Annex

Caracterització de perfils acadèmics

Marina Rosell Murillo

Pau Lozano García

Patricia Cabot Álvarez

Grau en Ciència i Enginyeria de Dades

Juny de 2020

Prof. Lluís Antonio Belanche Muñoz

Prof. Marta Arias Vicente

Índex

Comentari

Aquest és un document complementari al treball sobre la caracterització de perfils acadèmics, on es troba un recull de les gràfiques obtingudes per aclarir la interpretació de l'estudi.

Durant la lectura del projecte pot trobar diverses referències a l'annex caracteritzades per l´ús de *Cursiva* i un seguit de xifres darrere que determinen el nombre de la imatge o del material gràfic adicional.

Totes les imatges, gràfics i continguts han estat obtinguts en l'entorn de R utilitzant les dades de la database sobre la qual s'efectua l'estudi.

1. Procés d'exploració de les dades

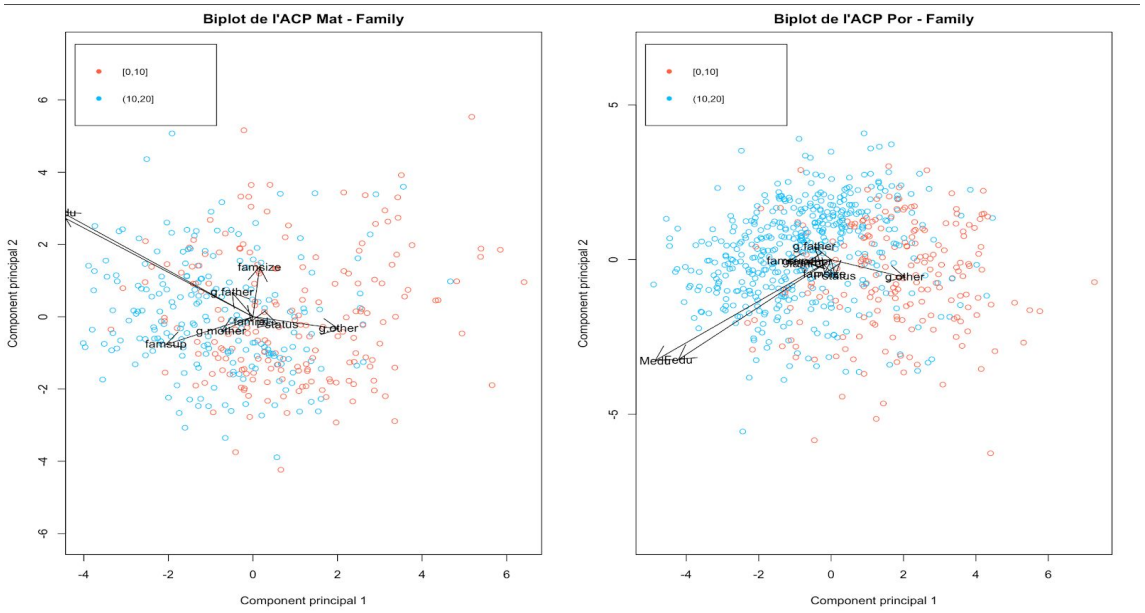
Atributs	Descripció i domini
Escola (school)	Escola de l'estudiant. Binària: "GP" - Gabriel Pereira; "MS" - Mousinho da Silveira.
Sexe (sex)	Sexe de l'estudiant. Binària: "*F*" - femení; "*M*" - masculí.
Edat (age)	Edat de l'estudiant. Numèrica: del 15 al 22.
Direcció (address)	Direcció de la casa de l'estudiant. Binària: "*U*" - urbana; "*R*" - rural.
Tamany de la família (famsize)	Tamany de la família. Binària: "*LE3*" - menys de 3 o igual; "*GT3*" - més de 3.
Situació dels pares (Psatus)	Estat de convivència dels pares. Binària: "*T*" - junts; "*A*" - separats.
Eduació de la mare (Medu)	Eduació de la mare. Numèrica: de 0 (cap) a 5 (educació superior).
Educació del pare (Feud)	Eduació del pare. Numèrica: de 0 (cap) a 5 (educació superior).
Feina de la mare (Mjob)	Feina de la mare. Nominal: professora - "*teacher*"; relacionat amb la salut - "*health*"; serveis civils - "*services*"; a casa "*at home*"; altre "*other*".
Feina del pare (Fjob)	Feina del pare. Nominal: professora - "*teacher*"; relacionat amb la salut - "*health*"; serveis civils - "*services*"; a casa "*at home*"; altre "*other*".
Raó (reason)	Raó per escollir l'escola. Nominal: a prop de l'escola - "*home*"; reputació de l'escola - "*reputation*"; preferència d'un curs "*course*; altre - "*other*.
Tutor (guardian)	Tutor de l'alumne. Nominal: mare - "*mother*"; pare - "*father*"; altre - "*other*".
Temps de viatge (traveltime)	Temps de viatge de l'escola a casa. Numèrica: 1 - <15 min; 2 - 15 a 30 min; 3 - 30 min - 1 h; 4 - > 1 h.
Temps d'estudi (studytime)	Temps d'estudi a la setmana. Numèrica: 1 - < 2 h; 2 - 2 a 5 h; 3 - 5 a 10 h; 4 - > 10 h.
Suspeses (failures)	Número d'assignatures suspeses. Numèrica: n si $1 \leq n < 3$; altrament 4.
Ajuda escolar (schoolsup)	Ajuda educativa extra de l'escola. Binària: sí - "*yes*"; no -"*no*".
Ajuda familiar (famsup)	Ajuda educativa de la família. Binària: sí - "*yes*"; no -"*no*".
Pagament (paid)	Classes pagades extra de l'assignatura. Binària: sí - "*yes*"; no -"*no*".
Activitats (activities)	Activitats extracurriculars. Binària: sí - "*yes*"; no -"*no*".
Guarderia (nursery)	Ha estat atès a la guarderia de l'escola. Binària: sí - "*yes*"; no -"*no*".
Superior (higher)	Vol cursar educació superior. Binària: sí - "*yes*"; no -"*no*".
Internet (internet)	Accés a internet a casa. Binària: sí - "*yes*"; no -"*no*".
Relació romàntica (romantic)	Està en una relació romàntica. Binària: sí - "*yes*"; no -"*no*".
Relació familiar (famrel)	Qualitat de la relació familiar. Numèrica: de 1 - molt dolenta a 5 - excel·lent.
Temps lliure (freetime)	Temps lliure després de l'escola. Numèrica: de 1 - molt poc a 5 - molt.
Quedar (goout)	Quedar amb els amics. Numèrica: de 1 - molt poc a 5 - molt.
Alcohol entre setmana (Dalc)	Consum d'alcohol entre setmana. Numèrica: de 1 - molt poc a 5 - molt.

Alcohol caps de setmana (Walc)	Consum d'alcohol els caps de setmana. Numèrica: de 1 - molt poc a 5 - molt.
Salud (health)	Salud actual de l'alumne. Numèrica: de 1 - molt dolenta a 5 - molt bona.
Absències (absences)	Nombre d'absències de l'assignatura. Numèrica: de 0 a 93.
Notes 1r trimestre (G1)	Notes del primer trimestre. Numèrica: de 0 a 20.
Notes 2n trimestre (G2)	Notes del segon trimestre. Numèrica: de 0 a 20.
Notes finals (G3)	Notes finals. Numèrica: de 0 a 20.

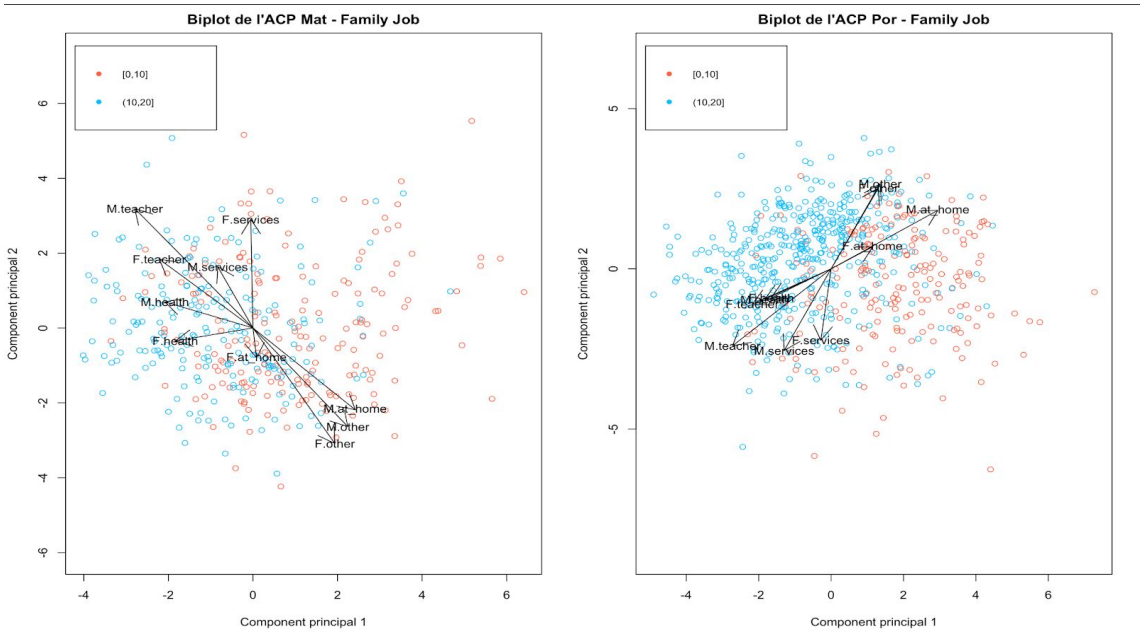
Taula 1. Descripció de les variables proporcionades per la base de dades.

2. Visualització

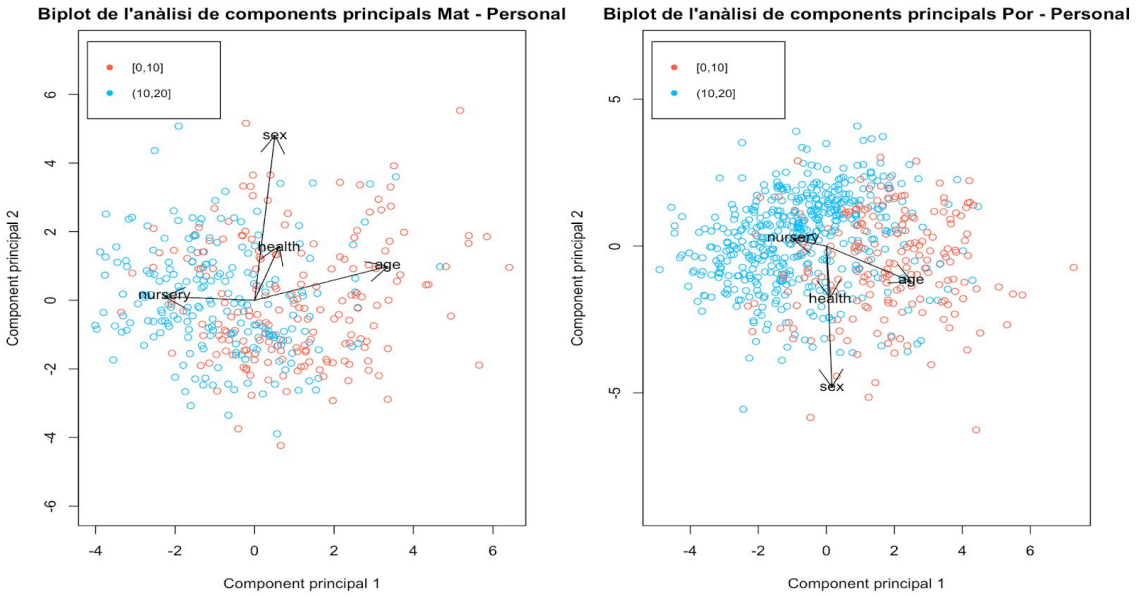
2.1 Anàlisi de components principals



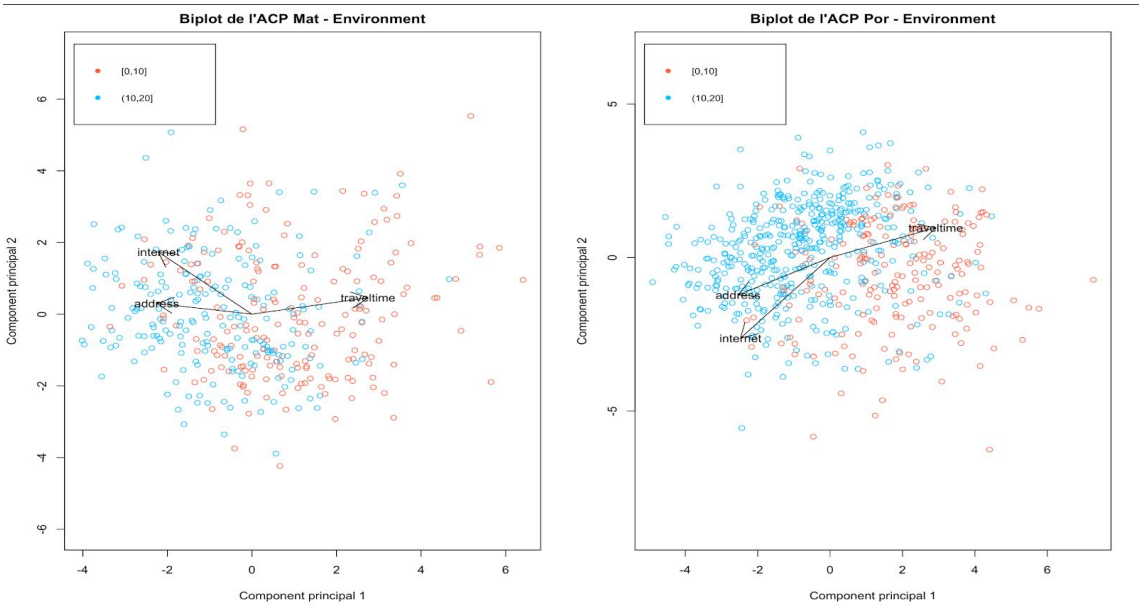
Imatge 1. ACP de les notes de Matemàtiques amb la relació familiar
Imatge 2. ACP de les notes de Portugués amb la relació familiar



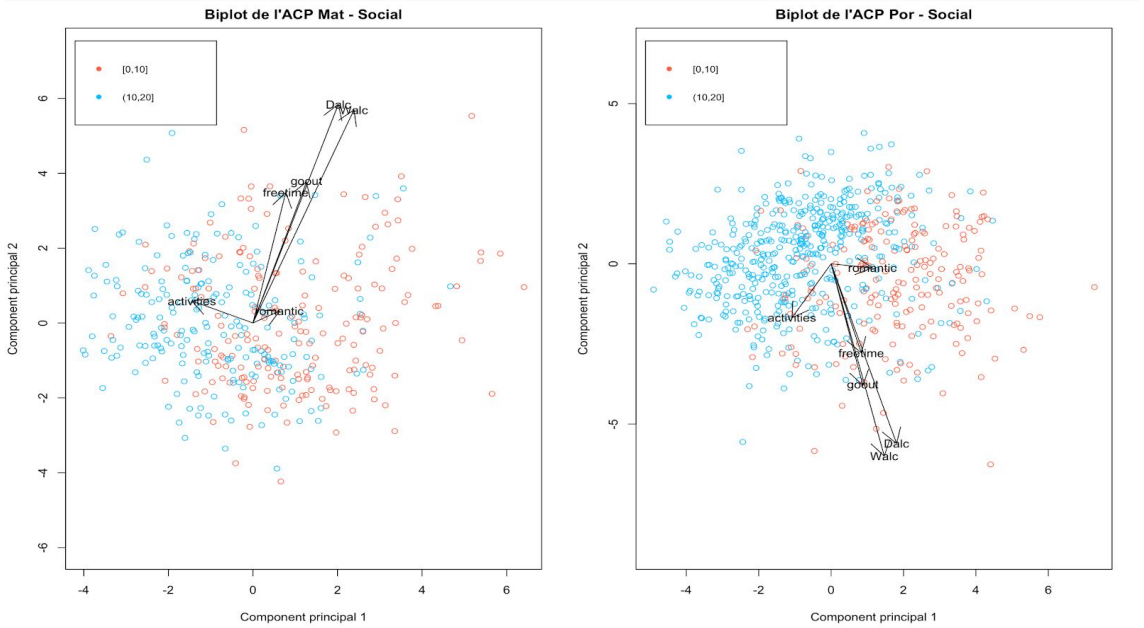
Imatge 3. ACP de les notes de Matemàtiques amb la relació de la professió dels pares
Imatge 4. ACP de les notes de Portugués amb la relació de la professió dels pares



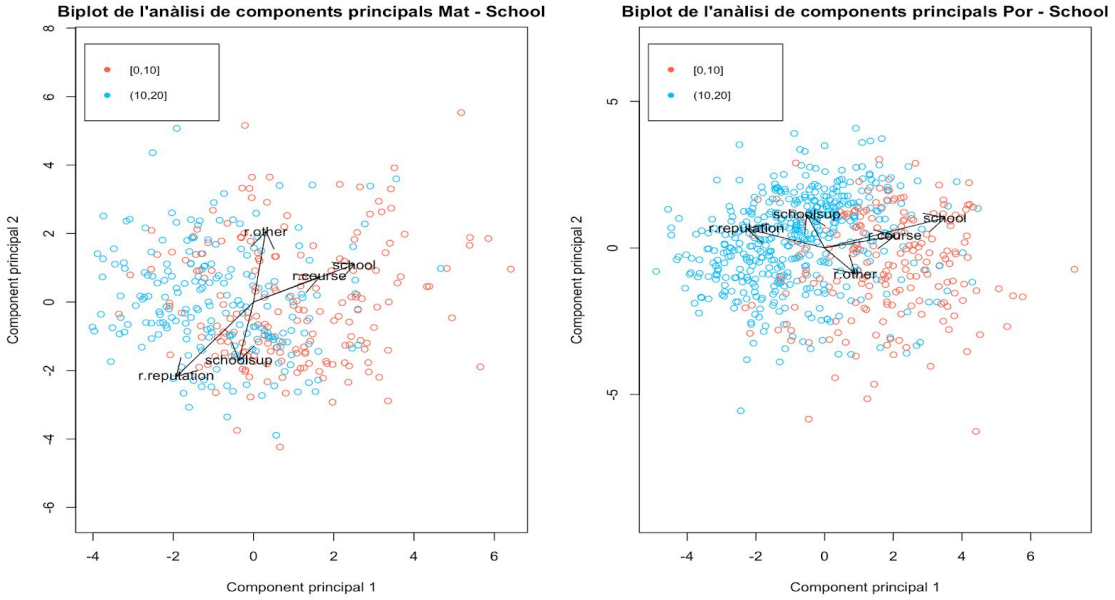
Imatge 5. ACP de les notes de Matemàtiques amb l'àmbit personal de l'alumne
Imatge 6. ACP de les notes de Portugués amb l'àmbit personal de l'alumne



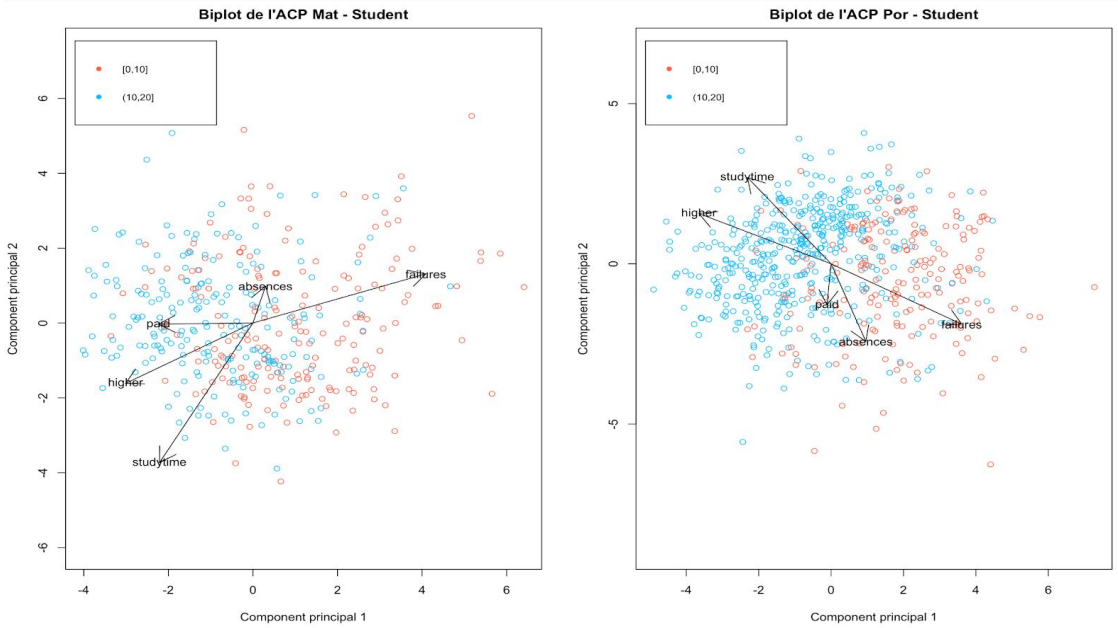
Imatge 7. ACP de les notes de Matemàtiques amb la relació de l'entorn de l'alumne
Imatge 8. ACP de les notes de Portugués amb la relació de l'entorn de l'alumne



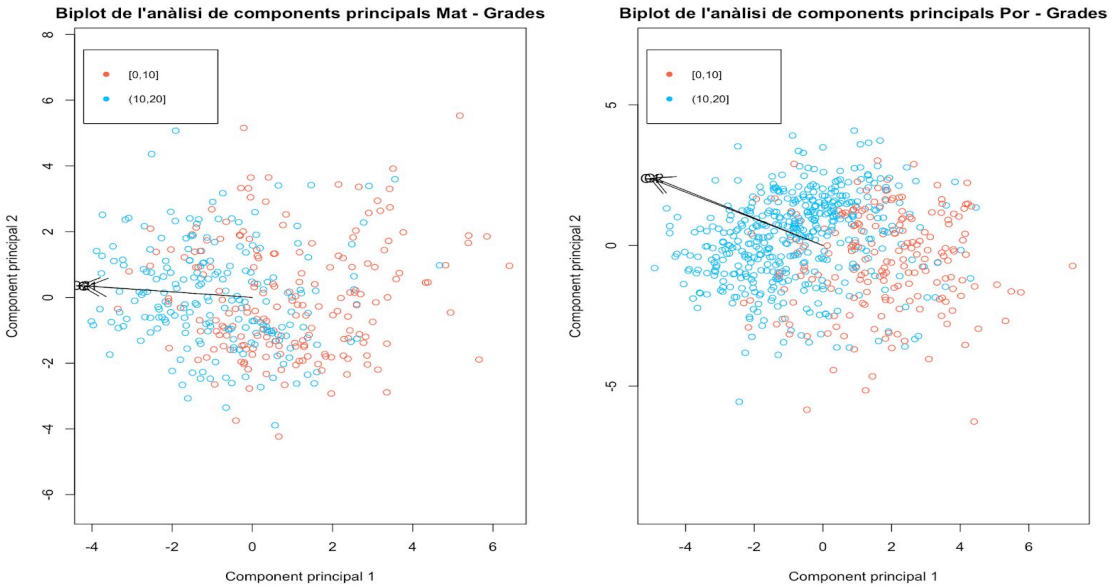
Imatge 9. ACP de les notes de Matemàtiques amb la relació de l'àmbit de l'alumne
Imatge 10. ACP de les notes de Portugués amb la relació de l'àmbit de l'alumne



Imatge 11. ACP de les notes de Matemàtiques amb la relació de escolar de l'alumne
Imatge 12. ACP de les notes de Portugués amb la relació de escolar de l'alumne

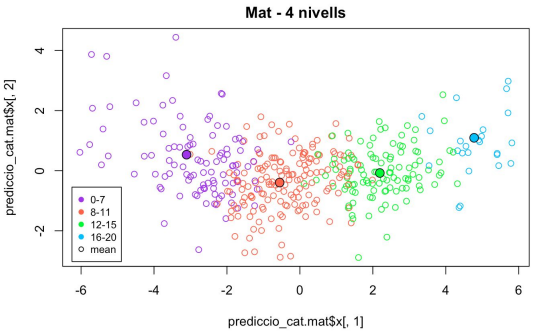


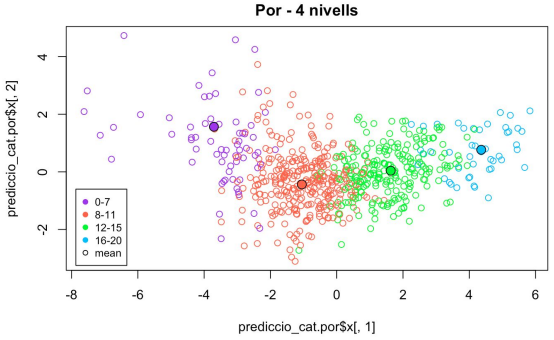
Imatge 13. ACP de les notes de Matemàtiques amb la relació de l'estudiant
Imatge 14. ACP de les notes de Portugués amb la relació de l'estudiant



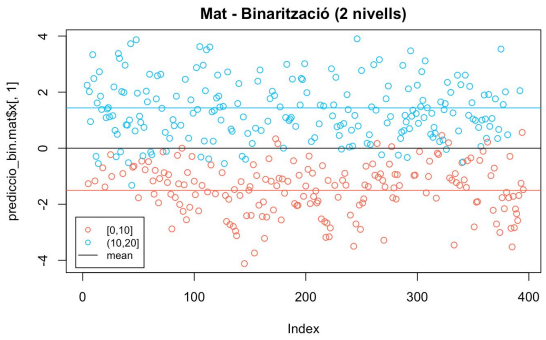
Imatge 15. ACP de les notes de Matemàtiques amb avaluacions anteriors de l'estudiant
Imatge 16. ACP de les notes de Portugués amb avaluacions anteriors de l'estudiant

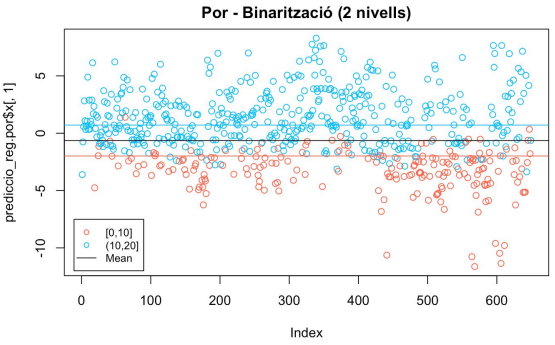
2.2 Anàlisi discriminant lineal





Imatge 17. LDA de les dades 4 nivells de qualificacions de les notes de Matemàtiques
Imatge 18. LDA de les dades 4 nivells de qualificacions de les notes de Portugués





Imatge 19. LDA de les dades 2 nivells de qualificacions de les notes de Matemàtiques
Imatge 20. LDA de les dades 2 nivells de qualificacions de les notes de Portugués

3. Protocol de re-mostreig i models considerats

Matemàtiques	Nivells	Train (%)	Test (%)
2 grups	[0,10)	49.66216	45.60811
	(10,20]	50.33784	54.39189
4 grups	[0,7]	26.35135	24.66216
	[8,11]	41.89189	41.89189
	[12,15]	25.33784	29.05405
	[16,20]	6.418919	4.391892

Taula 2. Taula on s'especifica la proporció de cada grup a les bases de dades dividides de les notes de matemàtiques

Portugués	Nivells	Train (%)	Test (%)
2 grups	[0,10)	30.8642	30.04115
	(10,20]	69.1358	69.95885
4 grups	[0,7]	9.259259	8.847737
	[8,11]	47.73663	47.73663
	[12,15]	36.21399	36.21399
	[16,20]	6.790123	7.201646

Taula 3. Taula on s'especifica la proporció de cada grup a les bases de dades dividides de les notes de portugués

4. Resultats preliminars i comparació

4.1 K-Nearest Neighbours

k	Precisió
1	95.94595
2	92.22973
3	92.22973
4	93.24324
5	92.22973
6	90.87838
7	91.55405
8	92.90541
9	91.55405
10	91.89189

Taula 4. Cerca del nombre k òptim per les notes en 2 categories de matemàtiques

k	Precisió
1	90.54054
2	81.75676
3	82.77027
4	83.78378
5	83.44595
6	82.77027
7	83.44595
8	82.43243
9	82.43243
10	82.09459

Taula 5. Cerca del nombre k òptim per les notes en 4 categories de matemàtiques

k	MSE
1	1.584459
2	3.131757
3	3.608108
4	4.415541
5	4.300676
6	4.618243
7	4.496622
8	4.070946
9	4.743243
10	4.621622

Taula 6. Cerca del nombre k òptim per les notes numèriques de matemàtiques

k	Precisió
1	94.85597
2	89.09465
3	90.94650
4	88.88889
5	88.88889
6	88.68313
7	90.94650
8	89.09465
9	90.12346
10	90.12346

Taula 7. Cerca del nombre k òptim per les notes en 2 categories de portugués

k	Precisió
1	90.94650
2	83.74486
3	84.56790

4	82.09877
5	81.68724
6	80.45267
7	82.30453
8	82.09877
9	81.68724
10	82.92181

Taula 8. Cerca del nombre k òptim per les notes en 4 categories de portugués

k	MSE
1	0.9465021
2	2.6502058
3	3.4979424
4	3.7386831
5	2.4238683
6	1.8292181
7	1.8703704
8	1.8168724
9	1.8456790
10	1.9259259

Taula 9. Cerca del nombre k òptim per les notes numèriques de portugués

4.2 Xarxa Neuronal

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	2	0	0.7989573	0.5969527	0.16933999	0.34044117
2	4	0	0.8703342	0.7406122	0.10659893	0.21308945
3	6	0	0.8934187	0.7868432	0.07500670	0.14998416
4	8	0	0.8972537	0.7945060	0.05206852	0.10402766
5	10	0	0.9002184	0.8004558	0.04883870	0.09762839
6	12	0	0.8925985	0.7852441	0.06135266	0.12256288
7	14	0	0.9025985	0.8051885	0.04910724	0.09820785

8	16	0	0.8959548	0.7918765	0.05459298	0.10915882
9	18	0	0.8973695	0.7947114	0.05508484	0.11008350
10	20	0	0.8952644	0.7904524	0.05374361	0.10751861

Taula 10. Cross-Validation cerca del valor size Matemàtiques 2 categories

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	20	0.01000000	0.8976749	0.7953439	0.05086179	0.10164973
2	20	0.01584893	0.8990197	0.7980034	0.05236601	0.10470908
3	20	0.02511886	0.8997094	0.7993911	0.04947775	0.09895335
4	20	0.03981072	0.8989967	0.7979719	0.05025575	0.10050644
5	20	0.06309573	0.8983760	0.7967310	0.04967813	0.09934396
6	20	0.10000000	0.8966404	0.7932414	0.04984792	0.09970309
7	20	0.15848932	0.9033760	0.8067147	0.04880654	0.09762240
8	20	0.25118864	0.9013530	0.8026833	0.04872474	0.09743626
9	20	0.39810717	0.9060542	0.8120668	0.04789480	0.09578688
10	20	0.63095734	0.9121125	0.8241975	0.04535719	0.09070689

Taula 11. Cross-Validation cerca del valor decay Matemàtiques 2 categories

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	2	0	0.6394291	0.4440492	0.14347470	0.2497498
2	4	0	0.6855337	0.5296401	0.11869183	0.1982269
3	6	0	0.6933374	0.5532335	0.08774317	0.1258062
4	8	0	0.7014168	0.5650975	0.09408575	0.1365361
5	10	0	0.7037091	0.5703083	0.08306617	0.1198994
6	12	0	0.7076058	0.5746172	0.08316332	0.1198502
7	14	0	0.7082938	0.5756688	0.07976520	0.1157802
8	16	0	0.7132679	0.5818593	0.08407364	0.1226160
9	18	0	0.7251747	0.5998539	0.08048032	0.1186761
10	20	0	0.7287996	0.6044626	0.07856598	0.1152094

Taula 12. Cross-Validation cerca del valor size Matemàtiques 4 categories

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	20	0.01000000	0.7588379	0.6475969	0.07266746	0.10730945
2	20	0.01584893	0.7599454	0.6484375	0.06831039	0.10078634
3	20	0.02511886	0.7565433	0.6445881	0.06483752	0.09415118
4	20	0.03981072	0.7643109	0.6553819	0.07251001	0.10727545

5	20	0.06309573	0.7643454	0.6547830	0.07538812	0.11088908
6	20	0.10000000	0.7668239	0.6585776	0.06979566	0.10385328
7	20	0.15848932	0.7612110	0.6501927	0.07260961	0.10726037
8	20	0.25118864	0.7665128	0.6578779	0.06694687	0.09928190
9	20	0.39810717	0.7654324	0.6566984	0.07298203	0.10744627
10	20	0.63095734	0.7736910	0.6684819	0.07303217	0.10750140

Taula 13. Cross-Validation cerca del valor decay Matemàtiques 4 categories

	size	decay	RMSE	Rsquared	MAE	RMSESD	RsquaredSD
1	2	0	4.19561	0.5673465	3.139921	0.978627	0.2638650
2	4	0	3.42870	0.6067587	2.551510	1.165848	0.2467151
3	6	0	3.00793	0.6365900	2.196542	1.029388	0.2184388
4	8	0	2.82246	0.6920288	1.985896	1.059654	0.1467127
5	10	0	2.69116	0.6997526	1.906855	0.832857	0.1696825
6	12	0	2.71933	0.6892884	1.944209	0.762933	0.1396638
7	14	0	2.59199	0.7254434	1.852432	0.767184	0.1445042
8	16	0	2.64366	0.7199852	1.893762	0.833539	0.1436234
9	18	0	2.73906	0.7080536	1.948794	0.968071	0.1567495
10	20	0	2.54555	0.7277836	1.822806	0.712440	0.1417828

Taula 14. Cross-Validation cerca del valor size Matemàtiques regressió

	size	decay	RMSE	Rsquared	MAE	RMSESD	Rsquared SD
1	20	0.01000000	2.070055	0.8137803	1.50769	0.4918	0.0894048
2	20	0.01584893	2.048871	0.8216886	1.49771	0.4344	0.0713643
3	20	0.02511886	1.968542	0.8297336	1.44199	0.4290	0.0690436
4	20	0.03981072	1.950366	0.8328121	1.42599	0.3985	0.0701977
5	20	0.06309573	1.914343	0.8358329	1.40281	0.3927	0.0643512
6	20	0.10000000	1.895642	0.8396761	1.37681	0.4118	0.0645050
7	20	0.15848932	1.928095	0.8360385	1.41359	0.3930	0.0617041
8	20	0.25118864	1.850029	0.8468461	1.35274	0.3907	0.0590907
9	20	0.39810717	1.803141	0.8519836	1.32008	0.3904	0.0620487

10	20	0.63095734	1.778126	0.8566487	1.28774	0.3935	0.0591382

Taula 15. Cross-Validation cerca del valor decay Matemàtiques regressió

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	2	0	0.8075553	0.4704819	0.08710001	0.3226549
2	4	0	0.8478614	0.6263557	0.05858714	0.1883178
3	6	0	0.8588265	0.6661834	0.04845333	0.1139940
4	8	0	0.8523682	0.6499024	0.04859411	0.1148996
5	10	0	0.8569260	0.6650518	0.04827243	0.1122193
6	12	0	0.8511522	0.6505954	0.05018244	0.1139434
7	14	0	0.8530782	0.6533143	0.04587119	0.1090611
8	16	0	0.8526956	0.6531790	0.04701917	0.1055772
9	18	0	0.8512840	0.6480108	0.04808734	0.1143705
10	20	0	0.8545111	0.6561311	0.04755185	0.1112279

Taula 16. Cross-Validation cerca del valor size Português 2 categories

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	20	0.01000000	0.8651956	0.6783063	0.05149788	0.1283211
2	20	0.01584893	0.8666199	0.6832626	0.04853983	0.1170521
3	20	0.02511886	0.8647619	0.6789803	0.04776036	0.1162795
4	20	0.03981072	0.8688733	0.6870877	0.05064735	0.1227982
5	20	0.06309573	0.8666582	0.6830884	0.04859493	0.1178362
6	20	0.10000000	0.8678571	0.6855914	0.04706845	0.1153459
7	20	0.15848932	0.8665859	0.6824296	0.04805154	0.1159241
8	20	0.25118864	0.8647662	0.6771208	0.04589849	0.1130512
9	20	0.39810717	0.8668325	0.6812245	0.04770216	0.1174035
10	20	0.63095734	0.8737968	0.6978951	0.04652610	0.1142313

Taula 17. Cross-Validation cerca del valor decay Português 2 categories

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	2	0	0.6832524	0.4354395	0.16153765	0.32611678
2	4	0	0.7012997	0.4978650	0.11373211	0.22638605
3	6	0	0.7235714	0.5567544	0.06983814	0.12302151
4	8	0	0.7077025	0.5341828	0.07137568	0.11999187
5	10	0	0.7126856	0.5418162	0.06292844	0.11520046

6	12	0	0.7295777	0.5704013	0.06338491	0.09919030
7	14	0	0.7102092	0.5421825	0.06635180	0.10387328
8	16	0	0.7191900	0.5554706	0.05861322	0.09154266
9	18	0	0.7116919	0.5449769	0.06274273	0.09671363
10	20	0	0.7183605	0.5536575	0.06455851	0.09863293

Taula 18. Cross-Validation cerca del valor size Português 4 categories

	size	decay	Accuracy	Kappa	AccuracySD	KappaSD
1	20	0.01000000	0.7735636	0.6383745	0.05941098	0.09345328
2	20	0.01584893	0.7840053	0.6545988	0.06357684	0.10096243
3	20	0.02511886	0.7798935	0.6474842	0.06047046	0.09467423
4	20	0.03981072	0.7796554	0.6472253	0.06243079	0.09769751
5	20	0.06309573	0.7838681	0.6537135	0.06287948	0.09845451
6	20	0.10000000	0.7858468	0.6565555	0.05797722	0.09208117
7	20	0.15848932	0.7832065	0.6526869	0.05821800	0.09097852
8	20	0.25118864	0.7854468	0.6560287	0.05766312	0.09030119
9	20	0.39810717	0.7848457	0.6537350	0.06532767	0.10498134
10	20	0.63095734	0.7833290	0.6509771	0.06088218	0.09711045

Taula 19. Cross-Validation cerca del valor decay Português 4 categories

	size	decay	RMSE	Rsquared	MAE	RMSESD	Rsquared SD
1	2	0	2.482180	0.6476589	1.822793	0.8910304	0.2676987
2	4	0	2.308437	0.6291631	1.455735	1.3592103	0.2441777
3	6	0	2.056832	0.6538884	1.305256	1.2908315	0.2418223
4	8	0	2.646562	0.6329609	1.454679	3.3968883	0.2086122
5	10	0	2.266064	0.6484332	1.393668	1.4775126	0.1954040
6	12	0	2.528844	0.6342059	1.520474	2.0287548	0.1744712
7	14	0	2.418876	0.6013425	1.515747	1.2605486	0.1972295
8	16	0	2.424308	0.6137236	1.543309	0.8942104	0.1597443
9	18	0	2.544620	0.5744840	1.632460	0.8711346	0.1603406
10	20	0	2.758527	0.5276548	1.726313	1.1838827	0.1907766

Taula 20. Cross-Validation cerca del valor size Português regressió

	size	decay	RMSE	Rsquared	MAE	RMSESD	RsquaredSD
1	20	0.0100000	2.10655	0.6496166	1.50587	0.3580	0.09988789

2	20	0.0158489	2.00895	0.6693284	1.43170	0.35383	0.09814599
3	20	0.0251188	1.95462	0.6846113	1.39010	0.35091	0.09877314
4	20	0.0398107	1.94048	0.6805222	1.37536	0.36570	0.10099843
5	20	0.0630957	1.84231	0.7076419	1.30617	0.34708	0.09128394
6	20	0.1000000	1.80408	0.7149096	1.28216	0.31530	0.08123794
7	20	0.1584893	1.75346	0.7275051	1.23839	0.34819	0.08939725
8	20	0.2511886	1.72733	0.7335085	1.21561	0.32876	0.08423823
9	20	0.3981071	1.66296	0.7497688	1.17455	0.32754	0.08299332
10	20	0.6309573	1.64943	0.7532407	1.137819	0.32721	0.08361640

Taula 21. Cross-Validation cerca del valor decay Português regressió

4.3 Random Forest

ntrees	OOB error
2	0.12000000
4	0.17142857
8	0.11267606
16	0.09152542
32	0.07094595
64	0.06756757
128	0.06418919
256	0.05405405
512	0.06418919
1024	0.06418919

Taula 22. Cerca del nombre d'arbres per les notes en 2 categories de matemàtiques

ntrees	OOB error
2	0.4545455
4	0.3764706
8	0.3321678
16	0.3344595
32	0.2263514

64	0.2229730
128	0.1925676
256	0.2027027
512	0.1891892
1024	0.1689189

Taula 23. Cerca del nombre d'arbres per les notes en 4 categories de matemàtiques

ntrees	OOB error
2	0.7709497
4	0.8008130
8	0.7132867
16	0.7162162
32	0.6722973
64	0.6655405
128	0.6587838
256	0.6182432
512	0.5810811
1024	0.6013514

Taula 24. Cerca del nombre d'arbres per les notes numèriques de matemàtiques

ntrees	OOB error
2	0.16785714
4	0.16183575
8	0.15513627
16	0.15020576
32	0.10699588
64	0.10082305
128	0.09670782
256	0.09465021
512	0.10493827
1024	0.09259259

Taula 25. Cerca del nombre d'arbres per les notes en 2 categories de portugués

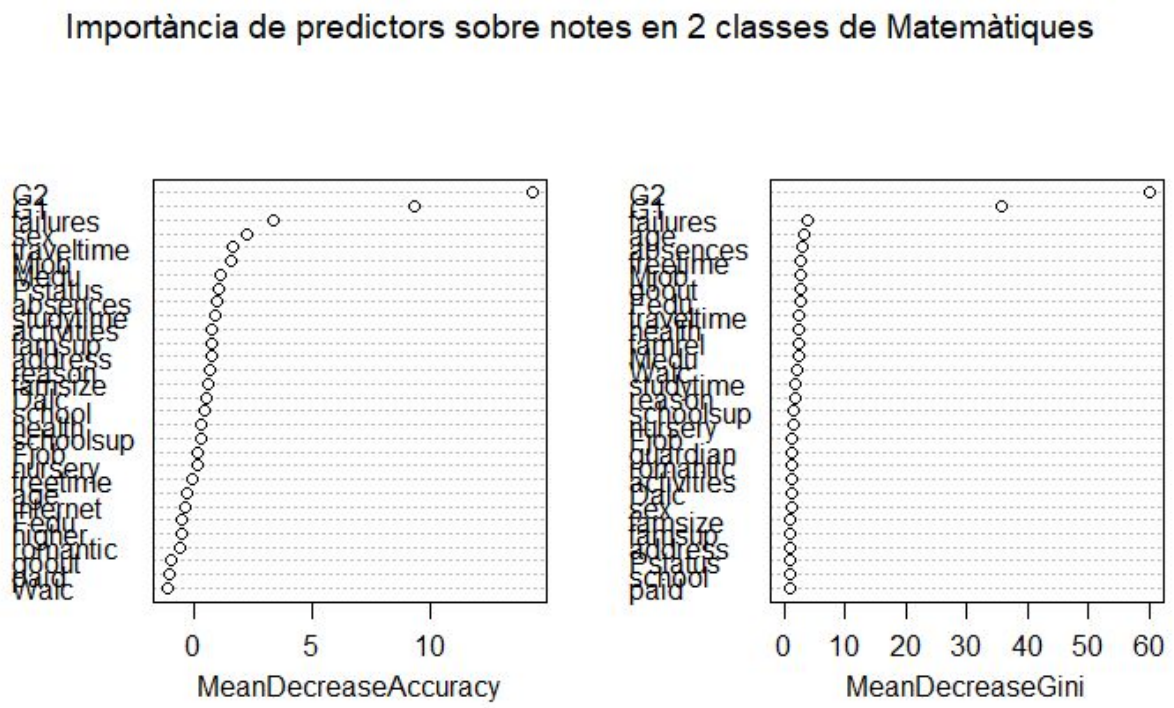
ntrees	OOB error
2	0.3401361
4	0.3569682
8	0.2521186
16	0.2078189
32	0.1872428
64	0.1790123
128	0.1851852
256	0.1790123
512	0.1831276
1024	0.1810700

Taula 26. Cerca del nombre d'arbres per les notes en 4 categories de portugués

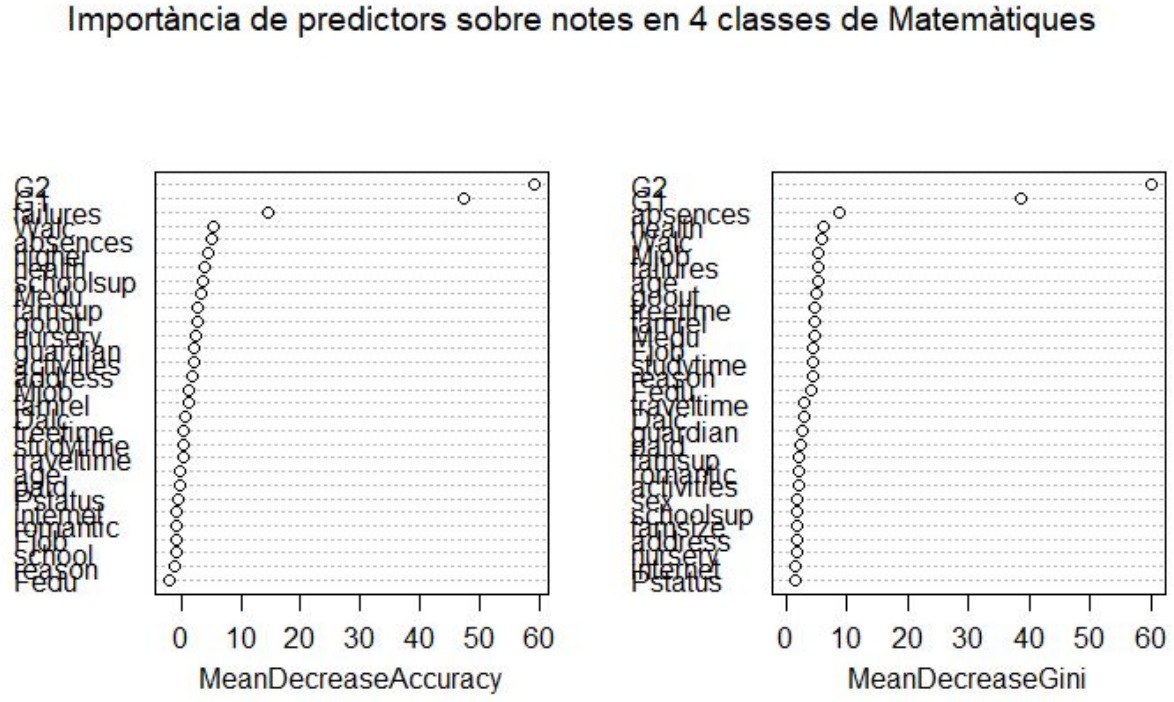
ntrees	OOB error
2	0.7517241
4	0.7865707
8	0.7172996
16	0.6975309
32	0.6563786
64	0.6378601
128	0.6378601
256	0.6193416
512	0.5864198
1024	0.6111111

Taula 27. Cerca del nombre d'arbres per les notes en numèriques de matemàtiques

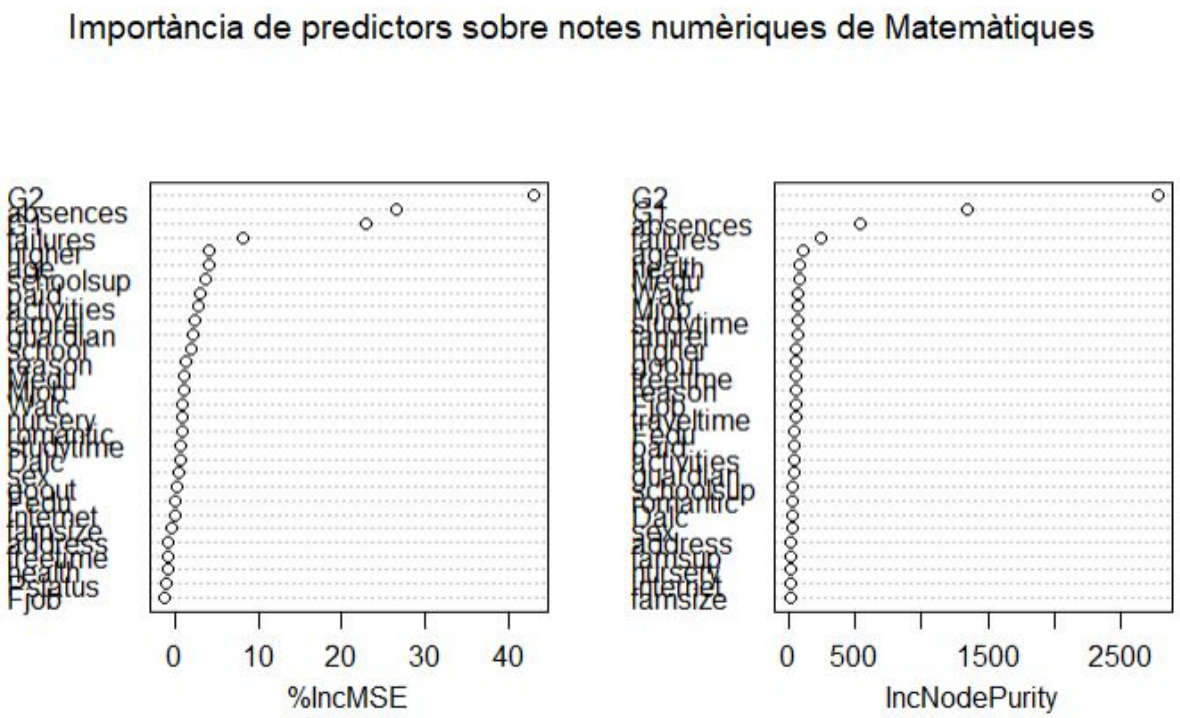
5. Test i elecció del millor model



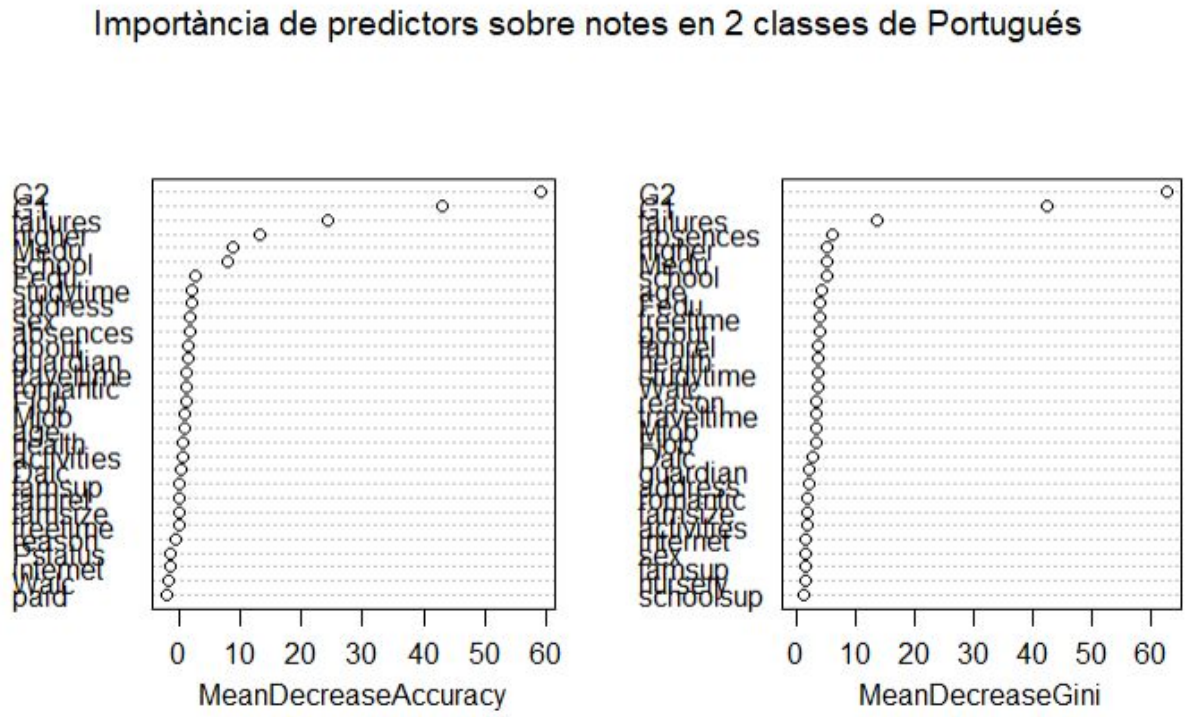
Imatge 21. Importància dels predictors sobre les notes en dues classes de matemàtiques



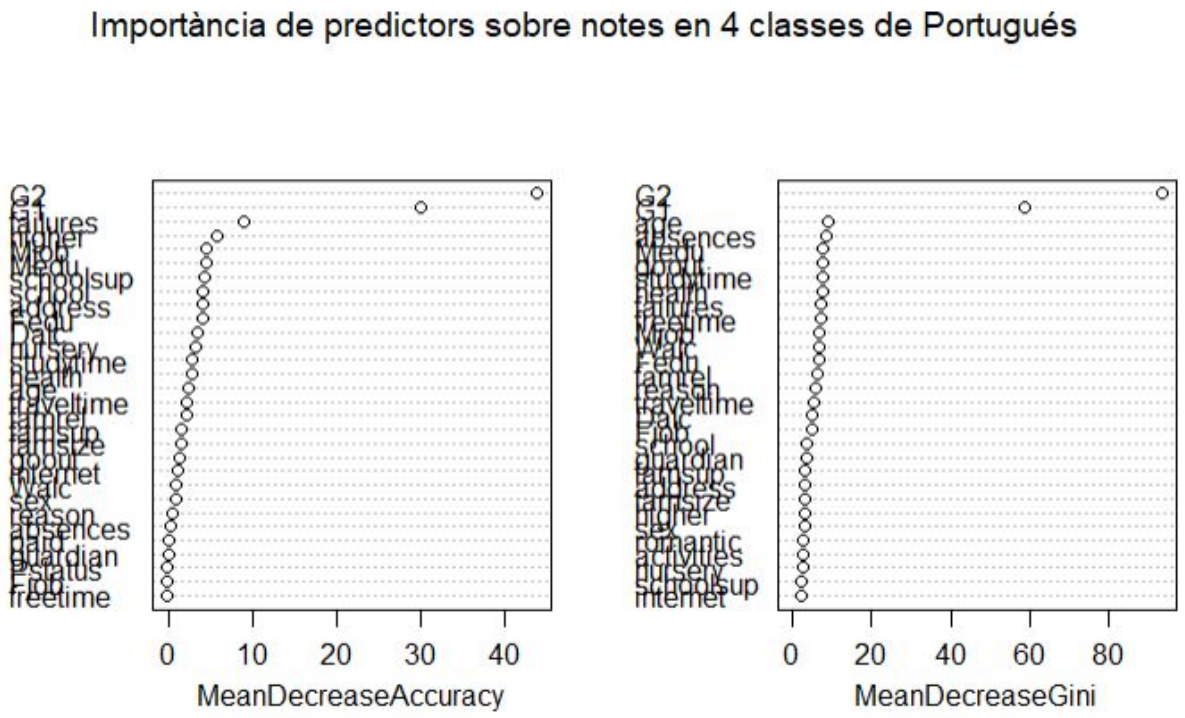
Imatge 22. Importància dels predictors sobre les notes en quatre classes de matemàtiques



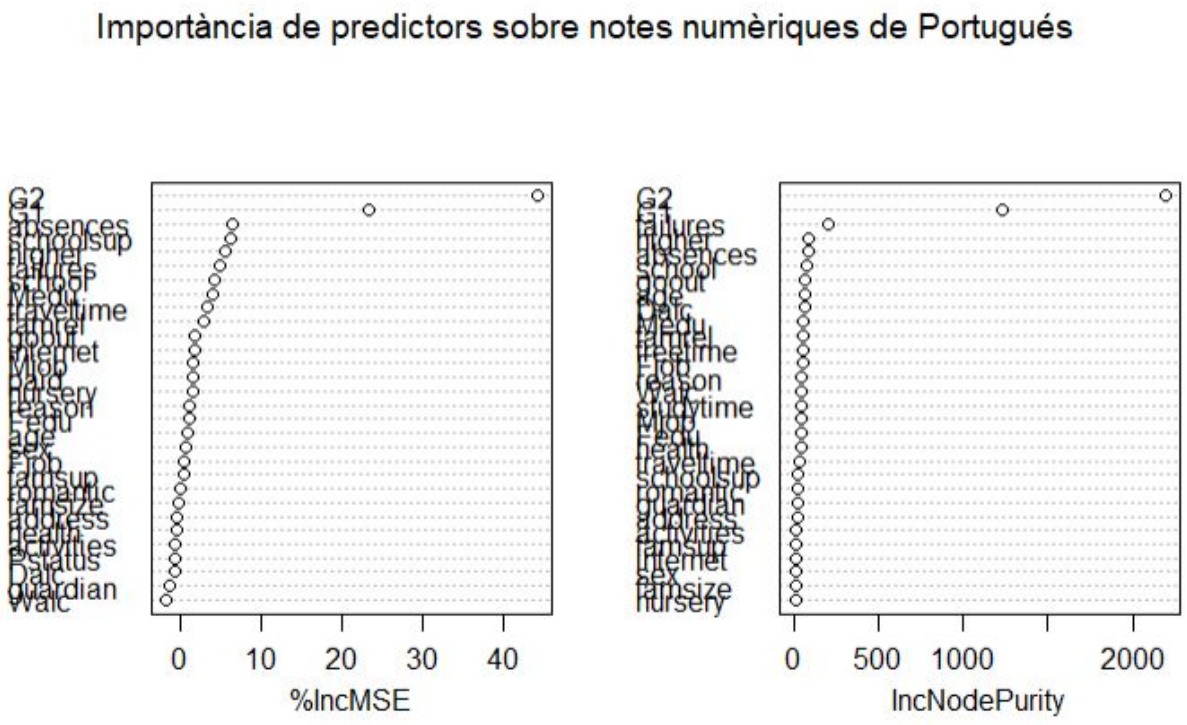
Imatge 23. Importància dels predictors sobre les notes numèriques de matemàtiques



Imatge 24. Importància dels predictors sobre les notes en dues classes de portugués



Imatge 25. Importància dels predictors sobre les notes en quatre classes de portugués



Imatge 26. Importància dels predictors sobre les notes numèriques de portugués