Using Data Mining to Predict Secondary School Student Performance

by Paulo Cortez and Alice Silva

LAMPROS KARAVIDAS – 4

PRODROMOS POLYCHRONIADIS – 9

DATASETS:

- *Student-mat.csv*
- *Student-por.csv*

Introduction

High student failure rates compared to other European countries

Lack of success in the core classes of Mathematics and the Portuguese

Secondary education students/ school reports /questionnaires

DM models

- Decision trees
- Random forest
- Neural networks
- Support vector machines

Results:

Good predictive accuracy through G1 and G2.

Our Analysis

- Models we used
  - **Linear (Multiple) Regression**: Find a relationship between one or more features
  - **Decision Tree**: Supervised learning model used for classification.
  - **Logistic Regression (Binary):** Statistical method for predicting classes (Pass/Fail)
  - **Logistic Regression (Ordinal)**: Target variable with more ordinal categories

Linear Regression

- Major features:
  - G1 = first period grade (numeric: 0 - 20)
  - G2 = second period grade (numeric: 0 - 20)
  - absences = number of school absences (numeric: 0 - 93)
  - failures = number of past class failures (numeric: n if 1 <= n <3, else 4)
  - age = students' age (numeric: from 15 to 22)

- **RMSE** (Root Mean Square Error) is the statistical metric.
  - How concentrated the data is around the line of best fit
  - Lower values of RMSE indicate better fit
  - It is the most popular metric

Linear Regression (Results)

Student-por.csv

	Test1	Test2	Test3	Test4	Test5		**Mean**
G2	0,97	1,46	1,28	**0,87**	1,44		**1,204**
G1	1,87	2,42	1,71	**1,71**	1,76		1,894
G1,G2	1,35	1,37	1,57	**1,12**	1,1		1,302
G2, Absences	1,53	1,36	1,07	1,22	**0,92**		1,22
Failures, Age	**2,81**	2,98	2,98	3,23	2,94		2,988
G1,G2,Absences	1,48	1,43	**1,1**	1,46	1,22		1,338
G1,G2,Absences, Failures, Age	1,26	1,25	1,2	**1,07**	1,38		1,232

Student-mat.csv

	Test1	Test2	Test3	Test4	Test5		**Mean**
G2	2,03	1,99	1,89	2	**1,52**		1,886
G1	**2,48**	2,72	2,61	2,9	2,77		2,696
G1,G2	2,23	2,31	1,63	**1,61**	2,05		1,966
G2, Absences	2,09	1,87	2,08	1,97	**1,71**		1,944
Failures, Age	4,08	4,12	4,15	4,55	**4,03**		4,186
G1,G2,Absences	1,84	1,71	1,77	1,96	**1,53**		**1,762**
G1,G2,Absences, Failures, Age	**1,59**	1,75	2,24	2,03	1,81		1,884

Decision Trees

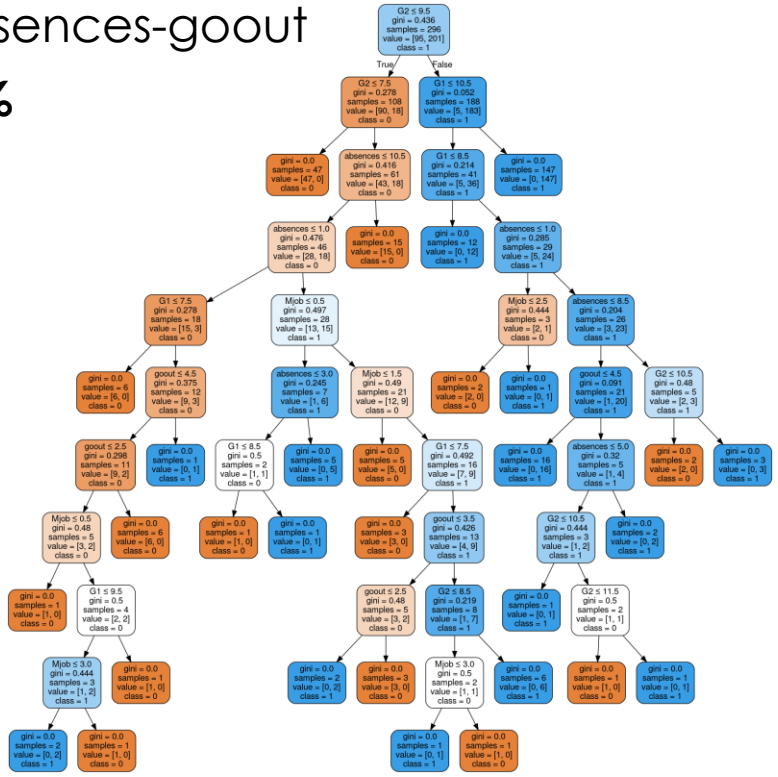
- **Major features:**
  - G1 = first period grade (numeric: 0 - 20)
  - G2 = second period grade (numberic: 0 - 20)
  - absences = number of school absences (numeric: 0 - 93)
  - mjob = mother's job (nominal)
  - goout = going out with friends (numberic: 1 (very low) - 5 (very high))

G1-G2-mjob

93%

G1-G2-mjob-absences-goout

87%

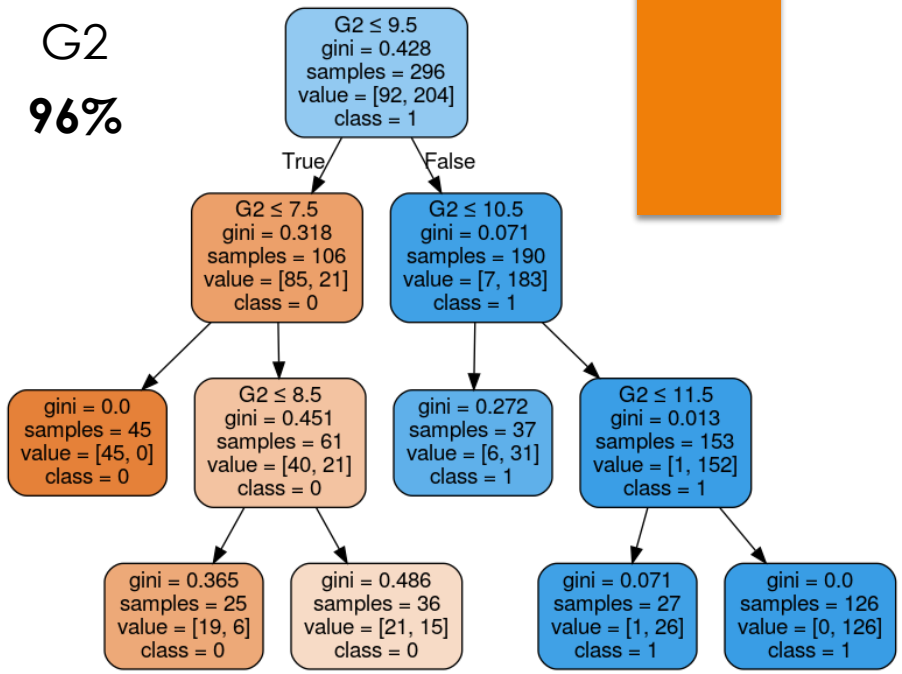


Decision Trees (Results) - Math

Decision Trees (Results) - Math

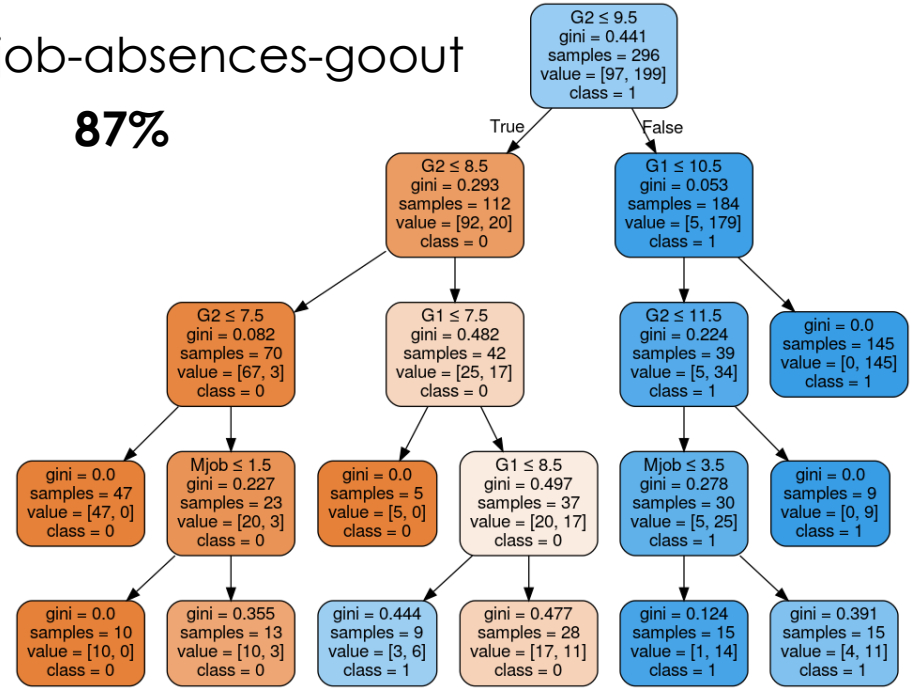
G2

96%



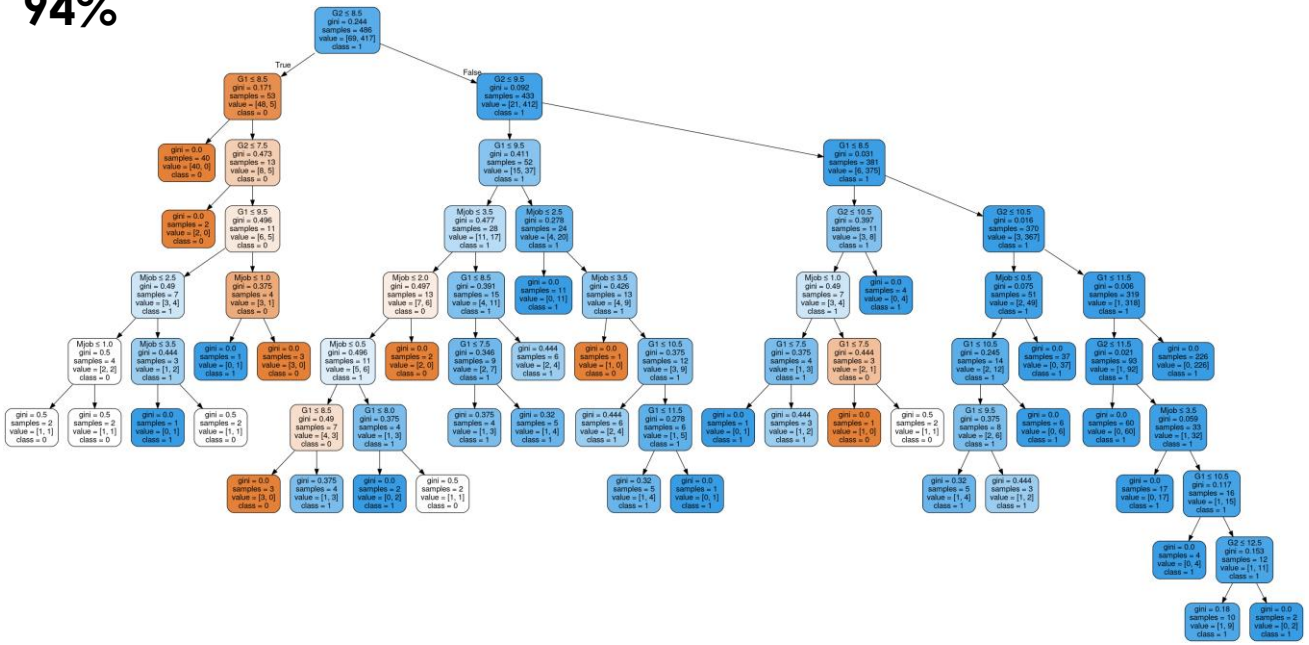
G1-G2-mjob-absences-goout

87%



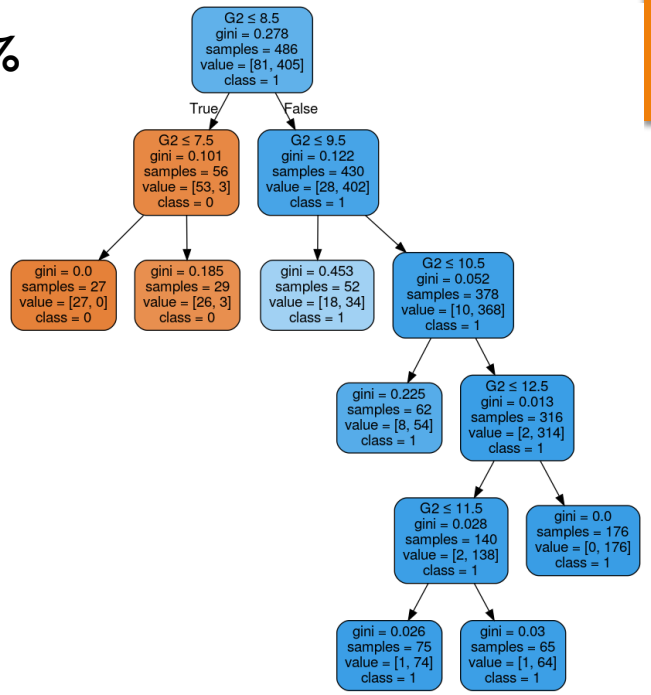
G1-G2-mjob

94%



G2

94%

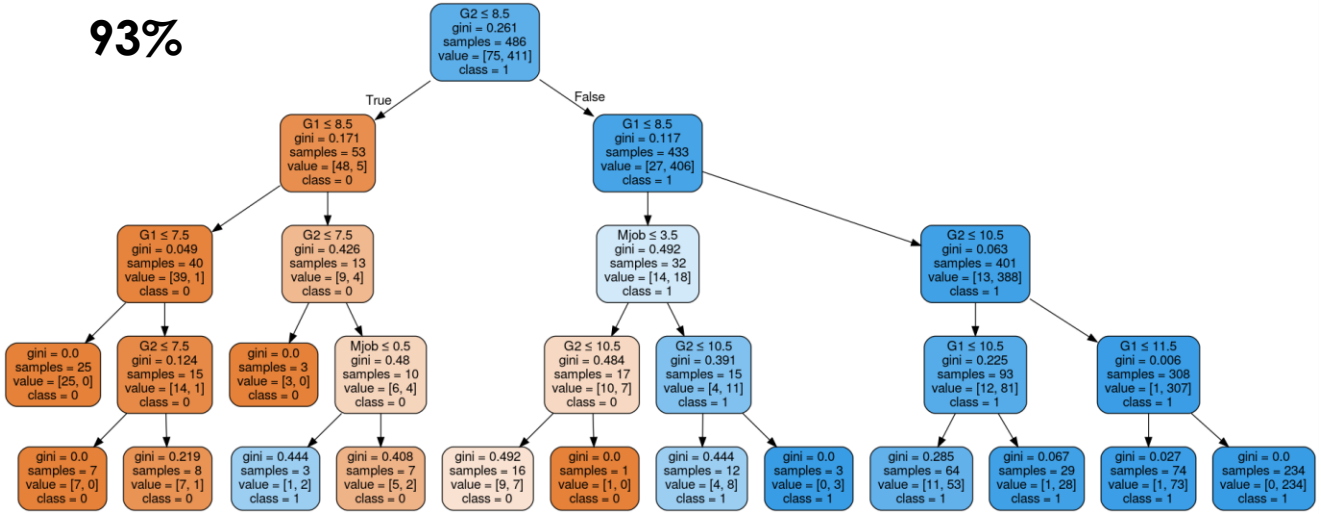


Decision Trees (Results) - Por

Decision Trees (Results) - Por

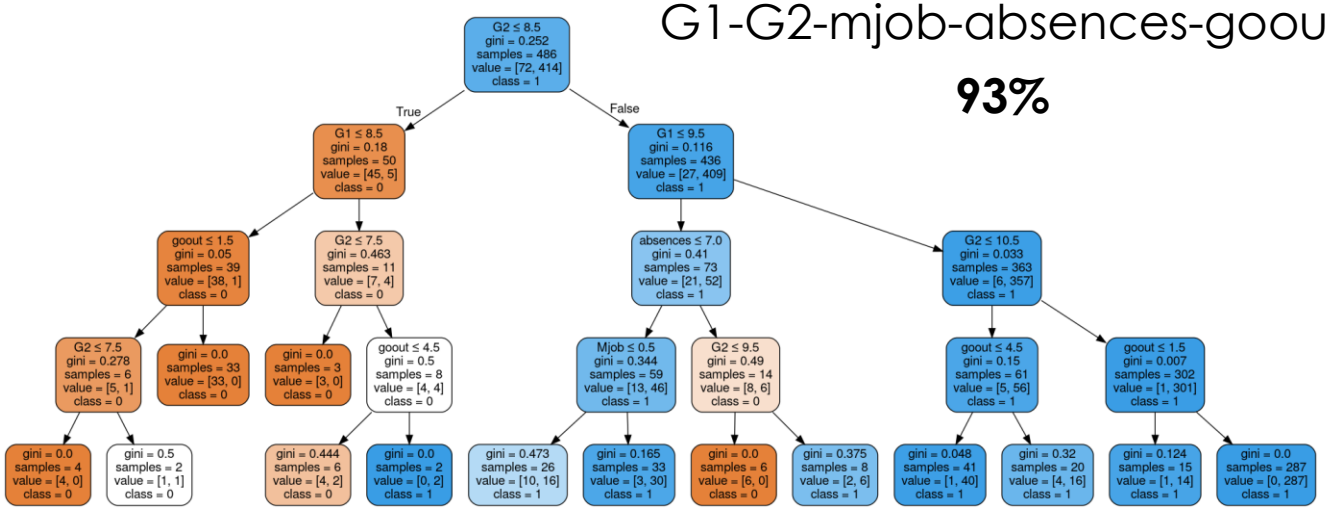
G1-G2-mjob

93%



G1-G2-mjob-absences-goout

93%



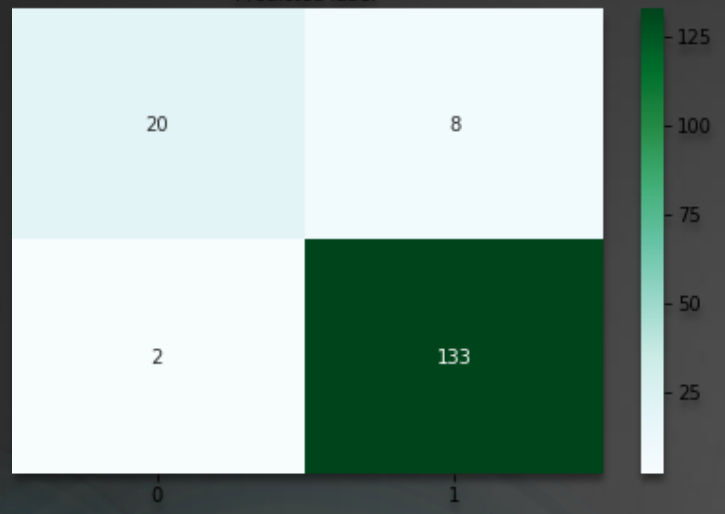
Logistic Regression (Binary)

- **Major features:**
  - G1 = first period grade (numeric: 0 - 20)
  - G2 = second period grade (numberic: 0 - 20)
  - absences = number of school absences (numeric: 0 - 93)
  - mjob = mother's job (nominal)
  - goout = going out with friends (numberic: 1 (very low) - 5 (very high))

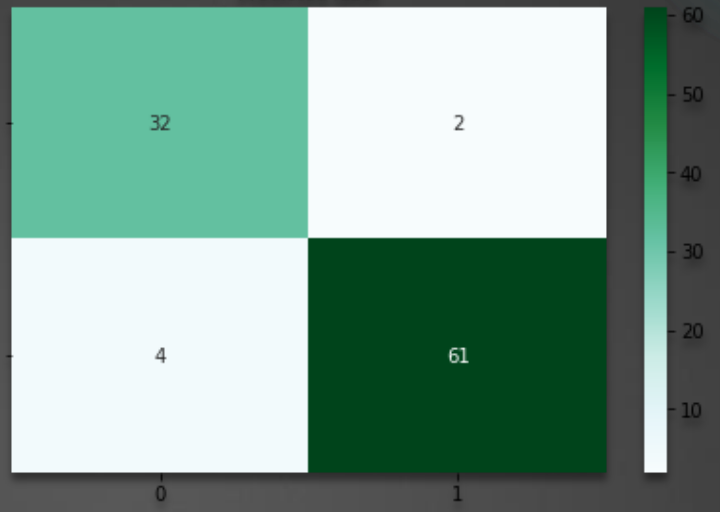
- **Confusion Matrix**
  - Describes the performance of our classification model
  - Accuracy: How often the classifier is correct

Logistic Regression (Binary – Results)

93,8%

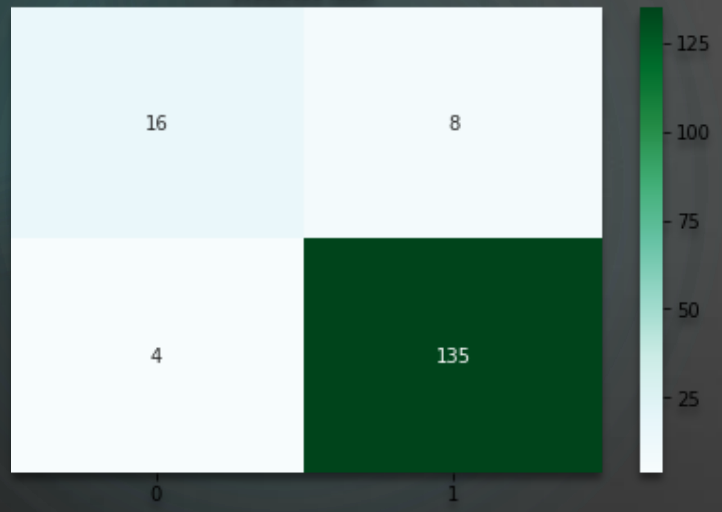


93,9%

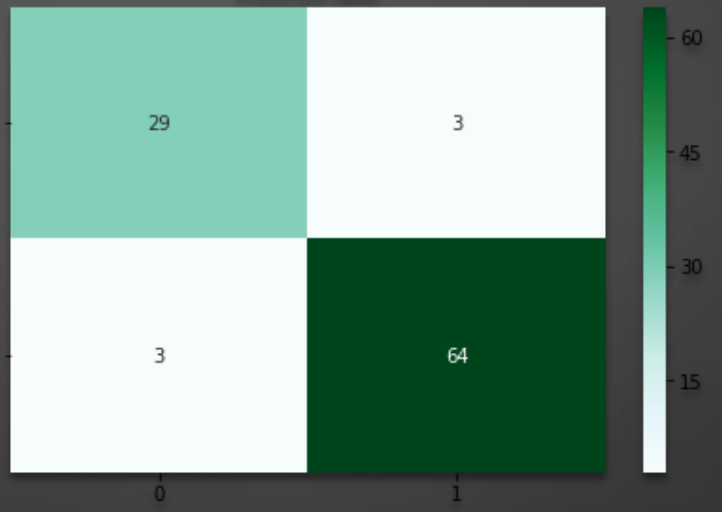


G2

92,6%



93,9%

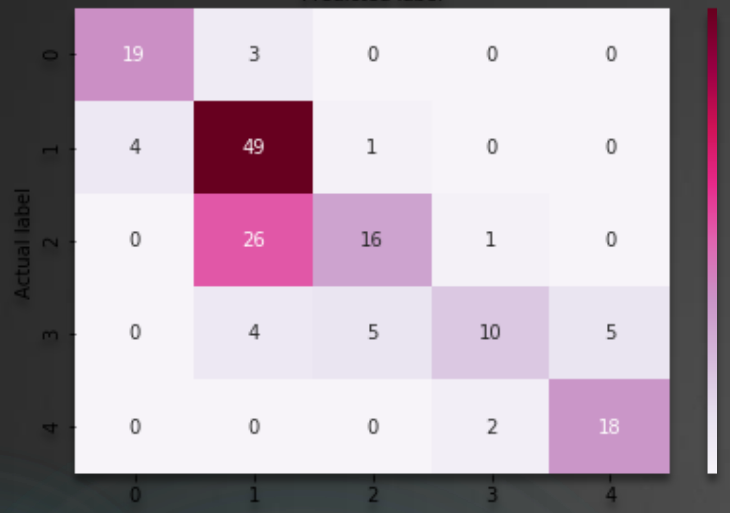


Logistic Regression (Ordinal)

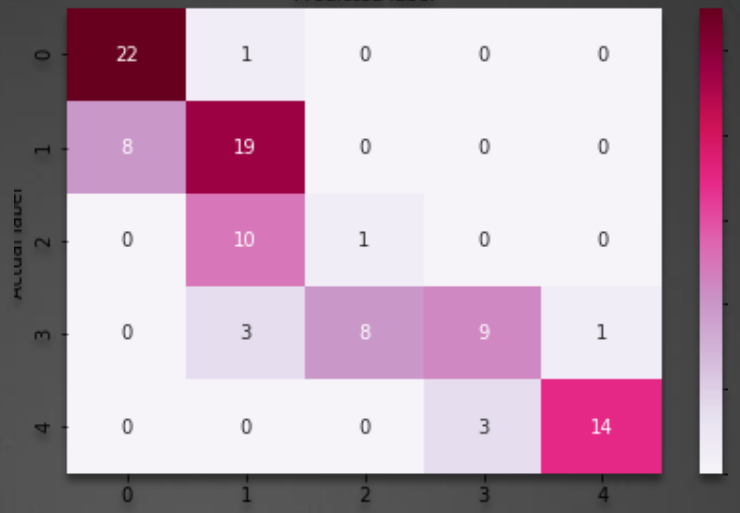
- **G3** values to 5 different classes
  - 0 - 9: transformed to 0 (fail)
  - 10 - 11: transformed to 1 (sufficient)
  - 12 - 13: transformed to 2 (satisfactory)
  - 14 - 15: transformed to 3 (good)
  - 16 - 20: transformed to 4 (excellent / very good)

Logistic Regression (Ordinal) - Results

68,7%

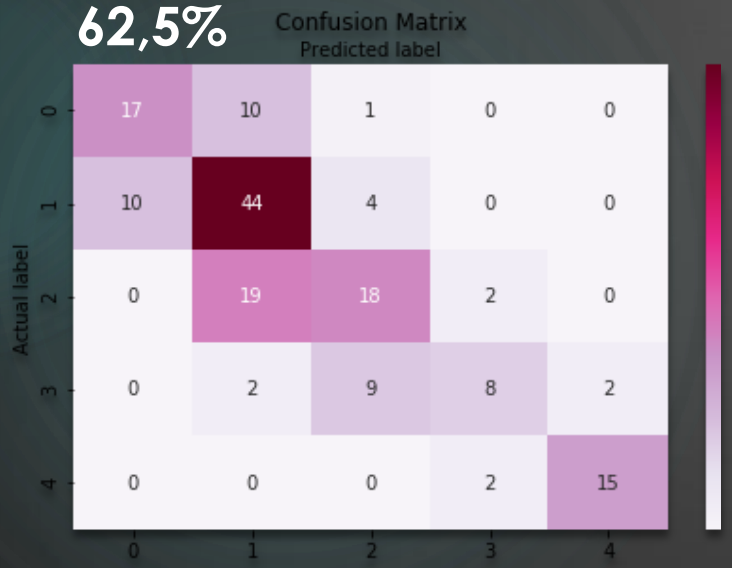


65,6%

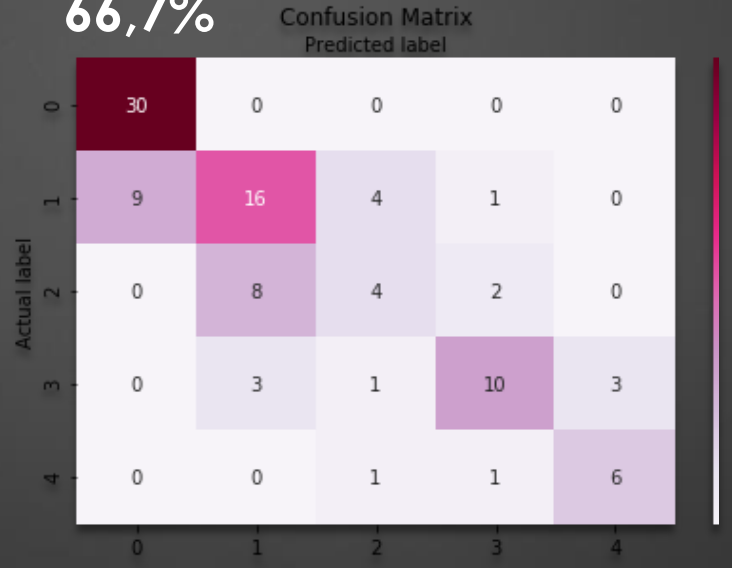


G2

62,5%



66,7%



Conclusions

- *G1* and *G2* are very important for a correct prediction of *G3*
- *G2* sufficies in order to have a pretty accurate prediction of *G3*
- *age* with *failures* is the worst predictor combination
- *G1*, *G2* and *mjob* is the best predictor combination of the ones mentioned in the paper for both classes (Mathematics and Portuguese language)
- *G2* provides us with predictions with the same or even, sometimes, higher accuracy
- *Binary* form is **much more** accurate than the *ordinal* one

All in all, we **confirm the findings of the paper** with the use of our own results and add that the ***G2* feature**, more or less, can always be a very **successful predictor** for the final grade *G3*.

ΣΑΣ ΕΥΧΑΡΙΣΤΟΥΜΕ