Relatório tarefa 3 INF1771

Professor: Augusto Baffa

Alunos: João Gabriel Cunha - 2211302
Luiz Augusto - 2210523

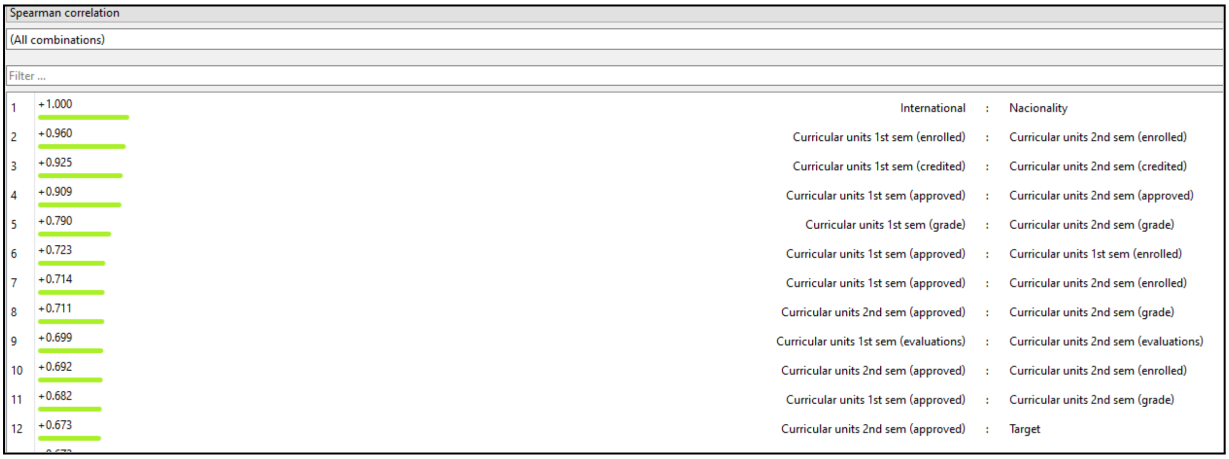
1. Introdução

O trabalho tem como objetivo analisar um dataset com informações de alunos, observando quais as principais características para um aluno abandonar os estudos. E a partir dos dados criar modelos capazes de prever a evasão com base nas informações disponíveis.

2. Metodologia

2.1. Análise exploratória dos dados

1. Vimos as distribuições para ter uma noção básica do dataset, observando as features e seus possíveis valores.
2. Filtramos os dados removendo os alunos que ainda estão cursando o curso, já que não precisamos de dados se eles vão ou não abandonar o curso, para não sujar os dados.
3. Analisamos as correlações entre as features, já pensando em remover as com alto nível de correlação, como "International" e "Nacionality" que tem correlação 1.0 na Spearman Correlation. Também já conseguimos observar algumas features que parecem ser importantes pela correlação com o target como o número de unidades aprovadas no 1 e no 2 semestre.

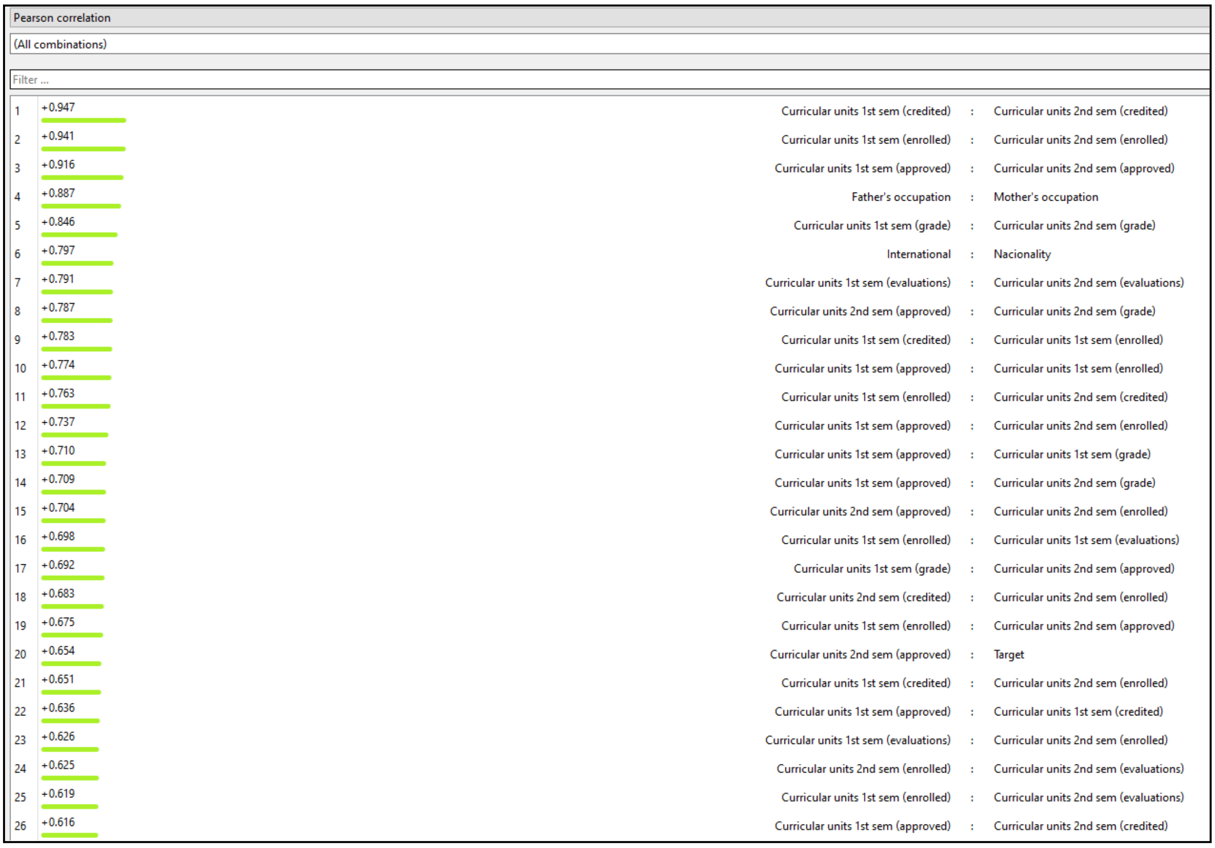


Spearman correlation

(All combinations)

Filter ...

1	+1.000	International	:	Nacionality
2	+0.960	Curricular units 1st sem (enrolled)	:	Curricular units 2nd sem (enrolled)
3	+0.925	Curricular units 1st sem (credited)	:	Curricular units 2nd sem (credited)
4	+0.909	Curricular units 1st sem (approved)	:	Curricular units 2nd sem (approved)
5	+0.790	Curricular units 1st sem (grade)	:	Curricular units 2nd sem (grade)
6	+0.723	Curricular units 1st sem (approved)	:	Curricular units 1st sem (enrolled)
7	+0.714	Curricular units 1st sem (approved)	:	Curricular units 2nd sem (enrolled)
8	+0.711	Curricular units 2nd sem (approved)	:	Curricular units 2nd sem (grade)
9	+0.699	Curricular units 1st sem (evaluations)	:	Curricular units 2nd sem (evaluations)
10	+0.692	Curricular units 2nd sem (approved)	:	Curricular units 2nd sem (enrolled)
11	+0.682	Curricular units 1st sem (approved)	:	Curricular units 2nd sem (grade)
12	+0.673	Curricular units 2nd sem (approved)	:	Target



Pearson correlation

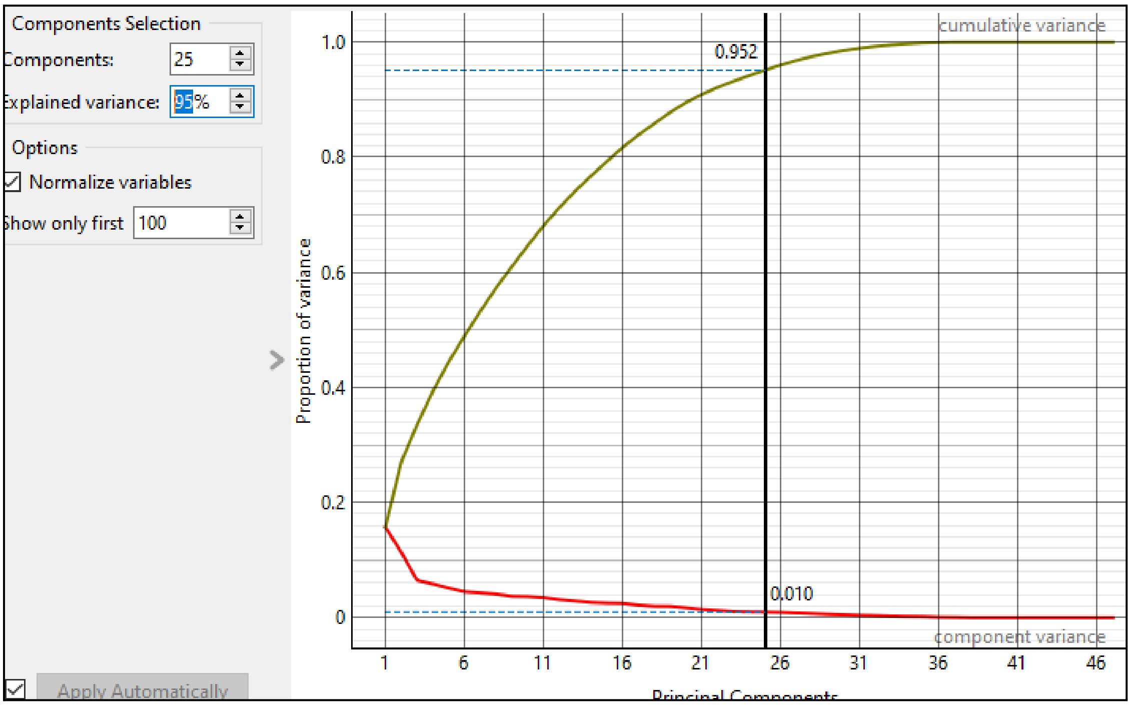
(All combinations)

Filter ...

1	+0.947	Curricular units 1st sem (credited)	:	Curricular units 2nd sem (credited)
2	+0.941	Curricular units 1st sem (enrolled)	:	Curricular units 2nd sem (enrolled)
3	+0.916	Curricular units 1st sem (approved)	:	Curricular units 2nd sem (approved)
4	+0.887	Father's occupation	:	Mother's occupation
5	+0.846	Curricular units 1st sem (grade)	:	Curricular units 2nd sem (grade)
6	+0.797	International	:	Nacionality
7	+0.791	Curricular units 1st sem (evaluations)	:	Curricular units 2nd sem (evaluations)
8	+0.787	Curricular units 2nd sem (approved)	:	Curricular units 2nd sem (grade)
9	+0.783	Curricular units 1st sem (credited)	:	Curricular units 1st sem (enrolled)
10	+0.774	Curricular units 1st sem (approved)	:	Curricular units 1st sem (enrolled)
11	+0.763	Curricular units 1st sem (enrolled)	:	Curricular units 2nd sem (credited)
12	+0.737	Curricular units 1st sem (approved)	:	Curricular units 2nd sem (enrolled)
13	+0.710	Curricular units 1st sem (approved)	:	Curricular units 1st sem (grade)
14	+0.709	Curricular units 1st sem (approved)	:	Curricular units 2nd sem (grade)
15	+0.704	Curricular units 2nd sem (approved)	:	Curricular units 2nd sem (enrolled)
16	+0.698	Curricular units 1st sem (enrolled)	:	Curricular units 1st sem (evaluations)
17	+0.692	Curricular units 1st sem (grade)	:	Curricular units 2nd sem (approved)
18	+0.683	Curricular units 2nd sem (credited)	:	Curricular units 2nd sem (enrolled)
19	+0.675	Curricular units 1st sem (enrolled)	:	Curricular units 2nd sem (approved)
20	+0.654	Curricular units 2nd sem (approved)	:	Target
21	+0.651	Curricular units 1st sem (credited)	:	Curricular units 2nd sem (enrolled)
22	+0.636	Curricular units 1st sem (approved)	:	Curricular units 1st sem (credited)
23	+0.626	Curricular units 1st sem (evaluations)	:	Curricular units 2nd sem (enrolled)
24	+0.625	Curricular units 2nd sem (enrolled)	:	Curricular units 2nd sem (evaluations)
25	+0.619	Curricular units 1st sem (enrolled)	:	Curricular units 2nd sem (evaluations)
26	+0.616	Curricular units 1st sem (approved)	:	Curricular units 2nd sem (credited)

2.2. Seleção de atributos

A partir das correlações e dos dados obtidos pelo PCA (usamos 95%) fomos removendo os atributos que acrescentam pouco. Analisamos o PCA com base nos valores que os atributos tinham nos componentes do PCA, principalmente nos primeiros componentes do PCA (PC1, PC2, PC3, ...).



Components Selection

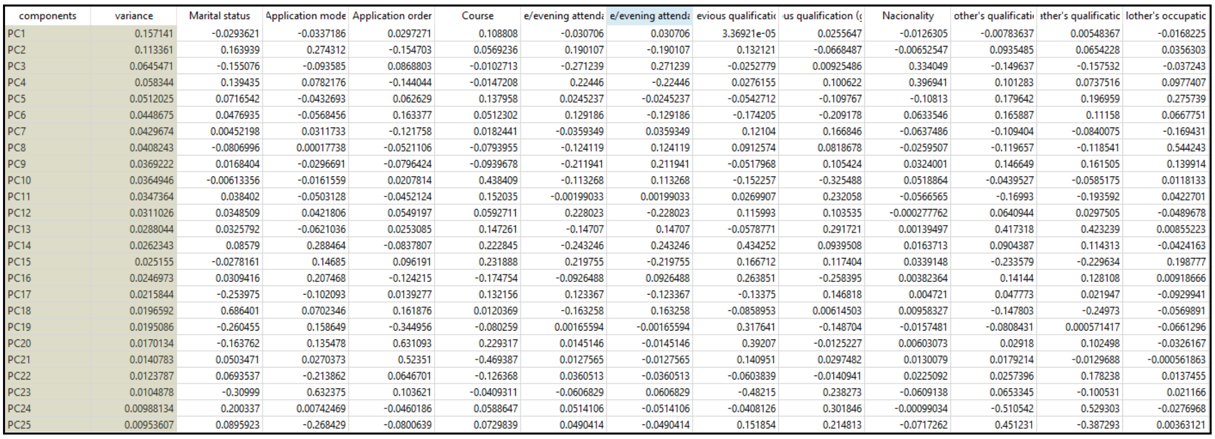
Components: 25

Explained variance: 95%

Options

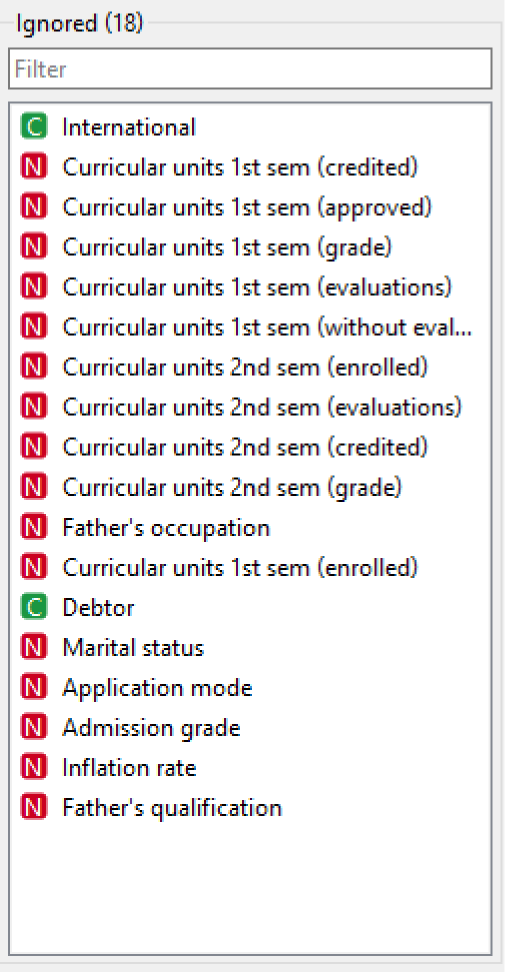
☑ Normalize variables

Show only first 100



components	variance	Marital status	Application mode	Application order	Course	e/evening attendа	e/evening attendа	evious qualificati	us qualification (	Nacionality	other's qualificati	ather's qualificatic	lother's occupatic
PC1	0.157141	-0.0293621	-0.0337186	0.0297271	0.108808	-0.030706	0.030706	3.36921e-05	0.0255647	-0.0126305	-0.00783637	0.00548367	-0.0168225
PC2	0.113361	0.163939	0.274312	-0.154703	0.0569236	0.190107	-0.190107	0.132121	-0.0668487	-0.00652547	0.0935485	0.0654228	0.0356303
PC3	0.0645471	-0.155076	-0.093585	0.0868803	-0.0102713	-0.271239	0.271239	-0.0252779	0.00925486	0.334049	-0.149637	-0.157532	-0.037243
PC4	0.058344	0.139435	0.0782176	-0.144044	-0.0147208	0.22446	-0.22446	0.0276155	0.100622	0.396941	0.101283	0.0737516	0.0977407
PC5	0.0512025	0.0716542	-0.0432693	0.062629	0.137958	0.0245237	-0.0245237	-0.0542712	-0.109767	-0.10813	0.179642	0.196959	0.275739
PC6	0.0448675	0.0476935	-0.0568456	0.163377	0.0512302	0.129186	-0.129186	-0.174205	-0.209178	0.0633546	0.165887	0.11158	0.0667751
PC7	0.0429674	0.00452198	0.0311733	-0.121758	0.0182441	-0.0359349	0.0359349	0.12104	0.166846	-0.0637486	-0.109404	-0.0840075	-0.169431
PC8	0.0408243	-0.0806996	0.00017738	-0.0521106	-0.0793955	-0.124119	0.124119	0.0912574	0.0818678	-0.0259507	-0.119657	-0.118541	0.544243
PC9	0.0369222	0.0168404	-0.0296691	-0.0796424	-0.0939678	-0.211941	0.211941	-0.0517968	0.105424	0.0324001	0.146649	0.161505	0.139914
PC10	0.0364946	-0.00613356	-0.0161559	0.0207814	0.438409	-0.113268	0.113268	-0.152257	-0.325488	0.0518864	-0.0439527	-0.0585175	0.0118133
PC11	0.0347364	0.038402	-0.0503128	-0.0452124	0.152035	-0.00199033	0.00199033	0.0269907	0.232058	-0.0566565	-0.16993	-0.193592	0.0422701
PC12	0.0311026	0.0348509	0.0421806	0.0549197	0.0592711	0.228023	-0.228023	0.115993	0.103535	-0.000277762	0.0640944	0.0297505	-0.0489678
PC13	0.0288044	0.0325792	-0.0621036	0.0253085	0.147261	-0.14707	0.14707	-0.0578771	0.291721	0.00139497	0.417318	0.423239	0.00855223
PC14	0.0262343	0.08579	0.288464	-0.0837807	0.222845	-0.243246	0.243246	0.434252	0.0939508	0.0163713	0.0904387	0.114313	-0.0424163
PC15	0.025155	-0.0278161	0.14685	0.096191	0.231888	0.219755	-0.219755	0.166712	0.117404	0.0339148	-0.233579	-0.229634	0.198777
PC16	0.0246973	0.0309416	0.207468	-0.124215	-0.174754	-0.0926488	0.0926488	0.263851	-0.258395	0.00382364	0.14144	0.128108	0.00918666
PC17	0.0215844	-0.253975	-0.102093	0.0139277	0.132156	0.123367	-0.123367	-0.13375	0.146818	0.004721	0.047773	0.021947	-0.0929941
PC18	0.0196592	0.686401	0.0702346	0.161876	0.0120369	-0.163258	0.163258	-0.0858953	0.00614503	0.00958327	-0.147803	-0.24973	-0.0569891
PC19	0.0195086	-0.260455	0.158649	-0.344956	-0.080259	0.00165594	-0.00165594	0.317641	-0.148704	-0.0157481	-0.0808431	0.000571417	-0.0661296
PC20	0.0170134	-0.163762	0.135478	0.631093	0.229317	0.0145146	-0.0145146	0.39207	-0.0125227	0.00603073	0.02918	0.102498	-0.0326167
PC21	0.0140783	0.0503471	0.0270373	0.52351	-0.469387	0.0127565	-0.0127565	0.140951	0.0297482	0.0130079	0.0179214	-0.0129688	-0.000561863
PC22	0.0123787	0.0693537	-0.213862	0.0646701	-0.126368	0.0360513	-0.0360513	-0.0603839	-0.0140941	0.0225092	0.0257396	0.178238	0.0137455
PC23	0.0104878	-0.30999	0.632375	0.103621	-0.0409311	-0.0606829	0.0606829	-0.48215	0.238273	-0.0609138	0.0653345	-0.100531	0.021166
PC24	0.00988134	0.200337	0.00742469	-0.0460186	0.0588647	0.0514106	-0.0514106	-0.0408126	0.301846	-0.00099034	-0.510542	0.529303	-0.0276968
PC25	0.00953607	0.0895923	-0.268429	-0.0800639	0.0729839	0.0490414	-0.0490414	0.151854	0.214813	-0.0717262	0.451231	-0.387293	0.00363121

A partir da análise de ambos filtramos os seguintes atributos inicialmente:



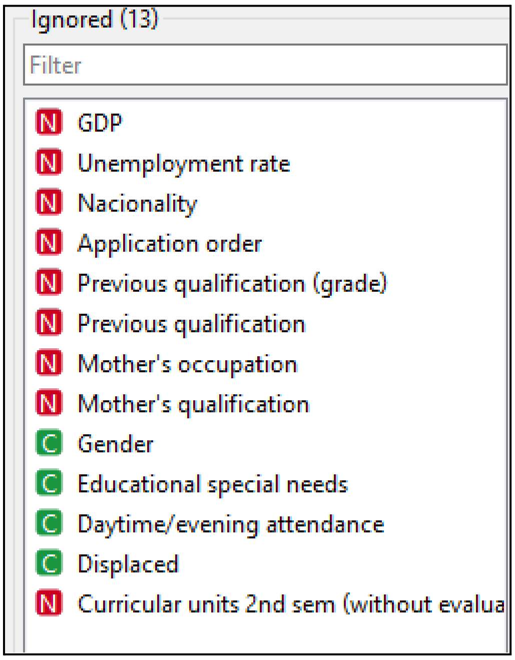
Ignored (18)

Filter

- C International
- N Curricular units 1st sem (credited)
- N Curricular units 1st sem (approved)
- N Curricular units 1st sem (grade)
- N Curricular units 1st sem (evaluations)
- N Curricular units 1st sem (without eval...
- N Curricular units 2nd sem (enrolled)
- N Curricular units 2nd sem (evaluations)
- N Curricular units 2nd sem (credited)
- N Curricular units 2nd sem (grade)
- N Father's occupation
- N Curricular units 1st sem (enrolled)
- C Debtor
- N Marital status
- N Application mode
- N Admission grade
- N Inflation rate
- N Father's qualification

Então repetimos novamente os passos de ver as correlações e os resultados do PCA para analisar com uma quantidade reduzida de atributos, facilitando a análise dos atributos restantes.

Com isso filtramos mais alguns atributos:

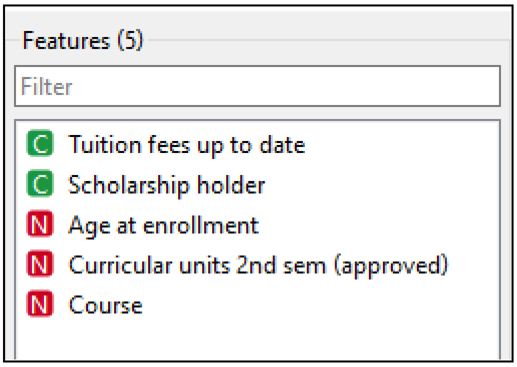


Ignored (13)

Filter

- N GDP
- N Unemployment rate
- N Nacionality
- N Application order
- N Previous qualification (grade)
- N Previous qualification
- N Mother's occupation
- N Mother's qualification
- C Gender
- C Educational special needs
- C Daytime/evening attendance
- C Displaced
- N Curricular units 2nd sem (without evalua

E ficamos com os seguintes atributos, sendo eles os mais importantes para determinar se o aluno irá concluir os estudos ou os abandonará no meio, de acordo com a nossa análise.



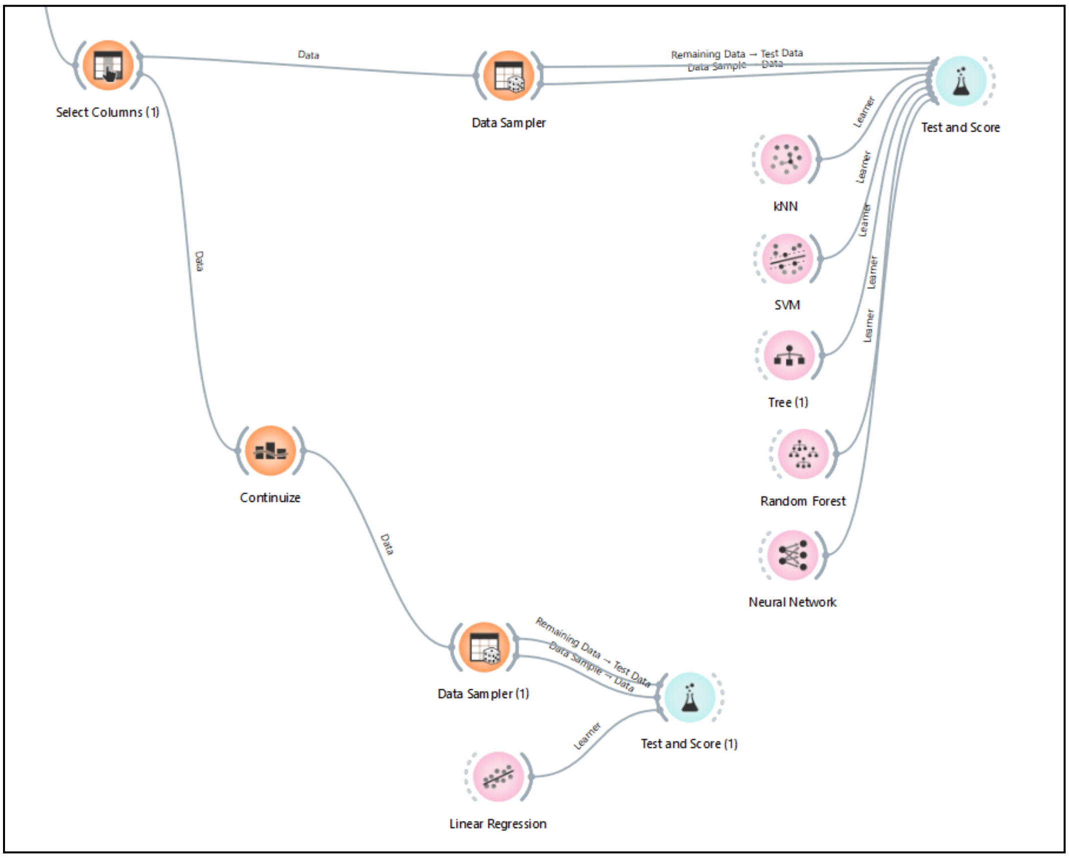
Features (5)

Filter

- C Tuition fees up to date
- C Scholarship holder
- N Age at enrollment
- N Curricular units 2nd sem (approved)
- N Course

2.3. Criação dos modelos

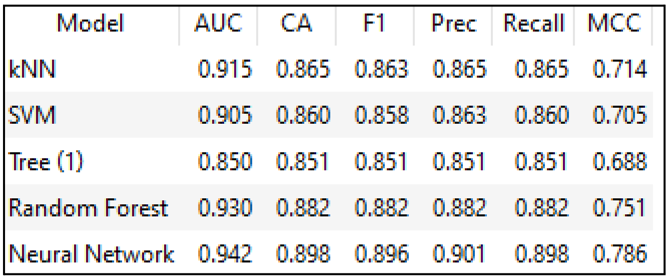
Usamos o data sampler para separar os dados para treino e teste, usando as features escolhidas. Criamos os seguintes modelos:



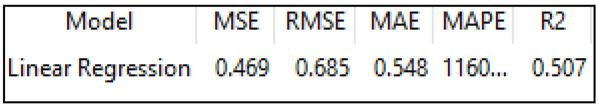
Fomos adaptando os parâmetros dos modelos em busca dos melhores resultados, já imaginando que modelos lineares não seriam muito bons pela natureza dos dados.

3. Resultados

Os resultados que obtivemos dos modelos foram os seguintes:



Model	AUC	CA	F1	Prec	Recall	MCC
kNN	0.915	0.865	0.863	0.865	0.865	0.714
SVM	0.905	0.860	0.858	0.863	0.860	0.705
Tree (1)	0.850	0.851	0.851	0.851	0.851	0.688
Random Forest	0.930	0.882	0.882	0.882	0.882	0.751
Neural Network	0.942	0.898	0.896	0.901	0.898	0.786



Model	MSE	RMSE	MAE	MAPE	R2
Linear Regression	0.469	0.685	0.548	1160...	0.507