Data mining

Students' Performance

Students:

- André Ferreira A81350
- Bruno Fonseca A83029
- Daniel Costa A81434
- Luis Silva A80981

Universidade do Minho
Escola de Engenharia

Teacher: Paulo Cortez

Project Theme

Helping Students through Data Mining

Education in Portugal has grown immensely, as well as the number of students placed in higher education. Despite the large growth, there are still flaws and a considerable percentage of students who simply give up.

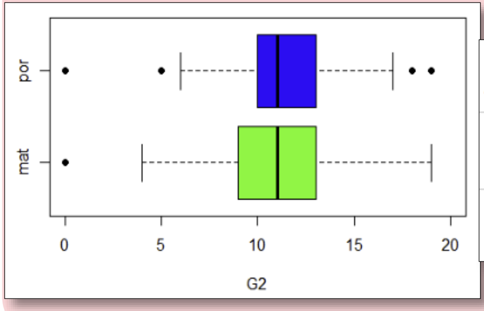
Business Understanding

Business Objectives:

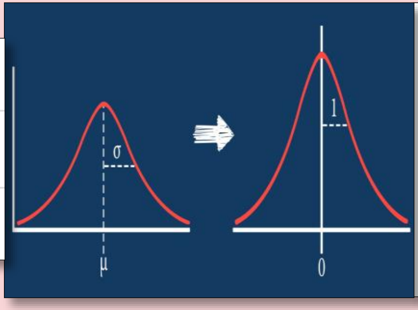
- Discover **which variables affect a student's success** in the subjects of Portuguese and Mathematics the most;

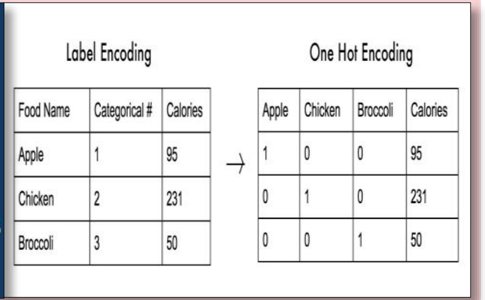
- **Categorize similar types of students** to identify target groups of students that have the highest risk of failure.

Data preparation



Address	
urban	0
rural	1





Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

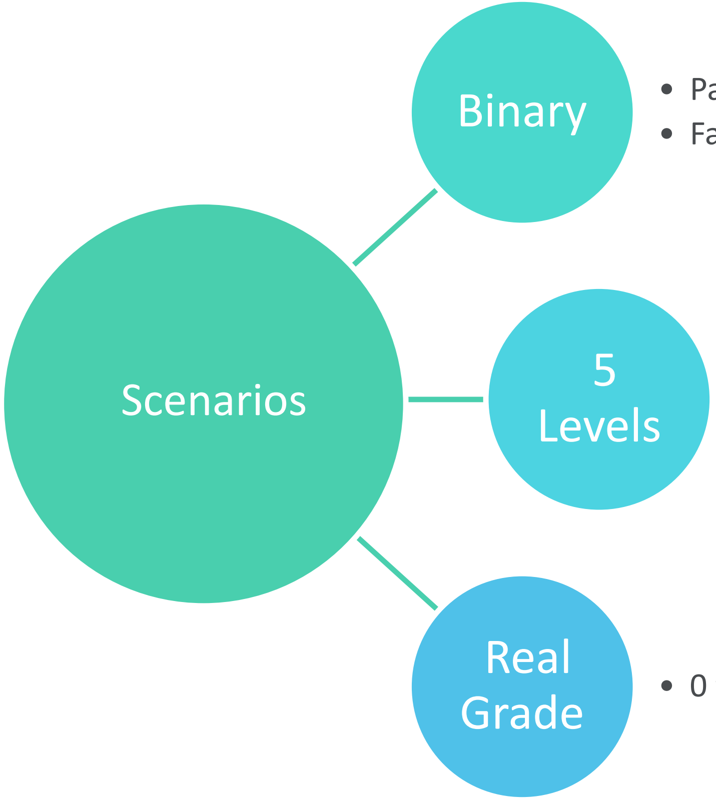
Outliers

Binary attributes

Normalization

One-hot encoding

Modeling



- Pass (1)
- Fail (0)

- 1 (16-20)
- 2 (14-15)
- 3 (12-13)
- 4 (10-11)
- 5 (0-9)

- 0 to 20

Modeling

Input Configurations

- InA
  - All attributes.
- InB
  - Without G1
- InC
  - Without G2
- InD
  - Without G1 and G2

Modeling

Classification and Regression

- Random Forest
- Naive
- Nearest-Neighbor
- Decision Tree
- Support Vector Machine

Association Rules

- Apriori

Clustering

- K-Means
- DBSCAN

Classification and Regression

Binary

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	91.0	68.3	**91.6**	77.9	88.5	**91.3**	86.2	90.9	86.7	89.3
InB	90.9	68.3	89.5	68.9	**91.8**	89.0	86.2	**92.1**	86.1	85.6
InC	80.3	68.3	**87.9**	72.6	79.8	85.9	86.2	**88.2**	86.6	85.1
InD	65.6	68.3	65.1	57.6	**69.0**	**86.6**	86.2	85.6	85.5	86.4

5-Level

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	75.0	47.1	**86.9**	42.6	56.6	67.3	28.9	**74.1**	44.4	55.1
InB	65.6	47.1	**85.2**	38.5	43.1	67.0	28.9	**74.4**	35.8	51.0
InC	60.3	47.1	**64.7**	40.3	47.6	54.5	28.9	**60.4**	36.8	43.6
InD	45.6	47.1	**47.2**	28.0	41.3	42.2	28.9	38.8	30.9	**43.1**

RMSE

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	**1.735**	4.590	1.996	3.635	2.214	**1.313**	3.233	1.476	2.301	1.468
InB	**1.906**	4.590	1.996	4.015	2.279	**1.429**	3.233	1.476	2.549	1.481
InC	**2.451**	4.590	2.664	4.189	2.979	1.785	3.233	**1.730**	2.713	1.891
InD	**3.930**	4.590	4.361	4.828	4.237	**2.665**	3.233	2.934	2.668	2.713

MAE

Input	Math					Portuguese				
Model	RF	NV	DT	NN	SVM	RF	NV	DT	NN	SVM
InA	**1.134**	3.438	1.218	2.671	1.386	**0.8144**	2.409	0.8551	1.639	0.8866
InB	1.282	3.438	**1.218**	2.992	1.400	**0.8915**	2.409	0.8551	1.835	0.9109
InC	**1.776**	3.438	1.886	3.097	1.973	1.2190	2.409	**1.199**	1.933	1.210
InD	**2.966**	3.438	3.264	3.644	3.119	**1.9270**	2.409	2.163	1.915	1.933

Evaluation

Confusion Matrix

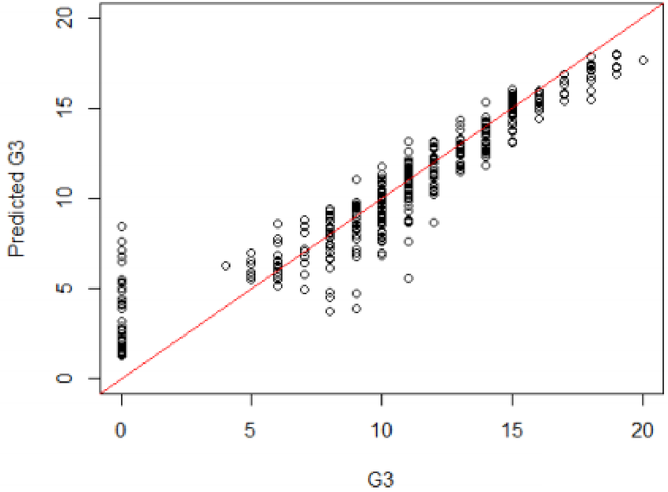
Predictions \ References	0	1
0	**29**	7
1	3	**80**

Acc= 91.60%
Sens=90.63%

Predictions \ References	5	4	3	2	1
5	**61**	3	0	0	0
4	4	**16**	0	0	0
3	0	2	**12**	2	0
2	0	0	0	**15**	1
1	0	0	0	0	**2**

Acc= 86.9%
Sens5=93.85%
Sens4=76,19%

Residual Plot



Outliers Removal vs Non-Outliers Removal

	Without outliers				With outliers			
Scenario	Binary	5-Levels	RMSE	MAE	Binary	5-Levels	RMSE	MAE
InA config	**92.23%** (SVM)	**86.92%** (DT)	1.780 (RF)	1.150 (RF)	91.6% (DT)	86.9% (DT)	**1.735** (RF)	**1.134** (RF)
InD config	**76.79%** (RF)	41.44% (RF)	**3.640** (RF)	**2.760** (RF)	69.0% (SVM)	**47.2%** (DT)	3.930 (RF)	2.966 (RF)

Association Rules

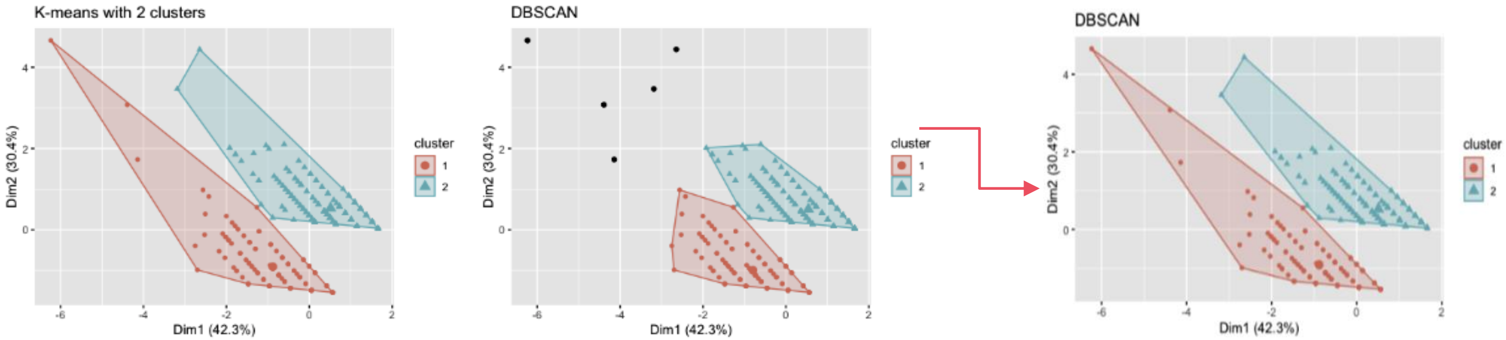
lhs	support	conf	coverage	lift	count
famsup_no, G2_pass, schoolsup_no	0.2278	1.000	0.2278	1.4907	90
activities_no, G2_pass,romantic_no	0.2076	1.000	0.2076	1.4907	82
activities_no, Dalc_1,G2_pass	0.2025	1.000	0.2025	1.4907	80
G2_pass, guardian_mother, paid_no	0.2051	1.000	0.2051	1.4907	81
Dalc_1, G1_pass, G2_pass	0.4152	1.000	0.4177	1.4815	164

lhs	support	conf	coverage	lift	count
famsup_no, G2_pass, schoolsup_no	0.2278	1.0000	0.2278	1.4906	90
activities_no, G2_pass, romantic_no	0.2076	1.0000	0.2076	1.4906	82
activities_no, Dalc_1, G2_pass	0.2025	1.0000	0.2025	1.4906	80
G2_pass, guardian_mother, paid_no	0.2051	1.0000	0.2052	1.4906	81
Dalc_1, G1_pass, G2_pass	0.4152	0.9939	0.4177	1.4815	164

lhs	support	conf	coverage	lift	count
G2_pass	0.6101	0.9679	0.6304	1.4427	241
G2_pass, higher_yes	0.5949	0.9671	0.6152	1.4415	235
G1_pass	0.5747	0.8972	0.6405	1.3374	227
G1_pass, G2_pass	0.5595	0.9822	0.5696	1.4641	221
G1_pass, higher_yes	0.5595	0.8984	0.6228	1.3391	221

Rules Association for G3 and pass

Clustering



Internal Evaluation
- Dunn index
- 0.894

External Evaluation
- Rand Index
- 0.850

Deployment

The models won't properly deployed, but it is possible to outline a plan to preemptively help the students:

- Beginning of the academic year;
- End of first period;
- End of second period.

At the end of the year, the new data can be added to the models so that these give even better prediction results, further enhancing the predictive algorithms.

Conclusion

- Prediction of students' performance based on socio-economical variables and school reports to support each student according to their education needs.

- Data Mining techniques were tested and analyzed, in order to achieve better results.

Conclusion

- **Classification task** – Decision Tree algorithm was generally better.
- **Regression task** – Random Forest algorithm performed better than the other parts.
- **Association Rules** – The rules with the greatest support and confidence are pass/fail in G1 and G2.
- **Clustering algorithms** – K-Means and DBSCAN were implemented and presented similar results.