En el cas de l'assignatura equacions diferencials, els resultats de *F1* sense tècniques de mostreig, que podem veure al Gràfic 13, augmenten progressivament a mida que augmenta *C* fins estabilitzar-se en el punt més alt quan *C* val 1. A la Taula 17 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,84	0,97	0,90
Suspesos	0,63	0,22	0,32

Taula 17: Mètriques per Equacions Diferencials sense tècniques de mostreig

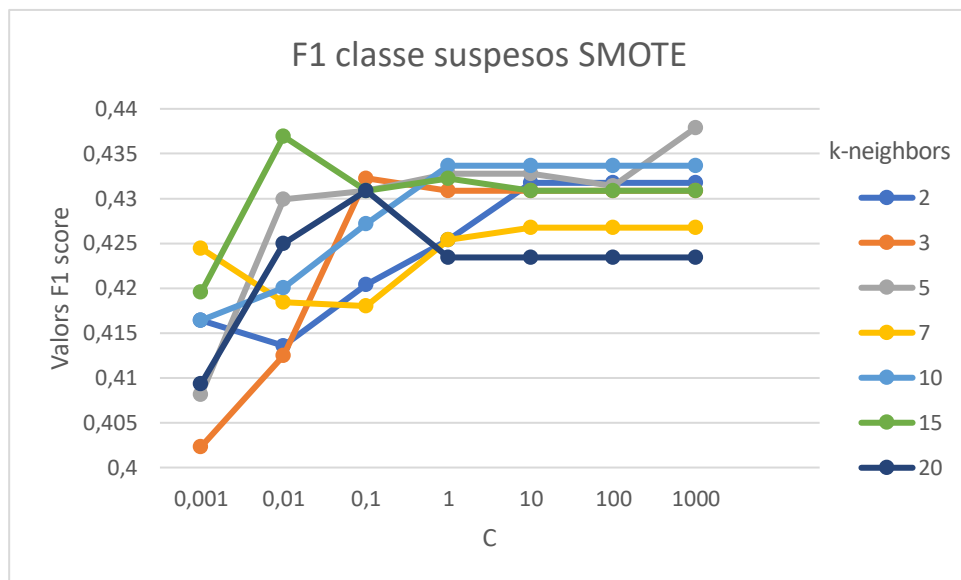
RandomOverSampler

Gràfic 14: Valors *F1 score* per a la classe suspesos per Equacions Diferencials amb *RandomOverSampler*

Tal i com podem veure al Gràfic 14 que mostra els resultats amb la tècnica *RandomOverSampler*, per a la majoria de valors de shrinkage, la mètrica *F1* es manté constant per als diferents valors de *C*. L'únic valor de shrinkage que no segueix el patró és el 0 que veiem que augmenta considerablement a mida que augmenta *C* fins estabilitzar-se quan *C* val 10 i destaca respecte als altres. A la Taula 18 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,90	0,64	0,75
Suspesos	0,32	0,69	0,43

Taula 18: Mètriques per Equacions Diferencials amb *RandomOverSampler*

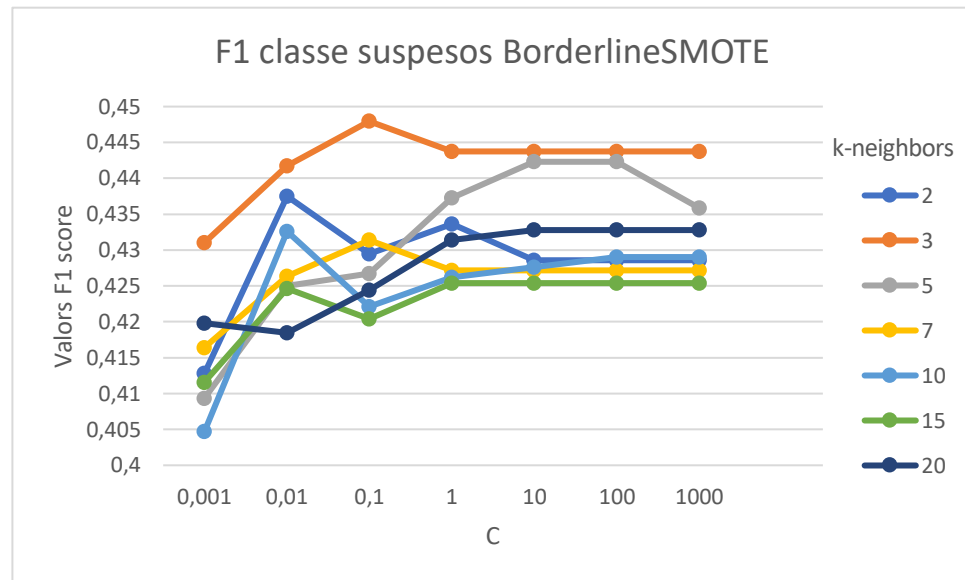
SMOTE

Gràfic 15: Valors *F1* score per a la classe suspesos per Equacions Diferencials amb SMOTE

En el Gràfic 15 es mostren els resultats de *F1* per a la tècnica SMOTE. En aquest cas s'observa una mateixa tendència per a la majoria de valors de nombre de veïns on *F1* va augmentant a mida que augmenta *C* fins establitzar-se a partir de *C* igual a 1. Destaca la configuració amb 15 veïns i *C* igual a 0,01 amb la que s'obté el millor resultat de *F1*. A la Taula 19 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,90	0,64	0,75
Suspesos	0,32	0,70	0,44

Taula 19: Mètriques per Equacions Diferencials amb SMOTE

BorderlineSMOTE

Gràfic 16: Valors F1 score per a la classe suspesos per Equacions Diferencials amb BorderlineSMOTE

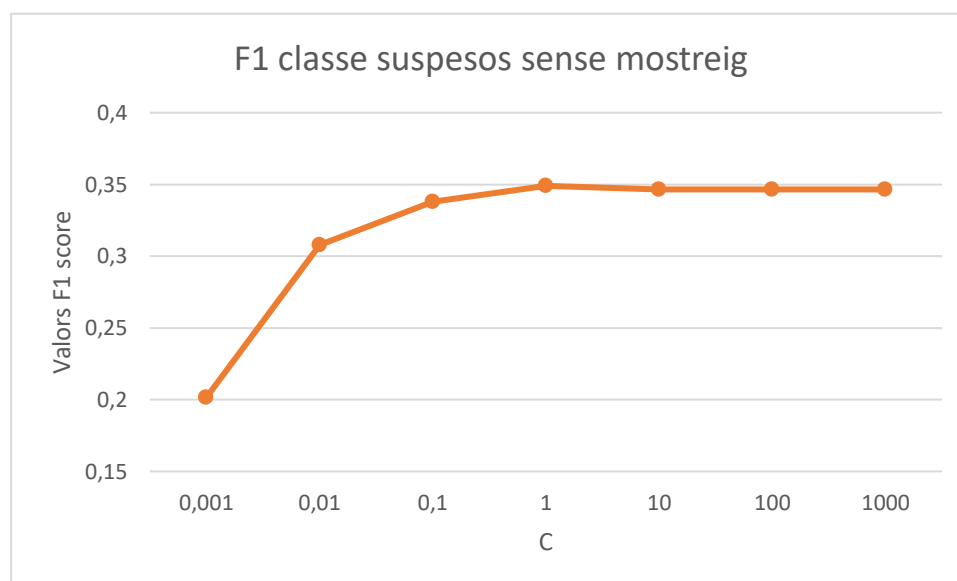
Els resultats amb la tècnica *BorderlineSMOTE* es mostren al Gràfic 16. Seguim veient la mateixa tendència on generalment *F1* augmenta quan augmenta *C*. En aquest cas destaca *C* igual a 0,1 amb 3 veïns on obtenim el major valor de *F1*. A la Taula 20 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,90	0,66	0,76
Suspesos	0,33	0,70	0,45

Taula 20: Mètriques per Equacions Diferencials amb BorderlineSMOTE

7.1.5. Informàtica

Sense tècniques de mostreig

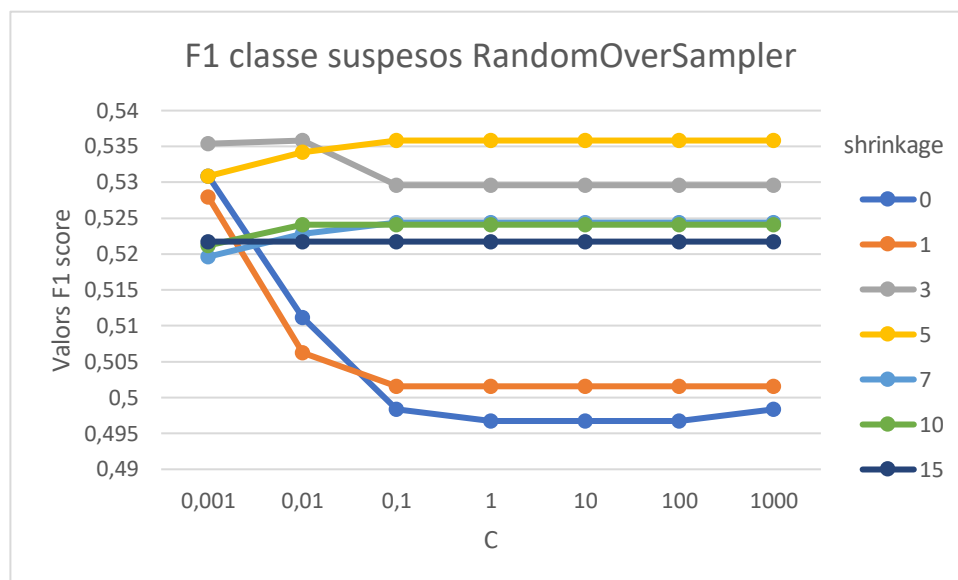


Gràfic 17: Valors *F1* score per a la classe suspesos per Informàtica sense tècniques de mostreig

En el cas de l'assignatura d'informàtica, els resultats de *F1* sense tècniques de mostreig, que podem veure al Gràfic 17, augmenten progressivament a mida que augmenta *C* fins establir-se en el punt més alt quan *C* val 1. A la Taula 21 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,82	0,98	0,89
Suspesos	0,76	0,23	0,35

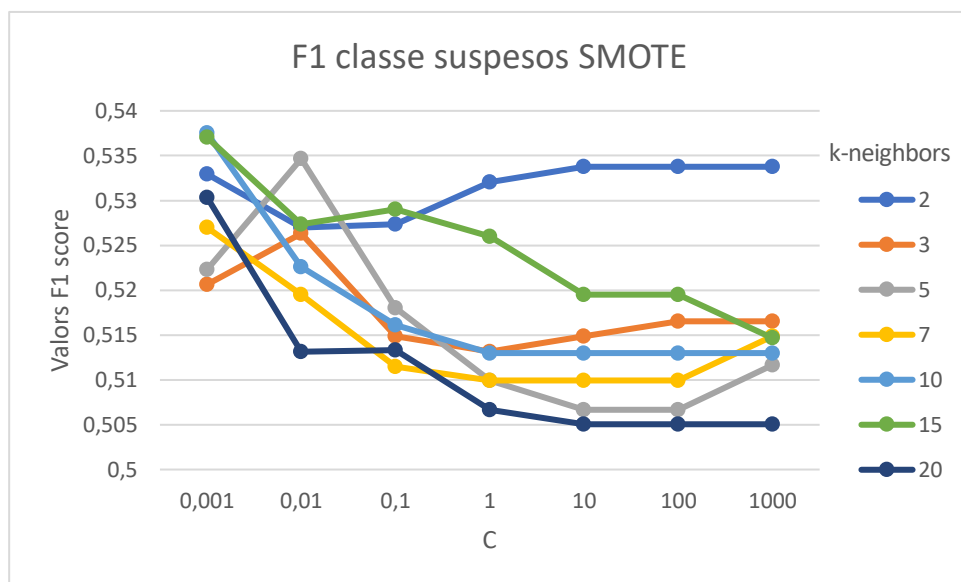
Taula 21: Mètriques per Informàtica sense tècniques de mostreig

RandomOverSamplerGràfic 18: Valors *F1 score* per a la classe suspesos per Informàtica amb *RandomOverSampler*

Tal i com podem veure al Gràfic 18, que mostra els resultats amb la tècnica *RandomOverSampler*, per a la majoria de valors de shrinkage, la mètrica *F1* es manté pràcticament constant per als diferents valors de *C* a excepció dels valors 0 i 1 on la mètrica *F1* disminueix notablement entre els valors de *C* 0,001 i 0,1. El millor resultat de *F1* es troba quan shrinkage val 3 i *C* 0,01. A la Taula 22 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,91	0,71	0,80
Suspesos	0,42	0,75	0,54

Taula 22: Mètriques per Informàtica amb *RandomOverSampler*

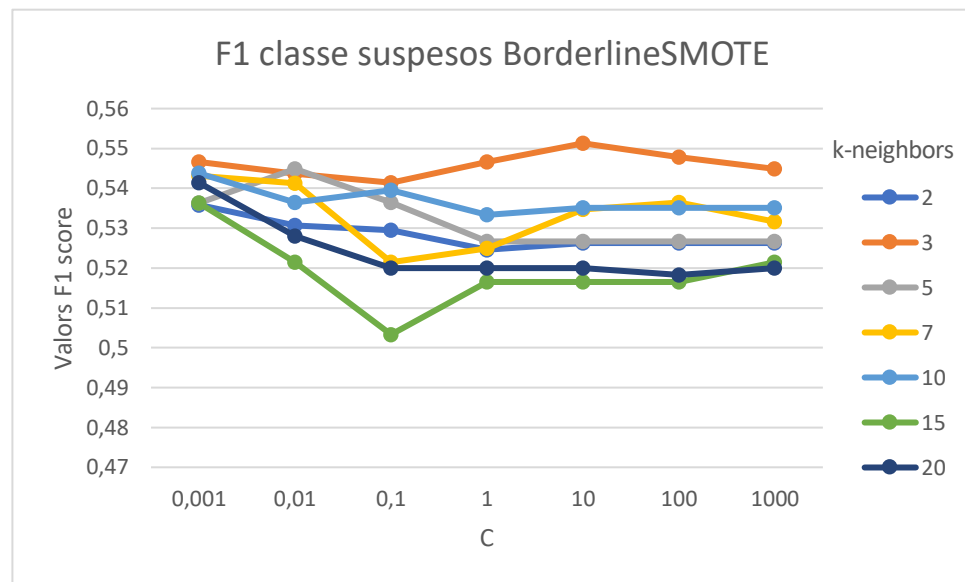
SMOTE

Gràfic 19: Valors *F1* score per a la classe suspesos per Informàtica amb SMOTE

En el Gràfic 19 es mostren els resultats de *F1* per a la tècnica SMOTE. En aquest cas es veu un mateix patró per al diferent nombre de veïns on *F1* va disminuint a mida que augmenta C menys quan el nombre de veïns és 2. Per a aquest valor s'observa un augment de *F1* a partir de C igual a 0,1. El valor de *F1* més elevat s'obté quan C val 0,001 i el nombre de veïns és 10. A la Taula 23 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,91	0,71	0,80
Suspesos	0,42	0,75	0,54

Taula 23: Mètriques per Informàtica amb SMOTE

BorderlineSMOTEGràfic 20: Valors *F1* score per a la classe suspesos per Informàtica amb BorderlineSMOTE

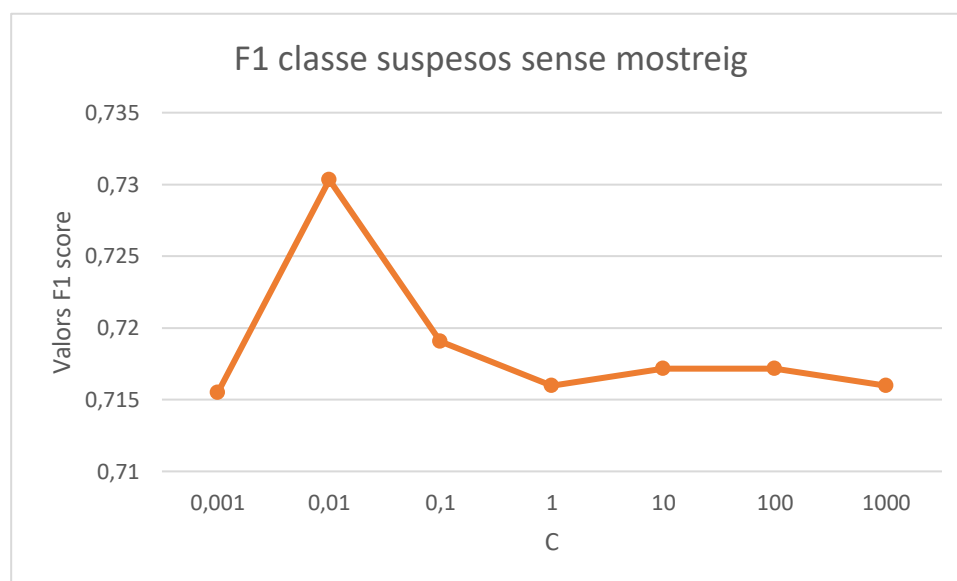
Els resultats de la tècnica *BorderlineSMOTE* es mostren al Gràfic 20. S'observa certa constància en la majoria de casos de nombre de veïns amb algun petit augment o disminució quan varia *C*. La tècnica amb 3 veïns és la que ens proporciona millors resultats de *F1* destacant *C* igual a 10 on trobem el valor més alt. No obstant, amb una *C* de 0,001, és a dir, amb un model més senzill s'obté un resultat pràcticament igual. A la Taula 24 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,91	0,71	0,80
Suspesos	0,43	0,77	0,55

Taula 24: Mètriques per Informàtica amb BorderlineSMOTE

7.1.6. Mecànica

Sense tècniques de mostreig

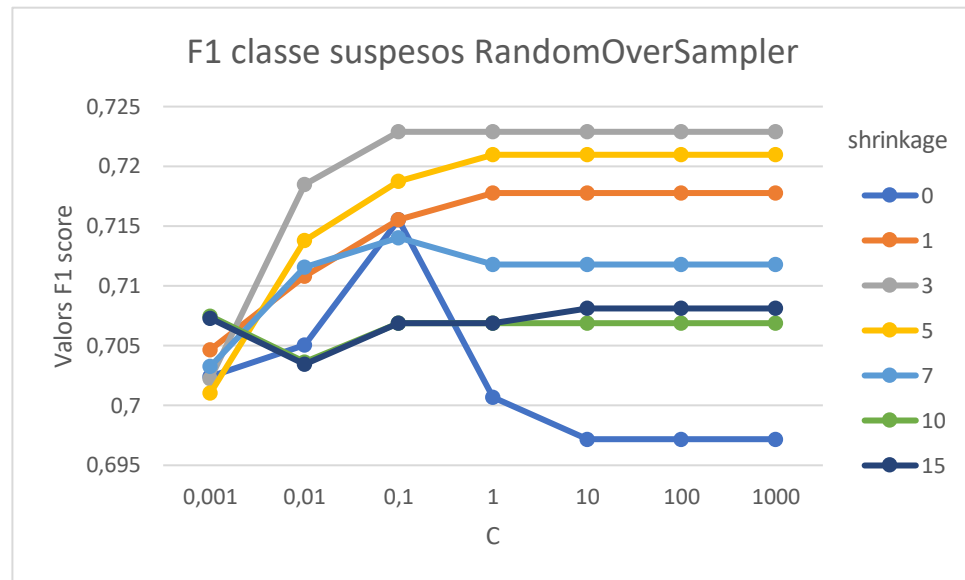


Gràfic 21: Valors F1 score per a la classe suspesos per Mecànica sense tècniques de mostreig

En el cas de l'assignatura de mecànica, els resultats de *F1* sense tècniques de mostreig que es mostren al Gràfic 21 són constants pels diferents valors de *C* a excepció de *C* igual a 0,01 on trobem un pic que ens proporciona el millor resultat de *F1*. A la Taula 25 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,72	0,58	0,64
Suspesos	0,68	0,79	0,73

Taula 25: Mètriques per Mecànica sense tècniques de mostreig

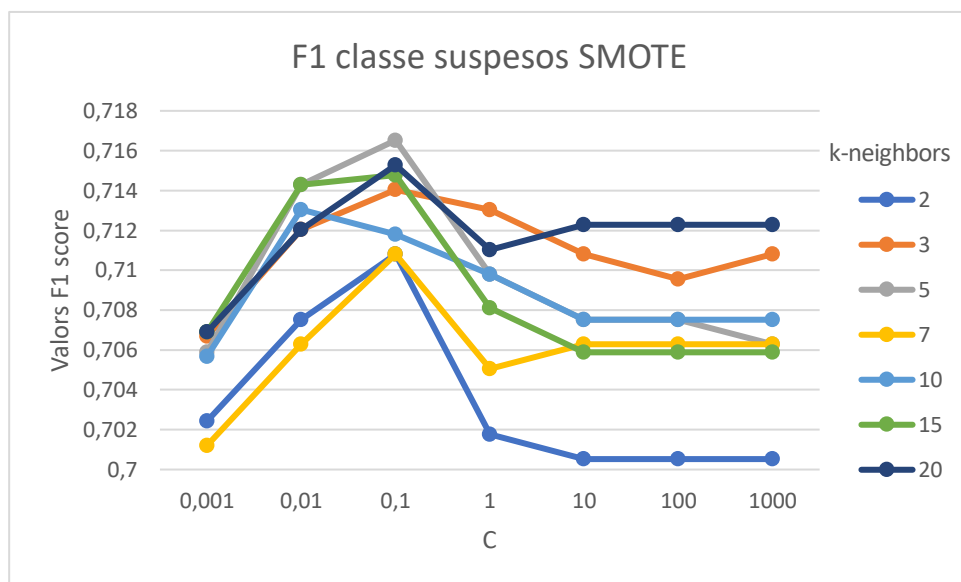
RandomOverSampler

Gràfic 22: Valors F1 score per a la classe suspesos per Mecànica amb RandomOverSampler

En el Gràfic 22 es poden observar els resultats de *F1* per a la tècnica *RandomOverSampler*. En aquest cas tornem a veure la tendència, en la majoria de casos de shrinkage, d'augmentar a mida que augmenta *C* estabilitzant-se quan *C* val 1. Destaca el cas de shrinkage 3 que ens proporciona els millors resultats. A la Taula 26 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,70	0,61	0,66
Suspesos	0,69	0,76	0,72

Taula 26: Mètriques per Mecànica amb RandomOverSampler

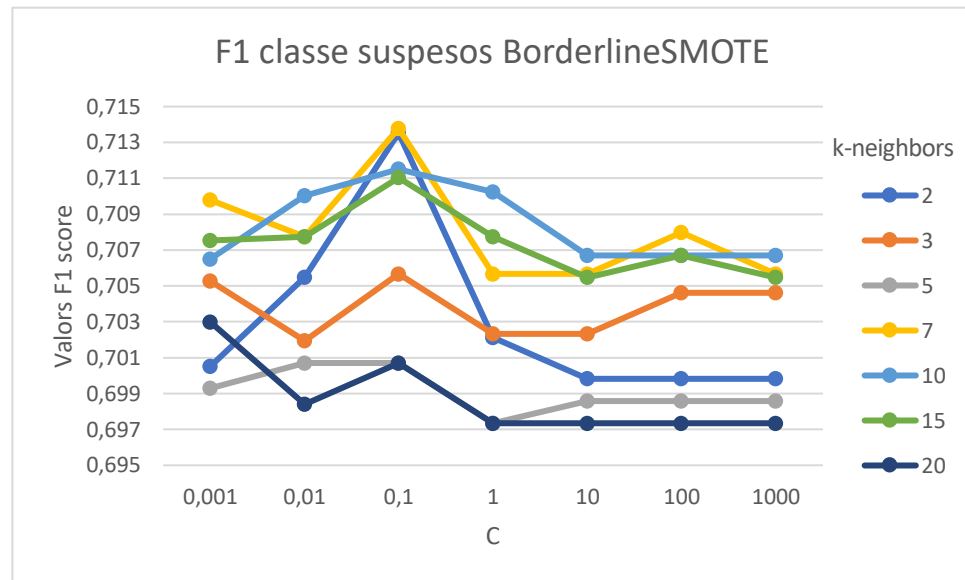
SMOTE

Gràfic 23: Valors F1 score per a la classe suspesos per Mecànica amb SMOTE

Tal i com podem veure al Gràfic 23, que mostra els resultats de *F1* per a la tècnica *SMOTE*, per a tots els valors de nombre de veïns es veu un pic en *C* igual a 0,1. La tècnica de mostreig amb 3 veïns és la menys complexa que ens proporciona millors resultats quan el paràmetre *C* val 0,1. A la Taula 27 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,69	0,61	0,65
Suspesos	0,68	0,75	0,71

Taula 27: Mètriques per Mecànica amb SMOTE

BorderlineSMOTE

Gràfic 24: Valors F1 score per a la classe suspesos per Mecànica amb BorderlineSMOTE

Els resultats de *F1* amb la tècnica *BorderlineSMOTE* es mostren al Gràfic 24. Es pot observar la mateixa tendència que en el cas anterior on es veuen pics per a *C* igual a 0,1. En aquest cas obtenim els millors resultats quan el nombre de veïns és 2. A la Taula 28 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,69	0,63	0,66
Suspesos	0,69	0,74	0,71

Taula 28: Mètriques per Mecànica amb BorderlineSMOTE

7.1.7. Resum resultats regressió logística

Per poder tenir una millor visió de tot el conjunt de resultats obtingut amb el model de regressió logística s'ha construït una taula que engloba els millors valors de $F1$ per a la classe suspesos i els respectius valors de *Precision* i *Recall*. També s'han afegit els resultats de la classe aprovats ja que, malgrat no ser l'objectiu principal del projecte, no es poden deixar de banda. Aquesta recopilació de resultats es mostra a la Taula 29. D'altra banda s'ha construït la Taula 30 on podem veure la variació de $F1$ quan apliquem la tècnica de mostreig respecte el valor quan no s'aplica.

En gairebé totes les assignatures es produeix un increment del rendiment de la predicció de suspesos aplicant els mètodes d'*oversampling* a excepció de l'assignatura de mecànica en la que la variació és insignificant. Per tant, arribem a la conclusió que era necessari equilibrar les dades de les dues classes ja que en tenir una gran quantitat d'aprovats al conjunt d'entrenament de dades, el model identificava molts aprovats que no eren correctes. A continuació es comentarà més detingudament com han variat els resultats en aplicar les tècniques de mostreig per a cada assignatura.

En l'assignatura d'electromagnetisme el rendiment de la predicció de suspesos computat amb $F1$ augmenta entre un 16% i un 19% quan apliquem les tècniques de mostreig. Aquesta millor predicció dels suspesos comporta un descens d'entre el 5% i el 9% del rendiment en la predicció d'aprovats.

En el cas de mètodes numèrics es produeix una millora molt més significativa que l'anterior. El rendiment de $F1$ per a la classe suspesos augmenta entre 175% i 183% mentre que el de la classe d'aprovats només disminueix entre un 15% i 17%. Aquesta millora tant significativa en la predicció de suspesos és deguda a que l'assignatura de mètodes numèrics era la més desequilibrada inicialment.

Continuem amb materials on tornem a veure una millora en la predicció de suspesos en aplicar les tècniques de mostreig. Es produeix un augment d'entre un 23% i un 25% en $F1$ dels suspesos. No obstant, també es torna a repetir la pèrdua de rendiment en els aprovats on trobem una disminució entre el 10% i 11,5%.

Per l'assignatura d'equacions diferencials es repeteix aquest augment en el rendiment de l'encert de suspesos incrementant-lo entre un 34% i 40% tenint també un decrement en el d'aprovats d'entre 15% i 17%.

En el cas d'informàtica la millora en la predicció dels suspesos és una mica més significativa trobant-se entre un 54% i un 57%. A més, cal destacar que la disminució del rendiment dels aprovats és menor que en el cas anterior tot i tenir-ne un augment major en els suspesos. Aquests descens és del 10%,

Per últim, trobem el cas de mecànica, l'única assignatura que es trobava inicialment equilibrada. Per aquest motiu, les variacions en el rendiment de la predicció tant dels suspesos com dels aprovats són insignificants sent d'entre un 1% i un 3% en ambdós casos. És important veure que, a mecànica, el rendiment en predir suspesos disminueix. Aquest fet és degut a que tot i que les classes es trobaven pràcticament equilibrades, la quantitat de suspesos superava la d'aprovats. Per aquest motiu, seria necessari considerar no aplicar tècniques de mostreig en aquest cas ja que empitjora els resultats.

No es pot detectar en quins casos funciona millor una tècnica de mostreig o una altra ja que per a cada assignatura obtenim millors resultats amb una tècnica diferent.

És necessari comentar que no es troba cap configuració òptima de valors dels paràmetres. En cada cas s'obté una combinació diferent de paràmetres que ens proporciona un millor resultat. No obstant això, es pot veure com, en la gran majoria de casos, els paràmetres es mantenen en valors baixos. En el cas del paràmetre *shrinkage*, no trobem cap superior a 10. El paràmetre *k-neighbors* també es manté baix però, en aquest cas, si que trobem alguna excepció on el nombre de veïns arriba a 15 o 20. Per últim, en el paràmetre C de la regressió logística s'observa una certa tendència a disminuir quan apliquem les tècniques de mostreig. Això significa que es necessita un model més senzill per arribar a un resultat òptim quan les dues classes estan equilibrades. És a dir, el desequilibri tendeix a fer que s'hagi de construir models més complexes i no necessàriament més predictius.

Com a conclusió podem dir que les tècniques de mostreig ens permeten obtenir un augment significatiu en la predicció de la classe minoritària quan les dades estan desequilibrades. Això comporta una pèrdua de rendiment per a la classe majoritària, però és molt inferior en comparació amb la millora. Per tant, concloem que val la pena el sacrifici sobretot si el nostre objectiu és la predicció dels suspesos.

				Classe aprovats			Classe suspesos				
				C	s/k	Precision	Recall	F1	Precision	Recall	F1
Electro- magnetisme	Sense mostreig	1	-	0,75	0,82	0,78	0,60	0,50	0,54		
	RandomOverSampler	0,01	10	0,84	0,62	0,71	0,53	0,79	0,63		
	SMOTE	0,1	5	0,84	0,66	0,74	0,55	0,76	0,64		
	BorderlineSMOTE	0,01	2	0,84	0,63	0,72	0,53	0,78	0,63		
Mètodes numèrics	Sense mostreig	0,1	-	0,89	0,99	0,94	0,57	0,06	0,12		
	RandomOverSampler	0,1	0	0,94	0,68	0,79	0,22	0,87	0,33		
	SMOTE	0,1	20	0,94	0,66	0,78	0,22	0,69	0,33		
	BorderlineSMOTE	0,1	2	0,94	0,68	0,79	0,22	0,68	0,34		
Materials	Sense mostreig	0,01	-	0,73	0,86	0,79	0,60	0,40	0,48		
	RandomOverSampler	0,001	5	0,82	0,61	0,70	0,50	0,74	0,60		
	SMOTE	0,001	5	0,80	0,64	0,71	0,51	0,70	0,59		
	BorderlineSMOTE	0,001	3	0,81	0,61	0,70	0,50	0,73	0,59		
Equacions diferencials	Sense mostreig	1	-	0,84	0,97	0,90	0,63	0,22	0,32		
	RandomOverSampler	1	0	0,90	0,64	0,75	0,32	0,69	0,43		
	SMOTE	0,01	15	0,90	0,64	0,75	0,32	0,70	0,44		
	BorderlineSMOTE	0,1	3	0,90	0,66	0,76	0,33	0,70	0,45		
Informàtica	Sense mostreig	1	-	0,82	0,98	0,89	0,76	0,23	0,35		
	RandomOverSampler	0,01	3	0,91	0,71	0,80	0,42	0,75	0,54		
	SMOTE	0,001	10	0,91	0,71	0,80	0,42	0,75	0,54		
	BorderlineSMOTE	0,001	3	0,91	0,71	0,80	0,43	0,77	0,55		
Mecànica	Sense mostreig	0,01	-	0,72	0,58	0,64	0,68	0,79	0,73		
	RandomOverSampler	0,1	3	0,70	0,61	0,66	0,69	0,76	0,72		
	SMOTE	0,1	3	0,69	0,61	0,65	0,68	0,75	0,71		
	BorderlineSMOTE	0,1	2	0,69	0,63	0,66	0,69	0,74	0,71		

Taula 29: Mètriques Regressió Logística (on s =shrinkage per RandomOverSampler i k = k -neighbors per SMOTE i BorderlineSMOTE)

		Classe aprovats	Classe suspesos
Electromagnetisme	RandomOverSampler	-8,97	16,67
	SMOTE	-5,13	18,52
	BorderlineSMOTE	-7,69	16,67
Mètodes numèrics	RandomOverSampler	-15,96	175,00
	SMOTE	-17,02	175,00
	BorderlineSMOTE	-15,96	183,33
Materials	RandomOverSampler	-11,39	25,00
	SMOTE	-10,13	22,92
	BorderlineSMOTE	-11,39	22,92
Equacions diferencials	RandomOverSampler	-16,67	34,38
	SMOTE	-16,67	37,50
	BorderlineSMOTE	-15,56	40,63
Informàtica	RandomOverSampler	-10,11	54,29
	SMOTE	-10,11	54,29
	BorderlineSMOTE	-10,11	57,14
Mecànica	RandomOverSampler	3,13	-1,37
	SMOTE	1,56	-2,74
	BorderlineSMOTE	3,13	-2,74

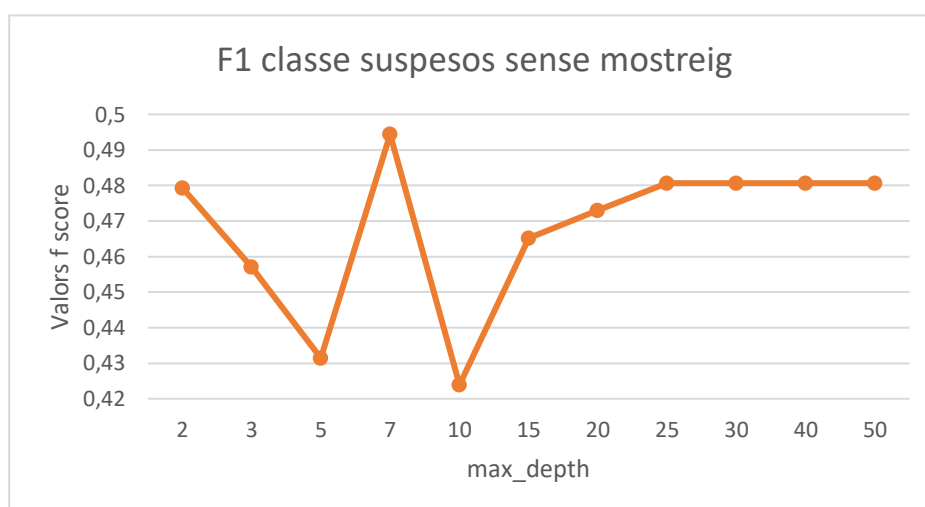
Taula 30: Variacions del valor F1 (en %) amb la tècnica de mostreig respecte al obtingut sense aplicar tècniques de mostreig

7.2. Arbres de decisió

En aquest apartat és durà a terme l'anàlisi amb el model predictiu arbres de decisió. Es mostraran els resultats obtinguts sense tècniques de mostreig i amb les tècniques *RandomOverSampler*, *SMOTE* i *BorderlineSMOTE*. Per a cadascun d'ells es mostrarà gràficament com varia la mètrica *F1* de la classe suspesos en funció dels paràmetres *max_depth* dels arbres de decisió, *shrinkage* de *RandomOverSampler* i *k-neighbors* de *SMOTE* i *BorderlineSMOTE*. Un cop vist com afecten els paràmetres a la predicció de suspesos, es seleccionarà la configuració que ens proporciona un millor rendiment i es mostraran les mètriques *Precision*, *Recall* i *F1* per ambdues classes.

7.2.1. Electromagnetisme

Sense tècniques de mostreig

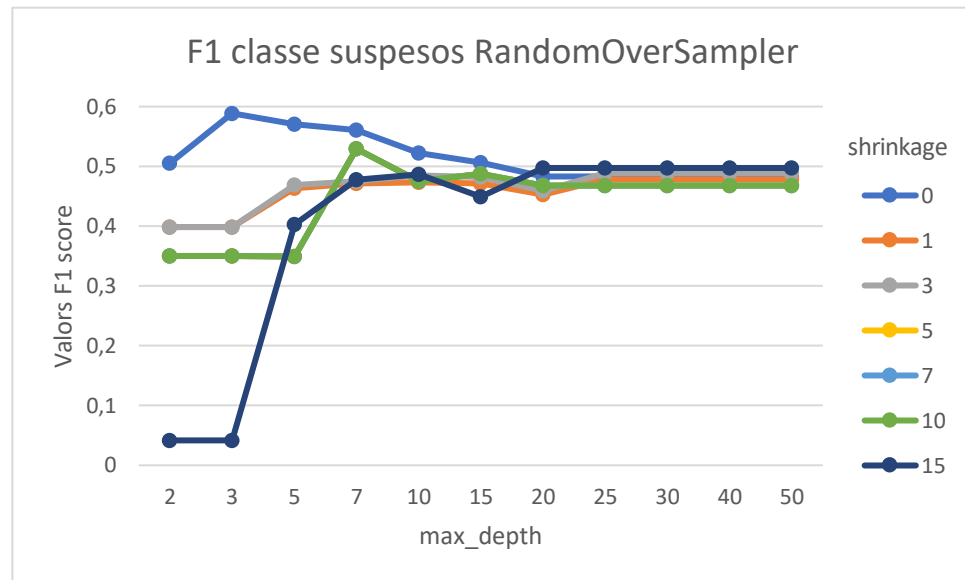


Gràfic 25: Valors *F1* score per a la classe suspesos per Electromagnetisme sense tècniques mostreig

En el cas de l'assignatura d'electromagnetisme sense tècniques de mostreig s'observen variacions per a les diferents profunditats de l'arbre. Com podem veure al Gràfic 25, la millor profunditat d'arbre és 7. A la Taula 31 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,73	0,76	0,74
Suspesos	0,51	0,48	0,49

Taula 31: Mètriques per Electromagnetisme sense tècniques de mostreig

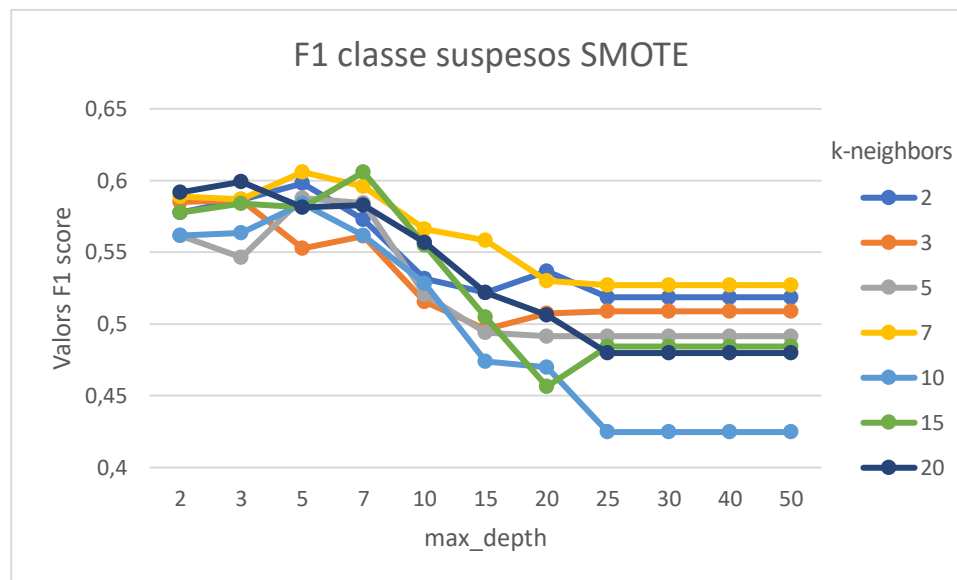
RandomOverSampler

Gràfic 26: Valors *F1* score per a la classe suspesos per Electromagnetisme amb *RandomOverSampler*

En el Gràfic 26 podem veure els resultats de *F1* obtinguts amb la tècnica de mostreig *RandomOverSampler*. Aquest ens mostra certa constància per als diferents valors de shrinkage a partir d'una profunditat d'arbre de 15. Els millors resultats de la mètrica *F1* s'obtenen amb una profunditat d'arbre 3 i un shrinkage de 0. A la Taula 32 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,80	0,58	0,68
Suspesos	0,49	0,74	0,59

Taula 32: Mètriques per Electromagnetisme amb *RandomOverSampler*

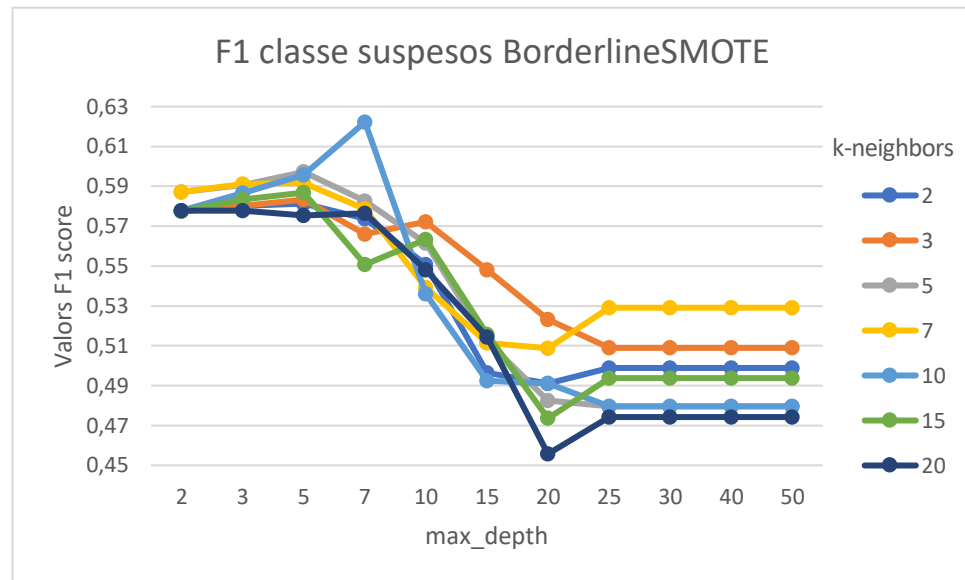
SMOTE

Gràfic 27: Valors *F1 score* per a la classe suspesos per Electromagnetisme amb SMOTE

Tal i com podem observar al Gràfic 27, els valors de *F1* amb la tècnica *SMOTE* disminueixen a mida que augmenta la profunditat d'arbre. Generalment s'obtenen millors resultats amb 7 veïns però cal destacar que amb 7 veïns i una profunditat d'arbre de 5 obtenim el valor de *F1* més elevat. A la Taula 33 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,82	0,58	0,68
Suspesos	0,50	0,77	0,61

Taula 33: Mètriques per Electromagnetisme amb SMOTE

BorderlineSMOTE

Gràfic 28: Valors F1 score per a la classe suspesos per Electromagnetisme amb BorderlineSMOTE

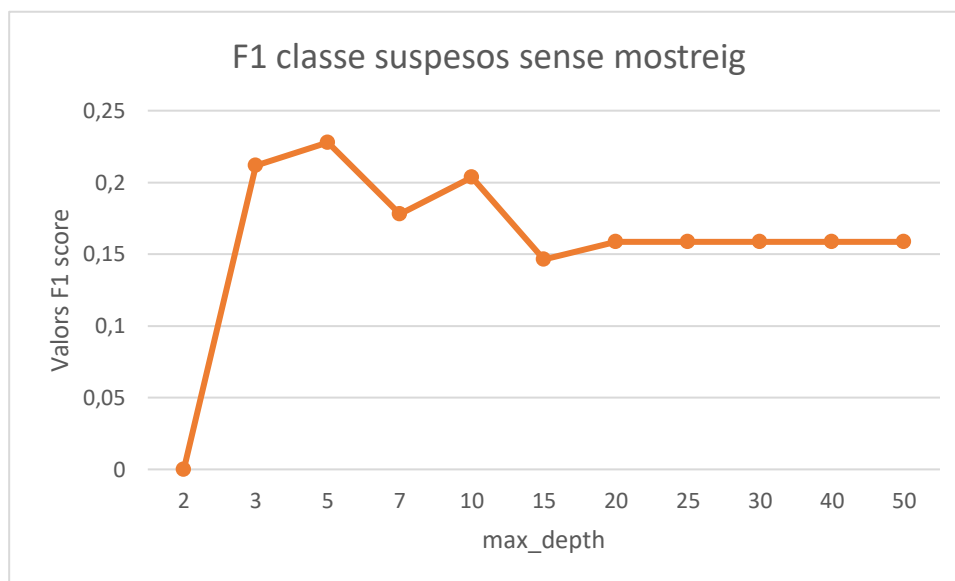
Analitzant el Gràfic 28 veiem que amb la tècnica *BorderlineSMOTE* s'observa la mateixa tendència que en el cas anterior on *F1* va disminuint quan augmenta la profunditat d'arbre. Destaca la configuració de 10 veïns i profunditat d'arbre 7 amb la que obtenim el millor valor de *F1*. A la Taula 34 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,83	0,63	0,72
Suspesos	0,53	0,76	0,62

Taula 34: Mètriques per Electromagnetisme amb BorderlineSMOTE

7.2.2. Mètodes numèrics

Sense tècniques de mostreig

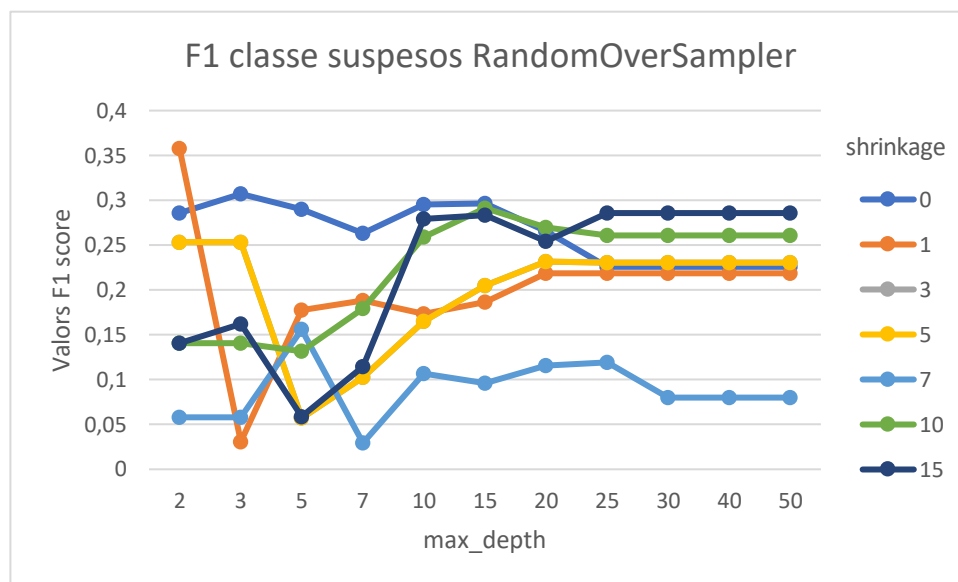


Gràfic 29: Valors *F1 score* per a la classe suspesos per Mètodes Numèrics sense tècniques de mostreig

En el cas de l'assignatura mètodes numèrics sense tècniques de mostreig s'observa que amb profunditat d'arbre 2 el model no és capaç de predir cap suspens correctament. Posteriorment, tal i com es mostra al Gràfic 29, a partir de profunditat 3 els valors de *F1* disminueixen estabilitzant-se a partir de profunditat 20. Els millors valors de *F1* s'obtenen amb una profunditat d'arbre de 5. A la Taula 35 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,90	0,98	0,94
Suspesos	0,53	0,15	0,23

Taula 35: Mètriques per Mètodes Numèrics sense tècniques de mostreig

RandomOverSampler

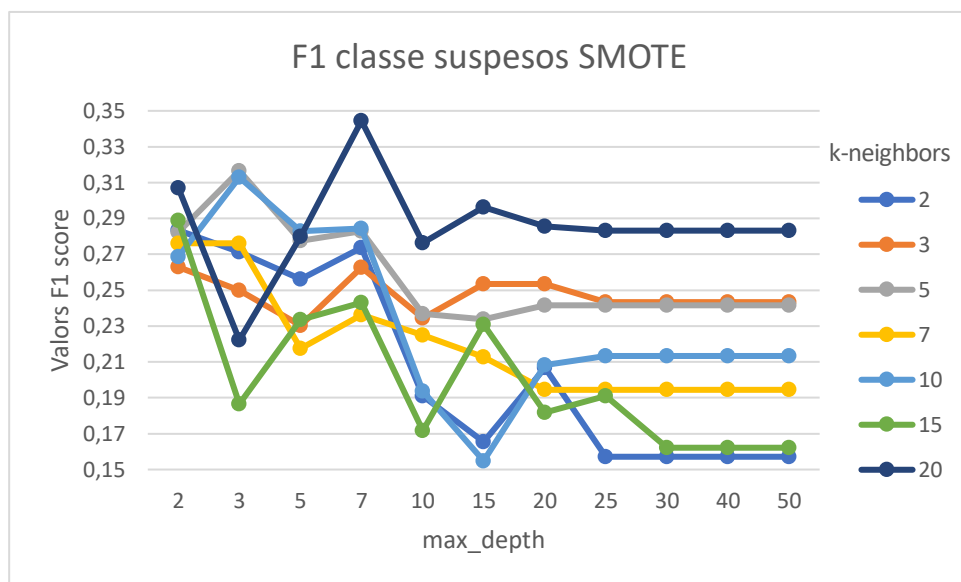
Gràfic 30: Valors F1 score per a la classe suspesos per Mètodes Numèrics amb RandomOverSampler

En el Gràfic 30 que mostra els resultats per a la tècnica *RandomOverSampler*, s'observen tendències diverses pels diferents valors de shrinkage. Cal destacar que per a tots els valors de shrinkage es produeix una estabilització a partir de la profunditat d'arbre 25. La configuració que ens proporciona millors resultats és shrinkage 1 i profunditat d'arbre 2. A la Taula 36 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,92	0,87	0,89
Suspesos	0,30	0,44	0,36

Taula 36: Mètriques per Mètodes Numèrics amb RandomOverSampler

SMOTE

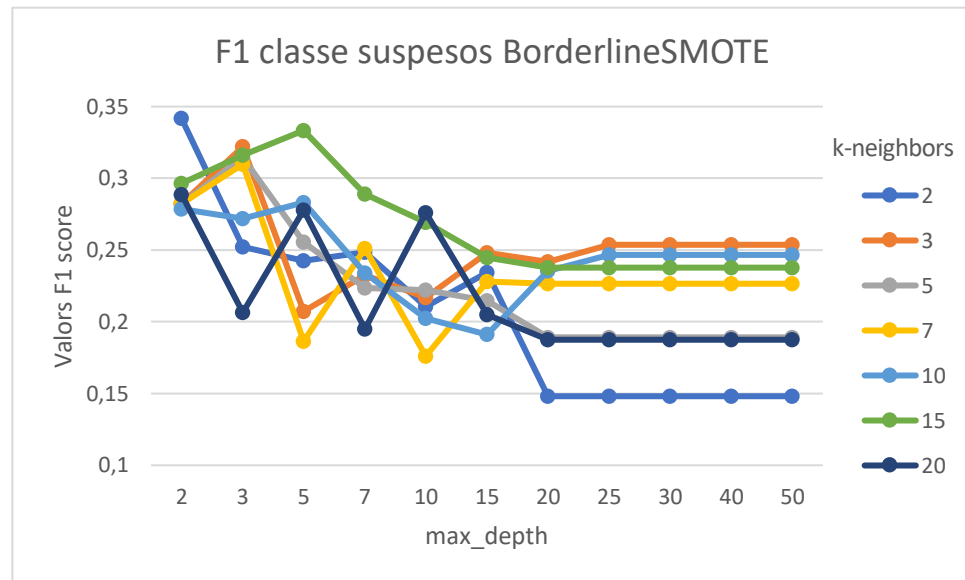


Gràfic 31: Valors *F1* score per a la classe suspesos per Mètodes Numèrics amb SMOTE

Els resultats de *F1* per a la tècnica SMOTE que es mostren al Gràfic 31 són diversos en funció de la profunditat i el nombre de veïns. L'únic patró que s'identifica és la constància a partir de profunditat 25 aproximadament. Destaca el cas de 20 veïns ja que el valor de *F1* es manté superior a la resta pràcticament per a totes les profunditats d'arbre, sent 7 la millor de totes. A la Taula 37 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,92	0,81	0,86
Suspesos	0,26	0,5	0,34

Taula 37: Mètriques per Mètodes Numèrics amb SMOTE

BorderlineSMOTE

Gràfic 32: Valors F1 score per a la classe suspesos per Mètodes Numèrics amb BorderlineSMOTE

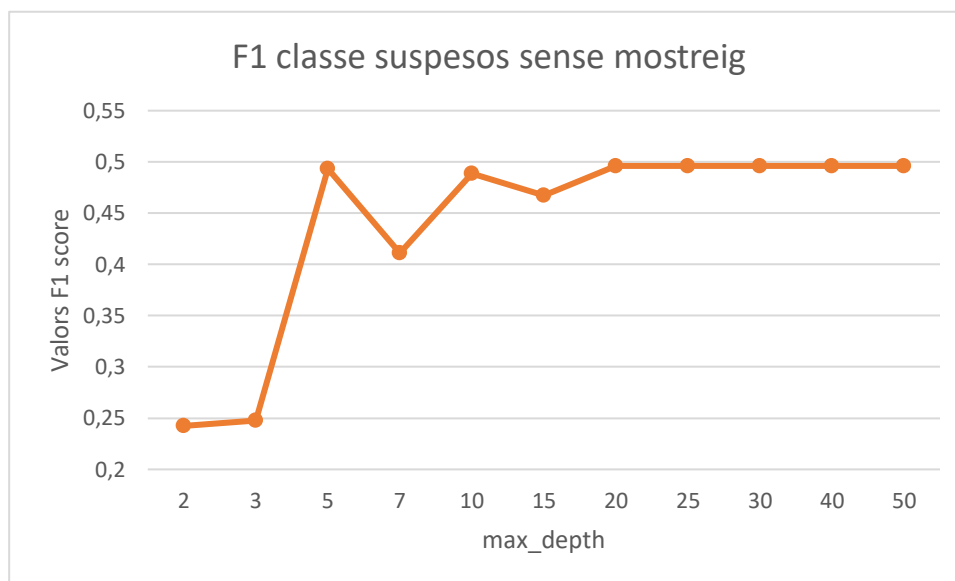
Tal i com podem observar al Gràfic 32, que mostra els resultats de la tècnica *SMOTE*, existeixen variacions dels valors de *F1* per a profunditats d'arbre inferiors a 20. A partir d'aquí els valors es mantenen constants. El millor resultat s'obté amb 2 veïns i una profunditat d'arbre 2. A la Taula 38 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovatats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,94	0,70	0,80
Suspesos	0,23	0,66	0,34

Taula 38: Mètriques per Mètodes Numèrics amb BorderlineSMOTE

7.2.3. Materials

Sense tècniques de mostreig

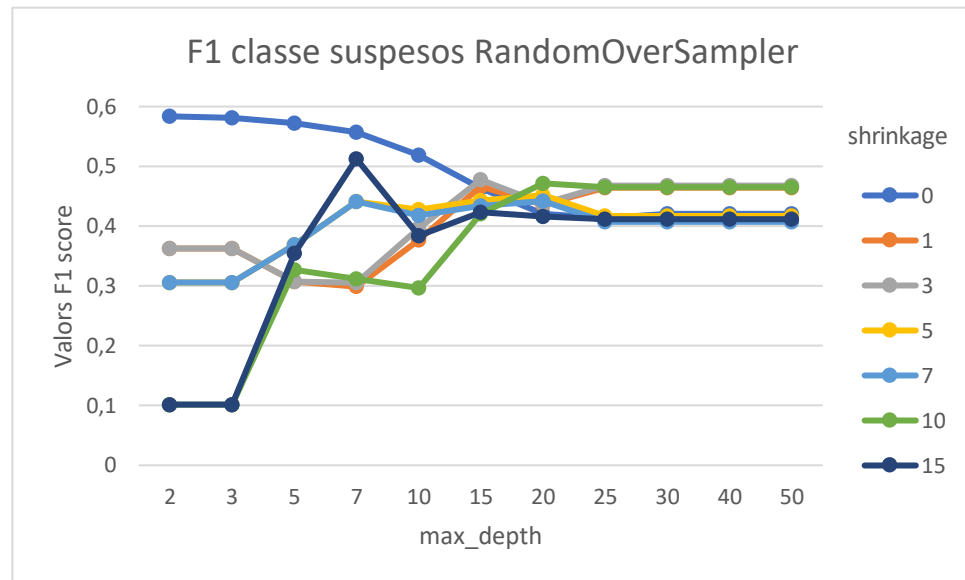


Gràfic 33: Valors F1 score per a la classe suspesos per Materials sense tècniques de mostreig

En el cas de l'assignatura de materials, els resultats de *F1* sense tècniques de mostreig que es mostren al Gràfic 33 augmenten significativament quan passem de profunditat 3 a 5. A profunditat 7 *F1* pateix una caiguda, però torna a recuperar a partir de 10 i ja es manté pràcticament constant. La profunditat d'arbre que ens proporciona millors resultats és 5. A la Taula 39 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,73	0,73	0,73
Suspesos	0,49	0,49	0,49

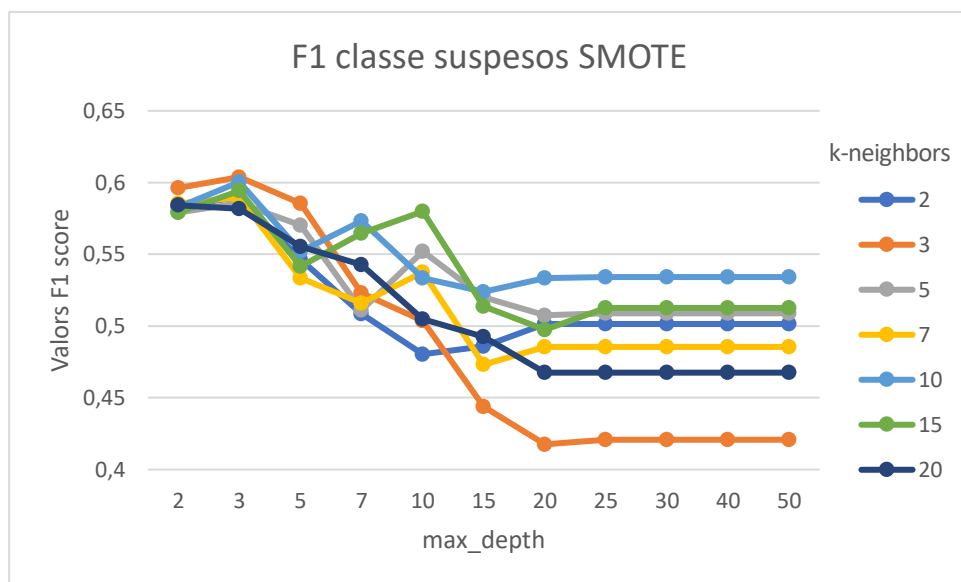
Taula 39: Mètriques per Materials sense tècniques de mostreig

RandomOverSamplerGràfic 34: Valors *F1* score per a la classe suspesos per Materials amb RandomOverSampler

En els resultats de *F1* per a la tècnica *RandomOverSampler* que es mostren al Gràfic 34, podem observar tendències diverses per als diferents valors de shrinkage, sobretot per a profunditats d'arbre petites. A partir de profunditat 15 els valors de *F1* es mantenen constants. Amb una profunditat d'arbre de 2 i un shrinkage de 0 s'obtenen els millors resultats de *F1*. A la Taula 40 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,83	0,47	0,60
Suspesos	0,45	0,82	0,58

Taula 40: Mètriques per Materials amb RandomOverSampler

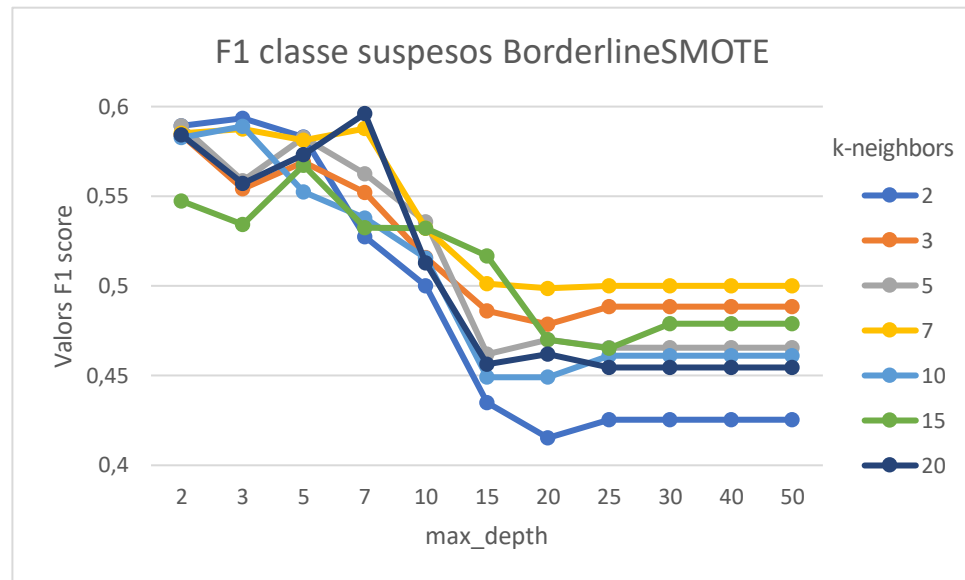
SMOTE

Gràfic 35: Valors *F1* score per a la classe suspesos per Materials amb SMOTE

Tal i com podem veure al Gràfic 35, els valors de *F1* amb la tècnica SMOTE segueixen la mateixa tendència per al diferent nombre de veïns. Aquest patró és que *F1* va disminuint a mesura que augmentem la profunditat de l'arbre i sent constant a partir de profunditat 20. Amb 3 veïns obtenim el valor de *F1* més elevat, concretament amb una profunditat d'arbre de 3. A la Taula 41 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,87	0,47	0,61
Suspesos	0,46	0,86	0,60

Taula 41: Mètriques per Materials amb SMOTE

BorderlineSMOTE

Gràfic 36: Valors F1 score per a la classe suspesos per Materials amb BorderlineSMOTE

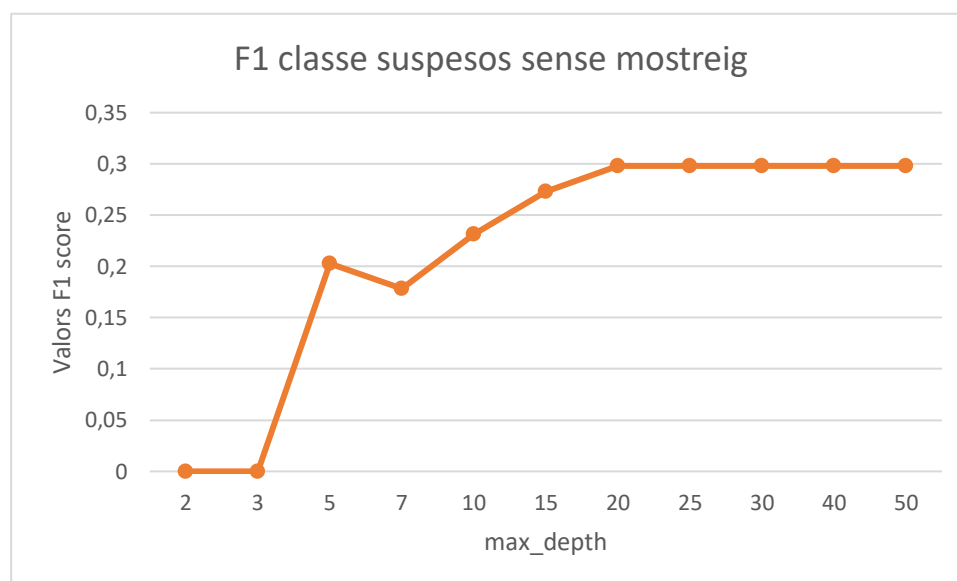
En el Gràfic 36 es mostren els resultats per a la tècnica *BorderlineSMOTE*. S'observa la mateixa tendència que en el cas anterior, on que *F1* va disminuint a mesura que augmentem la profunditat de l'arbre i sent constant a partir de profunditat 20. En aquest cas, la configuració òptima és amb 2 veïns i una profunditat d'arbre de 3. A la Taula 42 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,83	0,52	0,64
Suspesos	0,47	0,80	0,59

Taula 42: Mètriques per Materials amb BorderlineSMOTE

7.2.4. Equacions diferencials

Sense tècniques de mostreig

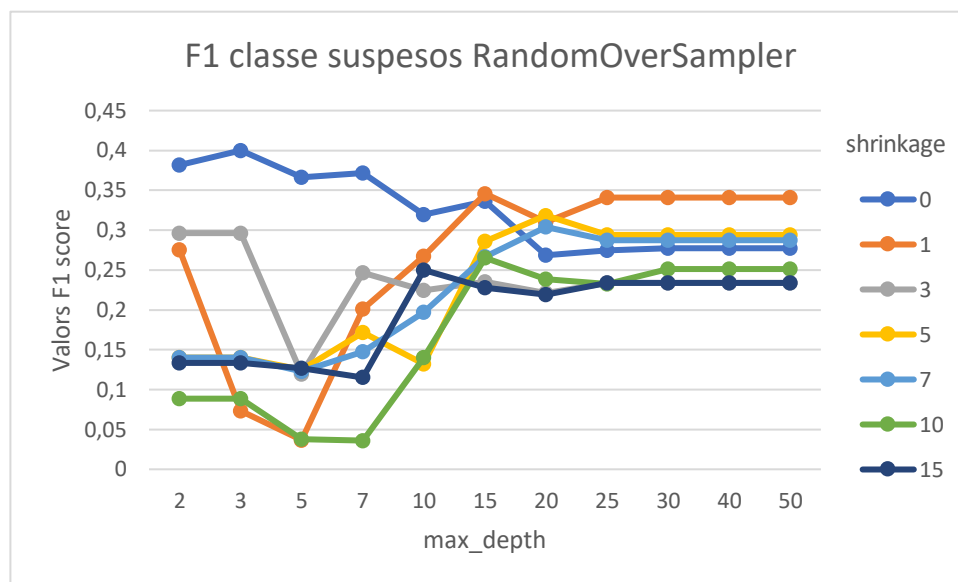


Gràfic 37: Valors F1 score per a la classe suspesos per Equacions Diferencials sense tècniques de mostreig

En el cas de l'assignatura equacions diferencials, els resultats de *F1* sense tècniques de mostreig, que podem veure al Gràfic 37, augmenten progressivament a mida que augmenta la profunditat de l'arbre fins estabilitzar-se en el punt més alt quan la profunditat és 20. A la Taula 43 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,83	0,82	0,83
Suspesos	0,29	0,31	0,30

Taula 43: Mètriques per Equacions Diferencials sense tècniques de mostreig

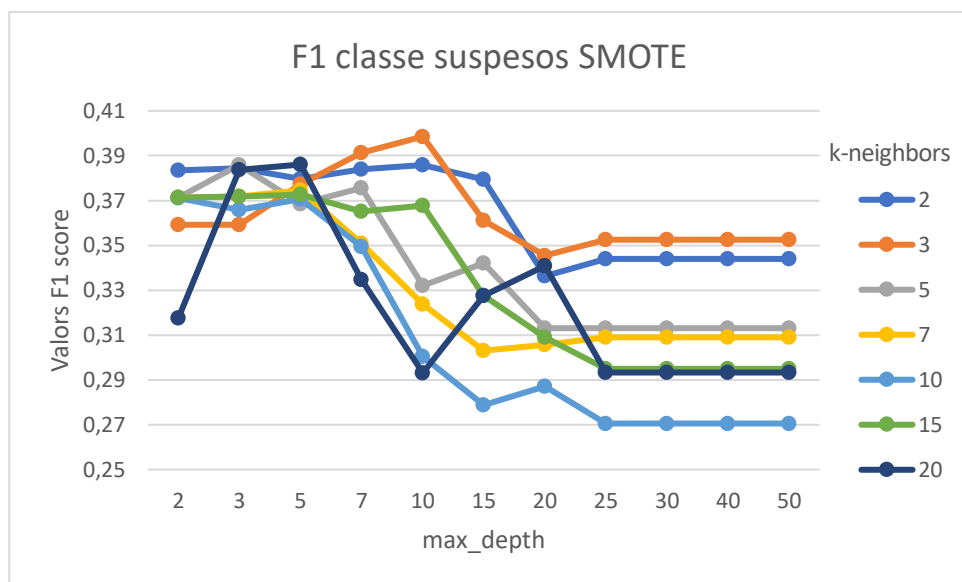
RandomOverSampler

Gràfic 38: Valors $F1$ score per a la classe suspesos per Equacions Diferencials amb *RandomOverSampler*

En el Gràfic 38 es mostren els valors de $F1$ amb la tècnica *RandomOverSampler*. Analitzant aquest gràfic observem un patró semblat per a gairebé tots els valors de shrinkage on els valors de $F1$ cauen per a profunditats d'arbre entre 3 i 7 i després augmenten fins mantenir-se constants a partir d'una profunditat de 20. En el cas de shrinkage 0 no veiem aquesta caiguda sinó que el valor de $F1$ va disminuint poc a poc fins mantenir-se constant. El valor de $F1$ més elevat s'obté amb una profunditat d'arbre de 3 i un shrinkage de 0. A la Taula 44 es mostren les mètriques *Precision*, *Recall* i $F1$ tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	$F1$
Aprovats	0,90	0,50	0,64
Suspesos	0,27	0,77	0,4

Taula 44: Mètriques per Equacions Diferencials amb *RandomOverSampler*

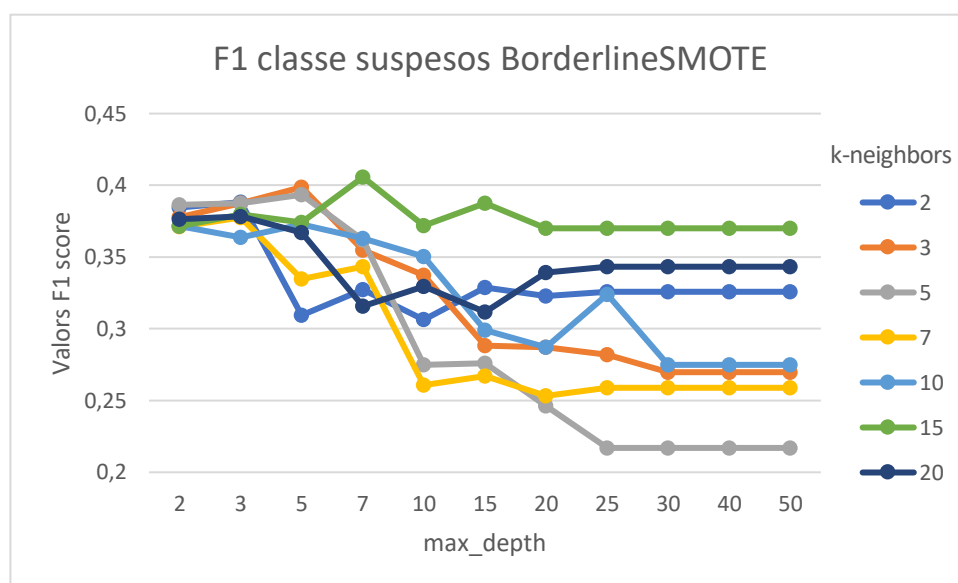
SMOTE

Gràfic 39: Valors *F1* score per a la classe suspesos per Equacions Diferencials amb SMOTE

Tal i com podem observar al Gràfic 39, que mostra els valors de *F1* obtinguts amb la tècnica SMOTE, aquests segueixen el mateix patró per als diferents nombres de veïns. Aquest patró està marcat per un descens a mesura que augmenta la profunditat fins mantenir-se un valor constant a partir d'una profunditat de 25. Cal destacar que s'obtenen millors resultats amb un nombre baix de veïns, sent el millor amb 3 veïns i una profunditat de 10. A la Taula 45 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,87	0,73	0,79
Suspesos	0,32	0,53	0,40

Taula 45: Mètriques per Equacions Diferencials amb SMOTE

BorderlineSMOTE

Gràfic 40: Valors *F1* score per a la classe suspesos per Equacions Diferencials amb *BorderlineSMOTE*

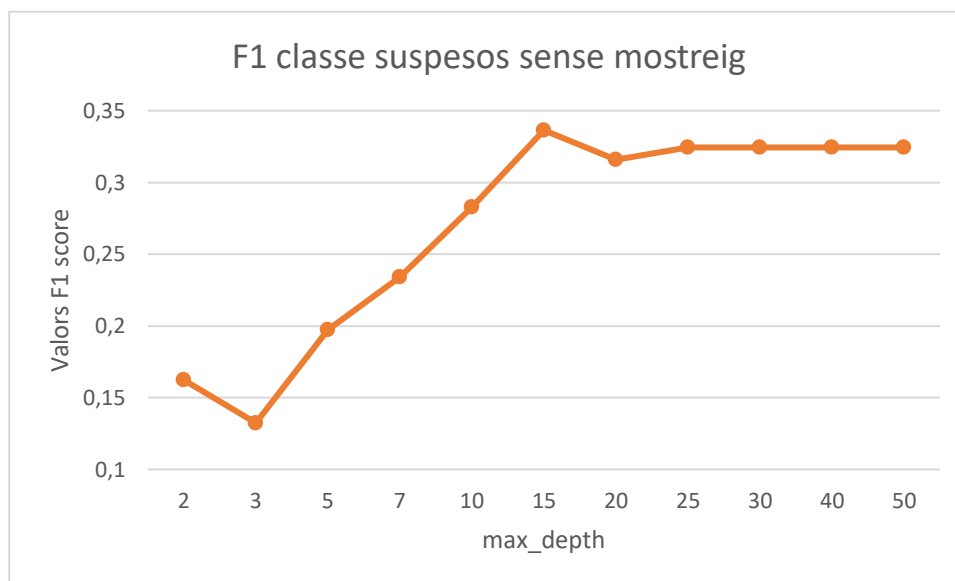
Els resultats de *F1* obtinguts amb la tècnica *BorderlineSMOTE* es mostren al Gràfic 40. Com hem vist en anteriors resultats es segueix la tendència de disminuir a mida que augmenta la profunditat de l'arbre fins mantenir-se constant a partir de 25 aproximadament. En aquest cas existeix una excepció i és pel cas de 15 veïns on veiem que *F1* es manté més o menys constant aconseguint el valor més alt amb una profunditat de 7. No obstant, amb un model de profunditat 5 i 3 veïns s'obté un valor pràcticament igual. A la Taula 46 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,87	0,67	0,76
Suspesos	0,30	0,59	0,40

Taula 46: Mètriques per Equacions Diferencials amb *BorderlineSMOTE*

7.2.5. Informàtica

Sense tècniques de mostreig

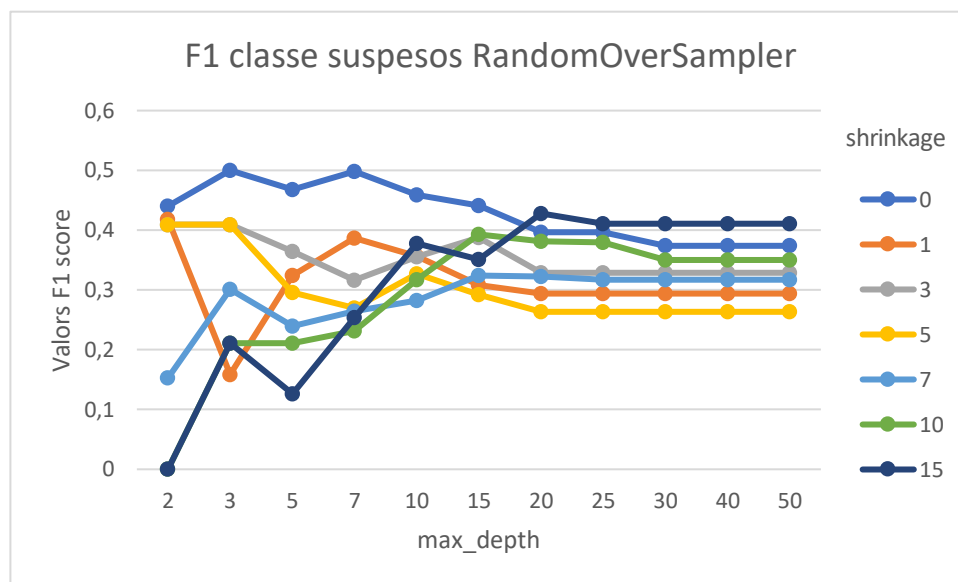


Gràfic 41: Valors *F1* score per a la classe suspesos per Informàtica sense tècniques de mostreig

En el cas de l'assignatura d'informàtica, els resultats de *F1* sense tècniques de mostreig, que podem veure al Gràfic 41, augmenten progressivament a mida que augmenta la profunditat de l'arbre fins establir-se quan la profunditat és 25. La profunditat que ens proporciona un valor de *F1* més elevat és 15. A la Taula 47 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,81	0,83	0,82
Suspesos	0,35	0,32	0,34

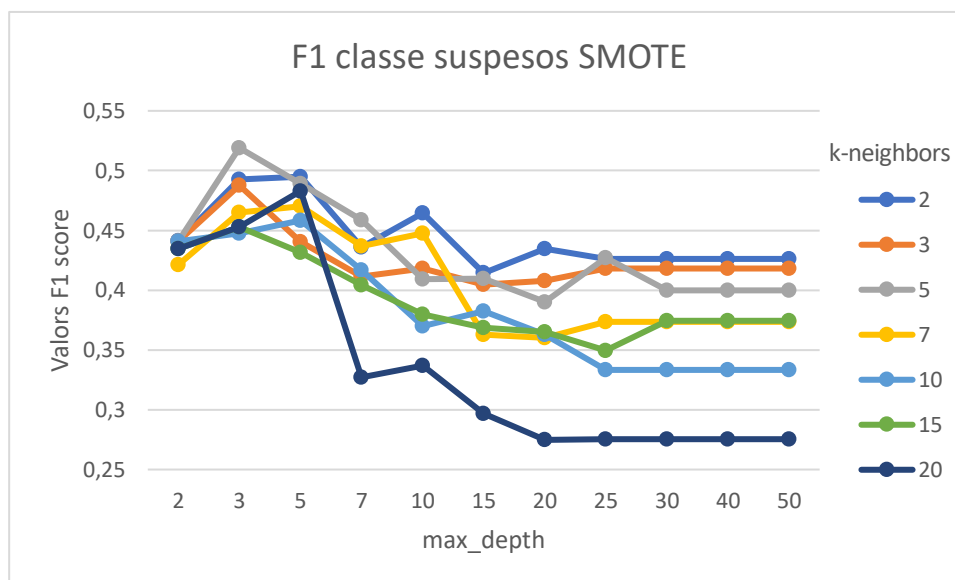
Taula 47: Mètriques per Informàtica sense tècniques de mostreig

RandomOverSamplerGràfic 42: Valors *F1* score per a la classe suspesos per Informàtica amb *RandomOverSampler*

En els resultats de *F1* per a la tècnica *RandomOverSampler* que es mostren al Gràfic 42, podem observar tendències diverses per als diferents valors de shrinkage, sobretot per a profunditats d'arbre petites. A partir de profunditat 20 els valors de *F1* es mantenen constants. Amb profunditats d'arbre de 3 i 7 i un shrinkage de 0 s'obtenen els millors resultats de *F1*. A la Taula 48 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,89	0,68	0,77
Suspesos	0,38	0,71	0,5

Taula 48: Mètriques per Informàtica amb *RandomOverSampler*

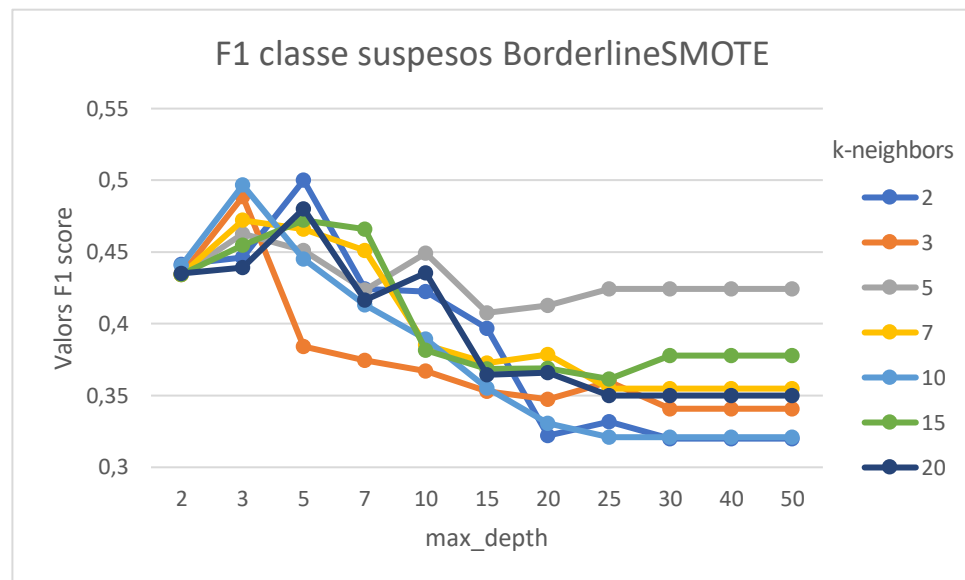
SMOTE

Gràfic 43: Valors *F1* score per a la classe suspesos per Informàtica amb SMOTE

Tal i com podem observar al Gràfic 43, que mostra els valors de *F1* obtinguts amb la tècnica *SMOTE*, aquests segueixen el mateix patró per als diferents nombres de veïns. Aquest patró està marcat per un descens a mesura que augmenta la profunditat fins mantenir-se un valor constant a partir d'una profunditat de 25. Cal destacar que s'obtenen millors resultats amb un nombre baix de veïns, sent el millor amb 5 veïns i una profunditat de 3. A la Taula 49 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,89	0,76	0,82
Suspesos	0,43	0,65	0,52

Taula 49: Mètriques per Informàtica amb SMOTE

BorderlineSMOTEGràfic 44: Valors *F1* score per a la classe suspesos per Informàtica amb BorderlineSMOTE

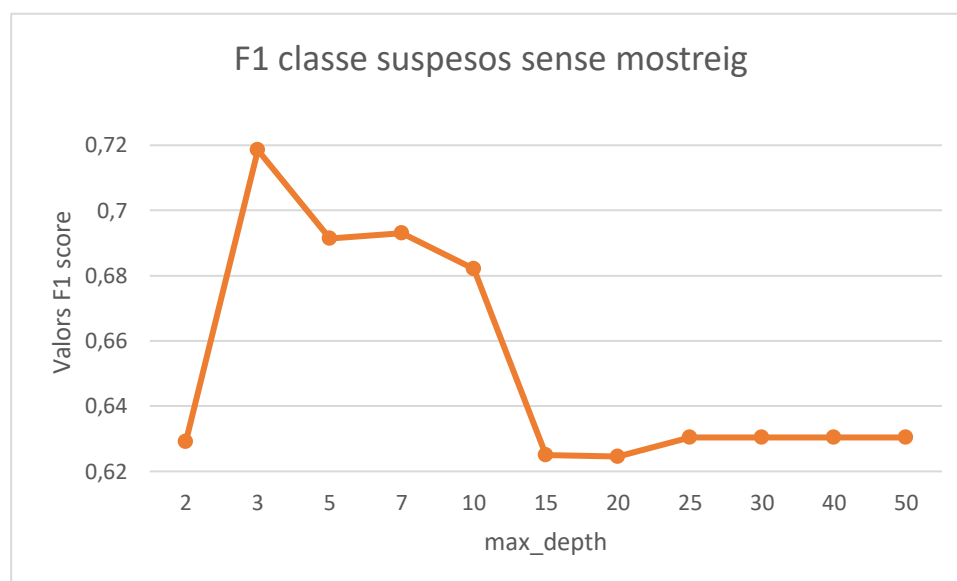
Analitzant el Gràfic 44 veiem que amb la tècnica *BorderlineSMOTE* s'observa la mateixa tendència que en el cas anterior on *F1* va disminuint quan augmenta la profunditat d'arbre encara que van sorgint diversos pics en el descens. Destaca la configuració de 2 veïns i profunditat d'arbre 5 amb la que obtenim el millor valor de *F1*. A la Taula 50 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovats com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,88	0,72	0,80
Suspesos	0,40	0,66	0,5

Taula 50: Mètriques per Informàtica amb BorderlineSMOTE

7.2.6. Mecànica

Sense tècniques de mostreig

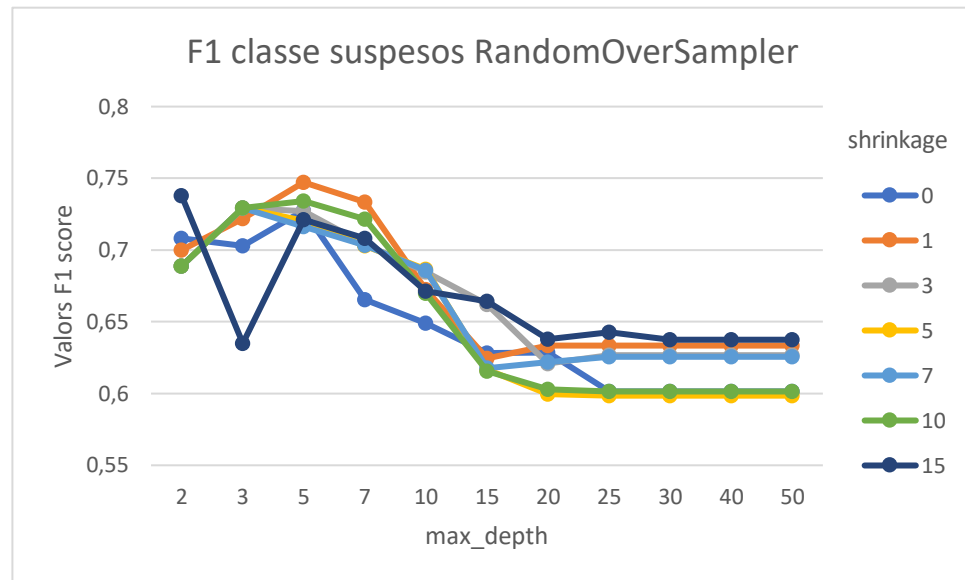


Gràfic 45: Valors F1 score per a la classe suspesos per Mecànica sense tècniques de mostreig

En el cas de l'assignatura de mecànica, els resultats de la qual es mostren al Gràfic 45, s'observa un gran pic a profunditat d'arbre 3 i un posterior descens a mida que augmenta la profunditat. A la Taula 51 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,72	0,45	0,55
Suspesos	0,63	0,84	0,72

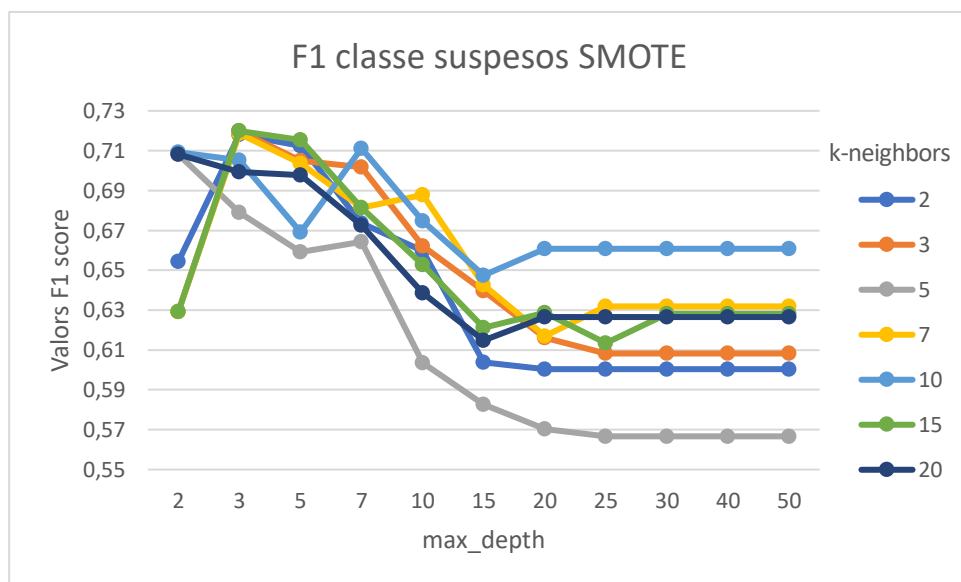
Taula 51: Mètriques per Mecànica sense tècniques de mostreig

RandomOverSamplerGràfic 46: Valors *F1* score per a la classe suspesos per Mecànica amb *RandomOverSampler*

En el Gràfic 46 podem veure els valors de *F1* obtinguts amb la tècnica *RandomOverSampler*. S'observa clarament que per a tots els valors de shrinkage es segueix un patró on inicialment els valors de *F1* augmenten fins a una profunditat de 5 i després disminueixen fins a mantenir-se constants a partir de profunditat 20. La configuració que ens proporciona un millor resultat és profunditat d'arbre 5 i shrinkage 1. A la Taula 52 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,76	0,52	0,62
Suspesos	0,66	0,85	0,75

Taula 52: Mètriques per Mecànica amb *RandomOverSampler*

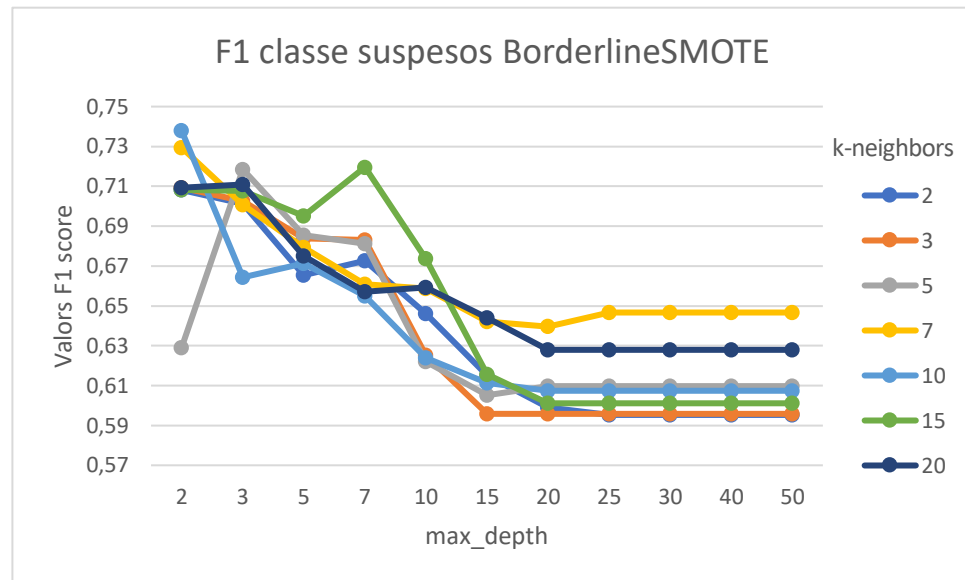
SMOTE

Gràfic 47: Valors *F1 score* per a la classe suspesos per Mecànica amb SMOTE

Tal i com podem observar al Gràfic 47, els valors de *F1* amb la tècnica SMOTE generalment disminueixen a mida que augmenta la profunditat d'arbre fins mantenir-se constants a partir de profunditat d'arbre 25. El valor més elevat de *F1* s'obté amb 2 veïns i una profunditat d'arbre de 3. A la Taula 53 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,72	0,43	0,54
Suspesos	0,62	0,85	0,72

Taula 53: Mètriques per Mecànica amb SMOTE

BorderlineSMOTE

Gràfic 48: Valors F1 score per a la classe suspesos per Mecànica amb BorderlineSMOTE

En el Gràfic 48 es mostren els resultats per a la tècnica *BorderlineSMOTE*. S'observa la mateixa tendència que en el cas anterior, on que *F1* va disminuint a mesura que augmentem la profunditat de l'arbre i sent constant a partir de profunditat 20 aproximadament. En aquest cas, la configuració òptima és amb 10 veïns i una profunditat d'arbre de 2. A la Taula 54 es mostren les mètriques *Precision*, *Recall* i *F1* tant de la classe d'aprovat com de la classe suspesos corresponents a aquesta configuració.

Classe	<i>Precision</i>	<i>Recall</i>	<i>F1</i>
Aprovats	0,76	0,47	0,58
Suspesos	0,64	0,87	0,74

Taula 54: Mètriques per Mecànica amb BorderlineSMOTE

7.2.7. Resum resultats arbres de decisió

Per poder tenir una millor visió de tot el conjunt de resultats obtingut amb el model d'arbres de decisió s'ha construït una taula que engloba els millors valors de $F1$ per a la classe suspesos i els respectius valors de *Precision* i *Recall*. També s'han afegit els resultats de la classe aprovats ja que, malgrat no ser l'objectiu principal del projecte, no es poden deixar de banda. Aquesta recopilació de resultats es mostra a la Taula 55. D'altra banda s'ha construït la Taula 56 on podem veure la variació de $F1$ quan apliquem la tècnica de mostreig respecte al valor quan no s'aplica.

En gairebé totes les assignatures es produeix un notable increment del rendiment de la predicció de suspesos aplicant els mètodes d'*oversampling* a excepció de l'assignatura de mecànica en la que la variació és insignificant. Per tant, arribem a la conclusió que era necessari equilibrar les dades de les dues classes ja que en tenir una gran quantitat d'aprovats al conjunt d'entrenament de dades, el model identificava molts aprovats que no eren correctes. A continuació es comentarà més detingudament com han variat els resultats en aplicar les tècniques de mostreig per a cada assignatura.

En l'assignatura d'electromagnetisme el rendiment de la predicció de suspesos computat amb $F1$ augmenta entre un 20% i un 26% quan apliquem les tècniques de mostreig. Aquesta millor predicció dels suspesos comporta un descens d'entre el 3% i el 8% del rendiment en la predicció d'aprovats.

En el cas de mètodes numèrics es produeix una millora més significativa que l'anterior. El rendiment de $F1$ per a la classe suspesos augmenta entre 47% i 57% mentre que el de la classe d'aprovats només disminueix entre un 5% i 15%. Aquesta millora tant significativa en la predicció de suspesos és deguda a que l'assignatura de mètodes numèrics era la més desequilibrada inicialment.

Continuem amb materials on tornem a veure una millora en la predicció de suspesos en aplicar les tècniques de mostreig. Es produeix un augment d'entre un 18% i un 22% en $F1$ dels suspesos. No obstant, també es torna a repetir la pèrdua de rendiment en els aprovats on trobem una disminució entre el 12% i 18%.

Per l'assignatura d'equacions diferencials es repeteix aquest augment en el rendiment de l'encert de suspesos incrementant-lo un 33% tenint també un decrement en el d'aprovats d'entre 5% i 23%.

En el cas d'informàtica la millora en la predicció dels suspesos és una mica més significativa trobant-se entre un 47% i un 53%. A més, cal destacar que la disminució del rendiment dels aprovats és menor que en el cas anterior tot i tenir-ne un augment major en els suspesos. El descens es troba entre un 0% i un 6%.

Per últim, trobem el cas de mecànica, l'única assignatura que es trobava inicialment equilibrada. Per aquest motiu, les variacions en el rendiment de la predicció tant dels suspesos com dels aprovats són insignificants sent d'entre un 1% i un 5% en ambdós casos. No obstant, cal destacar que amb la tècnica *RandomOverSampler* el rendiment de la predicció d'aprovats augmenta un 12,3%. També és important veure que augmenta el rendiment de predicció tant d'aprovats com de suspesos menys amb *SMOTE* on el rendiment de predicció d'aprovats disminueix.

No es pot detectar en quins casos funciona millor una tècnica de mostreig o una altra ja que per a cada assignatura obtenim millors resultats amb una tècnica diferent.

És necessari comentar que no es troba cap configuració òptima de valors dels paràmetres. En cada cas s'obté una combinació diferent de paràmetres que ens proporciona un millor resultat. No obstant això, es pot veure com, en la gran majoria de casos, els paràmetres es mantenen en valors baixos. En el cas del paràmetre *shrinkage*, no trobem cap superior a 1. El paràmetre *k-neighbors* també es manté baix però, en aquest cas, si que trobem alguna excepció on el nombre de veïns arriba a 10 o 20. Per últim, en el paràmetre *max-depth* dels arbres de decisió s'observa una certa tendència a disminuir quan apliquem les tècniques de mostreig. Això significa que es necessita un model més senzill per arribar a un resultat òptim quan les dues classes estan equilibrades. És a dir, el desequilibri tendeix a fer que s'hagi de construir models més complexos per compensar la falta d'informació. Un altre fet que corrobora aquesta afirmació és que la profunditat de l'arbre quan apliquem les tècniques de mostreig a l'assignatura de mecànica pràcticament no varia i es manté en nombres baixos ja que aquesta matèria ja tenia dades equilibrades.

Com a conclusió podem dir que les tècniques de mostreig ens permeten obtenir un augment significatiu en la predicció de la classe minoritària quan les dades estan desequilibrades. Això comporta una pèrdua de rendiment per a la classe majoritària, però és molt inferior en comparació amb la millora. Per tant, concloem que val la pena el sacrifici sobretot si el nostre objectiu és la predicció dels suspesos.

				Classe aprovats			Classe suspesos		
		m_d	s/k	Precision	Recall	F1	Precision	Recall	F1
Electro-magnetisme	Sense mostreig	7	-	0,73	0,76	0,74	0,51	0,48	0,49
	RandomOverSampler	3	0	0,80	0,58	0,68	0,49	0,74	0,59
	SMOTE	5	7	0,82	0,58	0,68	0,50	0,77	0,61
	BorderlineSMOTE	7	10	0,83	0,63	0,72	0,53	0,76	0,62
Mètodes numèrics	Sense mostreig	5	-	0,90	0,98	0,94	0,53	0,15	0,23
	RandomOverSampler	2	1	0,92	0,87	0,89	0,30	0,44	0,36
	SMOTE	7	20	0,92	0,81	0,86	0,26	0,50	0,34
	BorderlineSMOTE	2	2	0,94	0,70	0,80	0,23	0,66	0,34
Materials	Sense mostreig	5	-	0,73	0,73	0,73	0,49	0,49	0,49
	RandomOverSampler	2	0	0,83	0,47	0,60	0,45	0,82	0,58
	SMOTE	3	3	0,87	0,47	0,61	0,46	0,86	0,60
	BorderlineSMOTE	3	2	0,83	0,52	0,64	0,47	0,80	0,59
Equacions diferencials	Sense mostreig	20	-	0,83	0,82	0,83	0,29	0,31	0,30
	RandomOverSampler	3	0	0,90	0,50	0,64	0,27	0,77	0,40
	SMOTE	10	3	0,87	0,73	0,79	0,32	0,53	0,40
	BorderlineSMOTE	5	3	0,87	0,67	0,76	0,30	0,59	0,40
Informàtica	Sense mostreig	15	-	0,81	0,83	0,82	0,35	0,32	0,34
	RandomOverSampler	3	0	0,89	0,68	0,77	0,38	0,71	0,50
	SMOTE	3	5	0,89	0,76	0,82	0,43	0,65	0,52
	BorderlineSMOTE	5	2	0,88	0,72	0,80	0,40	0,66	0,50
Mecànica	Sense mostreig	3	-	0,72	0,45	0,55	0,63	0,84	0,72
	RandomOverSampler	5	1	0,76	0,52	0,62	0,66	0,85	0,75
	SMOTE	3	2	0,72	0,43	0,54	0,62	0,85	0,72
	BorderlineSMOTE	2	10	0,76	0,47	0,58	0,64	0,87	0,74

Taula 55: Mètriques Arbres de Decisió (on s=shrinkage per RandomOverSampler i k=k-neighbors per SMOTE i BorderlineSMOTE)

		Classe aprovats	Classe suspesos
Electromagnetisme	RandomOverSampler	-8,11	20,41
	SMOTE	-8,11	24,49
	BorderlineSMOTE	-2,70	26,53
Mètodes numèrics	RandomOverSampler	-5,32	56,52
	SMOTE	-8,51	47,83
	BorderlineSMOTE	-14,89	47,83
Materials	RandomOverSampler	-17,81	18,37
	SMOTE	-16,44	22,45
	BorderlineSMOTE	-12,33	20,41
Equacions diferencials	RandomOverSampler	-22,89	33,33
	SMOTE	-4,82	33,33
	BorderlineSMOTE	-8,43	33,33
Informàtica	RandomOverSampler	-6,10	47,06
	SMOTE	0,00	52,94
	BorderlineSMOTE	-2,44	47,06
Mecànica	RandomOverSampler	12,73	4,17
	SMOTE	-1,82	0,00
	BorderlineSMOTE	5,45	2,78

Taula 56: Variacions del valor F1 (en %) amb la tècnica de mostreig respecte al obtingut sense aplicar tècniques de mostreig

7.3. Comparació entre els models predictius regressió logística i arbres de decisió

En aquest apartat es farà una comparació entre els dos models predictius emprats en aquest projecte, la regressió logística i els arbres de decisió. S'analitzarà el rendiment de la predicció dels resultats acadèmics obtingut amb els dos mètodes per veure quin dels dos ens proporciona millors resultats. S'estudiarà també com els afecta l'aplicació de les tècniques de mostreig per equilibrar la distribució de dades amb l'objectiu de descobrir si la diferència en el rendiment en les prediccions són causades per aquestes o bé per la diferència entre els models predictius.

Aquesta comparativa es farà mitjançant les dades de les taules Taula 29, Taula 30, Taula 55 i Taula 56 on han quedat recollits els resultats obtinguts per a cada mètode.

Començarem amb els resultats obtinguts sense aplicar tècniques de mostreig. Per a la majoria d'assignatures no es veu diferència entre els rendiments de $F1$ de la regressió logística i els arbres de decisió. No obstant, trobem dos casos d'excepció. El primer és electromagnetisme on la regressió logística funciona una mica millor. D'altra banda, tenim el cas de mètodes numèrics on succeeix el contrari, el rendiment dels arbres de decisió destaca respecte la regressió logística.

A continuació, s'estudiarà com han afectat als resultats l'aplicació de les diferents tècniques de mostreig que són *RandomOverSampler*, *SMOTE* i *BorderlineSMOTE*. Com a norma general podem veure que els resultats de $F1$ que s'obtenen quan apliquem les tècniques de mostreig són significativament millors amb el model de regressió logística. Tanmateix, en l'assignatura de mecànica, que ja estava inicialment equilibrada, s'observa que amb la regressió logística quan apliquem les tècniques de mostreig, el rendiment de predicció disminueix mentre que quan ho fem amb els arbres de decisió augmenta. Un altre cas on l'arbre de decisió obté millor resultat que la regressió logística és quan apliquem *RandomOverSampler* a l'assignatura de mètodes numèrics.

Per poder entendre el perquè dels resultats és necessari tenir en compte que la regressió logística és un model predictiu lineal, és a dir, intenta separar les dades mitjançant una línia. Per tant, quan la distribució de les dades pren aquesta forma, la regressió logística ens proporciona molt bons resultats. En canvi, quan les dades no s'ajusten d'aquesta forma es produeix el que s'anomena un error de biaix ja que el model no s'adaptarà al conjunt de dades perquè és massa rígid.

D'altra banda, el model d'arbres de decisió és un model molt més flexible que permet divisions de dades molt més complexes jugant amb les profunditats i les branques de l'arbre. Per aquest motiu, és un mètode molt sensible als canvis en el conjunt de dades. El fet que el model s'ajusti massa al conjunt de dades pot causar en ocasions *overfitting*. Tanmateix, és necessari mencionar que al model d'arbres de decisió es poden modificar molts més paràmetres a banda de la profunditat, per tant, d'aquesta manera s'ha estat restringint el model.

Com a conclusió, un cop equilibrades les dades mitjançant les tècniques de mostreig, podem dir que generalment la regressió logística proporciona millors resultats. No obstant això, com ja s'ha dit anteriorment, existeixen molts més paràmetres que es podrien estudiar i que, probablement, afectarien al rendiment dels diferents models.

7.4. Aplicació pràctica del model

En aquest apartat es comentaran alguns aspectes relacionats amb l'aplicació pràctica del model predictiu, és a dir, els possibles criteris per decidir quina tècnica de mostreig aplicar.

En els resultats obtinguts amb els diferents mètodes de mostreig, que es mostren a les taules Taula 29 i Taula 55, s'observa que la variació entre tècniques dels valors de $F1$ són mínimes, en la majoria de casos de 0,01 i no més de 0,03. Aquesta diferència, quan apliquem el model de forma pràctica, no és significativa ja que si suposem que en un quadrimestre trobem uns 200 matriculats en una assignatura, aquesta variació d'entre 0,01 i 0,03 implicaria aproximadament un o dos estudiants. Per tant, en un cas pràctic, es podria considerar aplicar el mètode de *sampling* més senzill com és el *RandomOverSampler* que duplica aleatòriament exemples ja existents de la classe minoritària o bé escollir la tècnica de mostreig que obtingui bons resultats simplificant l'algoritme predictiu, és a dir, amb una C més baixa per a la regressió logística o una menor profunditat en el cas dels arbres de decisió.

D'altra banda, trobem casos en que, ja sigui per a diferents algoritmes de predicció o bé diferents mètodes de *sampling*, els valors de $F1$ són molt similars però els de *recall* són força diferents. En aquests casos es podria utilitzar un altre criteri per decidir quin d'ells ens proporciona millors resultats. Com l'objectiu és que, dins de valors de $F1$ similars, el model encerti la major part dels suspesos, el criteri per seleccionar model seria promoure els *recalls* més alts de la classe suspesos.

8. Impacte ambiental

L'impacte ambiental d'aquest treball ha estat mínim. En tractar-se d'un projecte de caire informàtic on l'eina principal ha estat l'ordinador, no s'han produït gairebé residus ja que l'ús de paper ha estat mínim.

Si es contempla des del punt de vista energètic, s'hauria de considerar el consum d'energia elèctrica de l'ordinador i del *router* que proporciona la connexió a internet. També es podria tenir en compte l'ús de fonts lumíniques en els períodes on no ha estat possible aprofitar la llum natural.

No obstant això, es pot concloure que l'impacte del projecte és positiu ja que els residus que es generen i el consum elèctric són menyspreables en comparació amb els beneficis que es pot obtenir d'aquest estudi.

9. Planificació

En aquest apartat s'exposa la planificació de tot el projecte des del seu inici fins la seva finalització. S'ha organitzat el projecte en cinc etapes començant des de l'inici del projecte i acabant amb la seva presentació. Cadascuna de les etapes s'ha dividit en diverses activitats. La planificació completa s'exposa mitjançant el diagrama de Gantt de la Taula 57.

		Febrer-21				Març-21				Abril-21				Maig-21				Juny-21				Juliol-21			
INICI DEL PROJECTE	Comprensió del problema i definició d'objectius																								
	Instal·lació de les eines																								
	Familiarització amb la llibreria <i>Pandas</i>																								
COMPENSIÓ I PREPARACIÓ DE DADES	Comprensió de dades																								
	Selecció i neteja de dades																								
	Transformació de dades																								
MODELATGE I VALIDACIÓ	Estudi de models predictius i mètodes i mètriques de validació																								
	Estudi de tècniques de mostreig pel desequilibri de dades																								
	Selecció i construcció dels models																								
	Validació i anàlisi de resultats																								
ELABORACIÓ DE LA MEMÒRIA	Redacció de la memòria																								
	Conclusions																								
PRESENTACIÓ	Preparació de la presentació																								
	Presentació del projecte																								

Taula 57: Diagrama de Gantt del projecte

10. Pressupost

En aquest apartat es farà un càlcul aproximat dels costos associats a la realització del projecte. Aquests costos es desglossaran en costos de personal i els costos d'equips i material.

Els costos de personal fan referència al treball de l'estudiant que realitza el projecte i al professor que el guia i supervisa.

En el cas de l'estudiant, per calcular les hores invertides en l'estudi s'ha considerat l'equivalència entre crèdits ECTS i les hores de treball que impliquen. Així doncs, si el Treball de Fi de Grau consta de 12 crèdits ECTS i un crèdit implica 25 hores de treball, es conclou que el temps invertit són 300 hores. S'ha considerat el sou estàndard d'enginyer júnior de 20 €/h. Per tant, el cost total serà de 6000 €.

En relació a la tutoria del professor, es considera com a enginyer sènior. El sou mig d'enginyer sènior és de 40 €/h i s'ha estimat un total de 50 hores de guia i supervisió per part del professor. Per tant, el cost total resulta de 2000 €.

Pel que fa als costos d'equips i material es consideren els costos de l'ordinador utilitzat al llarg del projecte, el cost dels softwares utilitzats i el material d'oficina necessari.

Per determinar el cost associat a l'ordinador és necessari considerar l'amortització d'aquest. Per calcular-lo es considera que l'ordinador està valorat en 1000 € i que té una esperança de vida de 4 anys. Així doncs, si el projecte dura 20 setmanes, el cost associat a l'ús de l'ordinador es calcula:

$$\text{Cost amortització} = 20 \text{ setmanes} \cdot \frac{1 \text{ any}}{52 \text{ setmanes}} \cdot \frac{1000 \text{ €}}{4 \text{ anys}} = 96 \text{ €}$$

També, s'ha considerat el preu dels softwares utilitzats a l'estudi. Per una banda, el software de programació *Anaconda* és un software lliure i, per tant, no implica cap cost. D'altra banda, s'ha fet ús d'algunes de les eines de *Microsoft 365* com són *Word*, *Excel* i *Power Point*. Això implica disposar de *Microsoft 365* que té un preu de 7 €/mes. Per tant, el cost total de llicències de programes és de 35 €.

Per últim, s'ha afegit el cost associat a l'ús de material d'oficina com poden ser fulls de paper, bolígrafs, etc. S'estima un cost total de 20 €.

Un cop calculats tots els costos implicats en la realització de l'estudi s'obté que el cost total d'aquest és de 8151 €. A la Taula 58 es mostra el desglossament de tots els costos i el cost total del projecte.

	Preu	Unitats	Cost
Costos de personal			
Enginyer júnior	20 €/h	300 h	6.000 €
Enginyer sènior	40 €/h	50 h	2.000 €
Costos d'equips i material			
Amortització Ordinador			96 €
Software Anaconda			0 €
Software Microsoft Office 365	7 €/mes	5 mesos	35 €
Material d'oficina			20 €
TOTAL			8.151 €

Taula 58: Desglossament dels costos del projecte

11. Conclusions

Després d'haver realitzat el present treball, es considera que s'han complert tots els objectius que s'havien plantejat inicialment. S'ha estudiat el rendiment de diverses tècniques de mostreig per a predir si un estudiant aprovarà o suspendrà una determinada assignatura del tercer quadrimestre del Grau en Enginyeria en Tecnologies Industrials de l'ETSEIB.

Al llarg del projecte s'ha seguit de forma rigorosa la metodologia CRISP-DM per al desenvolupament de projectes de mineria de dades adaptant cadascuna de les seves fases a l'estudi i documentant-les per que en un futur pugui ser replicada a partir d'aquest document.

Per poder comparar el rendiment de diverses tècniques de mineria de dades en la predicció de resultats acadèmics, que era l'objectiu principal del projecte, s'ha utilitzat dos models predictius com són la regressió logística i els arbres de decisió als quals s'ha aplicat les diferents tècniques de mostreig: *RandomOverSampler*, *SMOTE* i *BorderlineSMOTE*. Un cop construïts els models s'ha procedit a validar-los mitjançant el mètode *Holdout* i a analitzar tots els resultats obtinguts. Com l'interès de l'estudi era determinar el nombre de suspesos de les diferents assignatures s'ha utilitzat com a mètrica principal el *F1* de la classe suspesos amb el suport de les mètriques *Precision* i *Recall*. Tot i això, no s'ha deixat de banda la predicció dels aprovats ja que l'error en alguna de les dues variables provocaria una predicció falsa.

Després de realitzar tot l'estudi, es pot concloure que les tècniques de mostreig ens permeten obtenir un augment significatiu en la predicció de la classe minoritària quan les dades estan desequilibrades. Això comporta una pèrdua de rendiment per a la classe majoritària, però és molt inferior en comparació amb la millora. També, cal destacar que els valors dels paràmetres *C* de la regressió logística i *max_depth* dels arbres de decisió disminuïen amb l'aplicació d'aquestes tècniques. Per tant, les tècniques de mineria de dades permeten arribar a millors resultats amb models molt més senzills. Per últim, cal mencionar que els experiments mostren que s'ha demostrat que els valors dels paràmetres dels diferents mètodes de *sampling* es troben dins d'un rang. Això facilita la seva aplicació pràctica i també els experiments futurs en el cas que es continuï investigant.

Personalment, gràcies a aquest projecte s'ha pogut veure la importància de la mineria de dades i tot el seu potencial ja que permet obtenir informació molt valuosa a partir de conjunts de dades. D'altra banda, s'ha pogut aplicar en un cas pràctic tots els coneixements adquirits durant el grau, especialment els de programació. També ha permès la familiarització amb l'entorn de treball del software *Anaconda* i l'ús de *Jupyter Notebooks* per crear tot el codi de programació de l'estudi. També cal destacar el coneixement i aprenentatge de les diferents llibreries de *Python* com són *Pandas* i *Scikit-learn* amb les que s'ha pogut manipular les dades i aplicar els diferents models.

12. Treballs futurs

A causa de l'extensió del treball ha hagut aspectes que no s'ha pogut abordar en aquest estudi. En aquest apartat es plantejaren diverses alternatives que seria necessari tenir en compte si es continués amb l'estudi en un futur.

Per una banda, existeix una gran quantitat de models predictius que es podrien estudiar i veure com afecten les tècniques de mostreig en altres algoritmes de predicció. A més, els models que s'han utilitzat consten de més paràmetres modificables que podrien afectar als resultats. Per exemple, el model d'arbres de decisió disposa de paràmetres que poden fer variar l'estructura de l'arbre de manera que no totes les fulles estiguin a la mateixa profunditat i així adaptar-se millor a les dades. Per tant, també es podria fer un anàlisi més exhaustiu d'aquests paràmetres.

D'altra banda, es podria ampliar i afegir més tècniques de mineria de dades per equilibrar la distribució de les classes. Com cadascuna d'elles equilibra les dades de manera diferent, seria molt interessant veure com afecten al rendiment de la predicció. També cal destacar que només s'ha aplicat tècniques d'*oversampling* ja que amb les d'*undersampling* el model hauria quedat incomplet degut a la poca quantitat de dades. Per tant, es podria considerar ampliar el conjunt de dades per poder veure com es comporta el rendiment de la predicció quan en lloc de crear nous exemples de la classe minoritària, eliminem de la classe majoritària.

Bibliografia

Referències bibliogràfiques

- [1] J. Saltz, «Data Science Project Management,» 30 Novembre 2020. [En línia]. Available: <https://www.datascience-pm.com/crisp-dm-still-most-popular/>. [Últim accés: 22 Gener 2021].
- [2] David L. Olson, Dursun Delen, Advanced Data Mining Techniques, Springer-Verlag Berlin Heidelberg, 2008.
- [3] KDnuggets, «Top Algorithms and Methods Used by Data Scientists,» Setembre 2016. [En línia]. Available: <https://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html>. [Últim accés: Abril 2021].
- [4] Wikipedia, «Regresión Logística Wikipedia,» 20 Març 2021. [En línia]. Available: https://es.wikipedia.org/wiki/Regresi%C3%B3n_log%C3%ADstica. [Últim accés: Abril 2021].
- [5] Wikipedia, «K vecinos más próximos,» 6 Setembre 2020. [En línia]. Available: https://es.wikipedia.org/wiki/K_vecinos_m%C3%A1s_pr%C3%B3ximos. [Últim accés: Abril 2021].
- [6] MathWorks, «Support Vector Machine (SVM),» 2021. [En línia]. Available: <https://es.mathworks.com/discovery/support-vector-machine.html>. [Últim accés: Abril 2021].
- [7] «Resampling strategies for imbalanced datasets,» 15 Novembre 2017. [En línia]. Available: <https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t1>. [Últim accés: Maig 2021].
- [8] «SMOTE explained for noobs - Synthetic Minority Over-sampling Technique line by line,» 6 Novembre 2017. [En línia]. Available: https://rikunert.com/SMOTE_explained. [Últim accés: Maig 2021].
- [9] The imbalanced-learn developers, «Imbalanced-learn. User guide. Over-sampling,» 2014-2021. [En línia]. Available: https://imbalanced-learn.org/stable/over_sampling.html. [Últim accés: Maig 2021].

[10] J. A. Rodrigo, «Regresión logística simple y múltiple. Ciencia de datos,» [En línia]. Available: https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple. [Últim accés: Abril 2021].

Bibliografia complementària

Allibhai, E., 2018. Holdout vs. Cross-validation in Machine Learning. [En línia] Medium. Available: <https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f> [Últim accés: Abril 2021].

Amat Rodrigo, J., 2021. Arboles de decision python. [En línia] Cienciadedatos.net. Available: https://www.cienciadedatos.net/documentos/py07_arboles_decision_python.html [Últim accés: Abril 2021].

Aprendemachinelearning.com. 2017. Qué es overfitting y underfitting y cómo solucionarlo | Aprende Machine Learning. [En línia] Available: <https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/> [Últim accés: Abril 2021].

Brownlee, J., 2020. Logistic Regression for Machine Learning. [En línia] Machine Learning Mastery. Available: <https://machinelearningmastery.com/logistic-regression-for-machine-learning/> [Últim accés: Abril 2021].

Medium. 2021. A Guide to Decision Trees for Machine Learning and Data Science. [En línia] Available: <https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956> [Últim accés: Abril 2021].

Merkle. 2020. El algoritmo K-NN y su importancia en el modelado de datos | Blog | Merkle. [En línia] Available: <https://www.merkleinc.com/es/es/blog/algoritmo-knn-modelado-datos> [Últim accés: Abril 2021].

Scikit-learn.org. 2021. scikit-learn: machine learning in Python — scikit-learn 0.24.1 documentation. [En línia] Available: <https://scikit-learn.org/stable/> [Últim accés: Maig 2021].

Seif, G., 2021. A Guide to Decision Trees for Machine Learning and Data Science. [En línia] Medium. Available: <<https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956>> [Últim accés: Abril 2021].

Singh Chauhan, N., 2020. Métricas De Evaluación De Modelos En El Aprendizaje Automático. [En línia] DataSource.ai. Available: <<https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico>> [Últim accés: Abril 2021].