



# 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

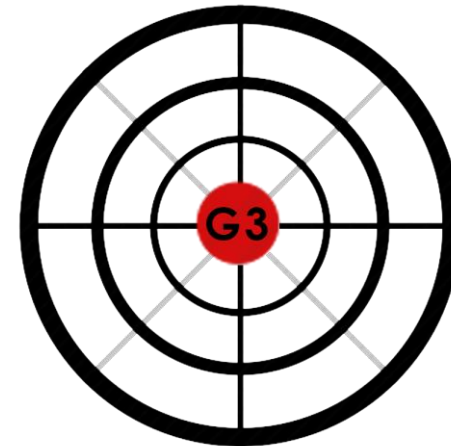


Results:

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
- ▶ Our target was to predict the grades of the third semester



# 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  $1 \leq n < 3$ , 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



Student-por.csv

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

Student-mat.csv

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

# Linear Regression (Results)

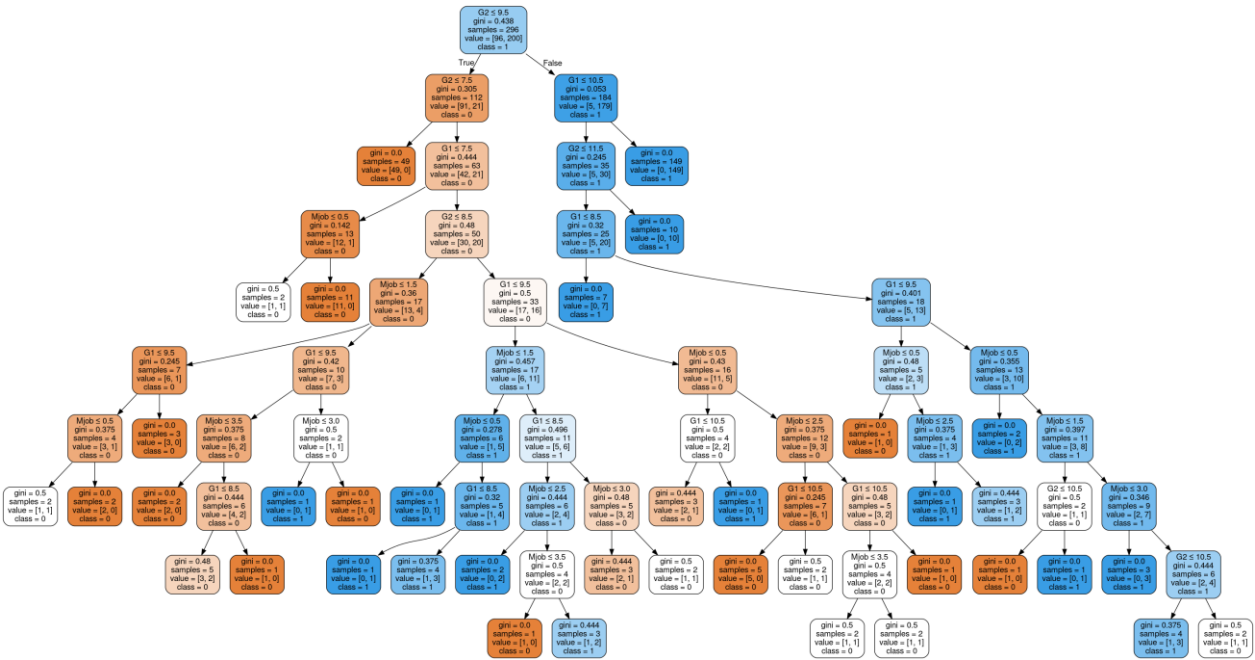
# Decision Trees

- ▶ **Major features:**

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

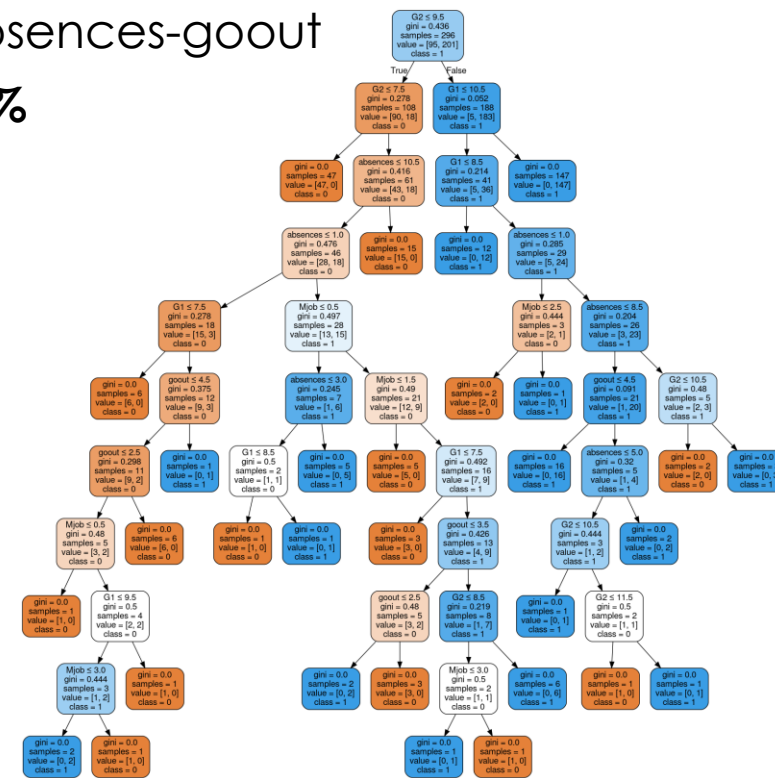
G1-G2-mjob

93%



G1-G2-mjob-absences-goout

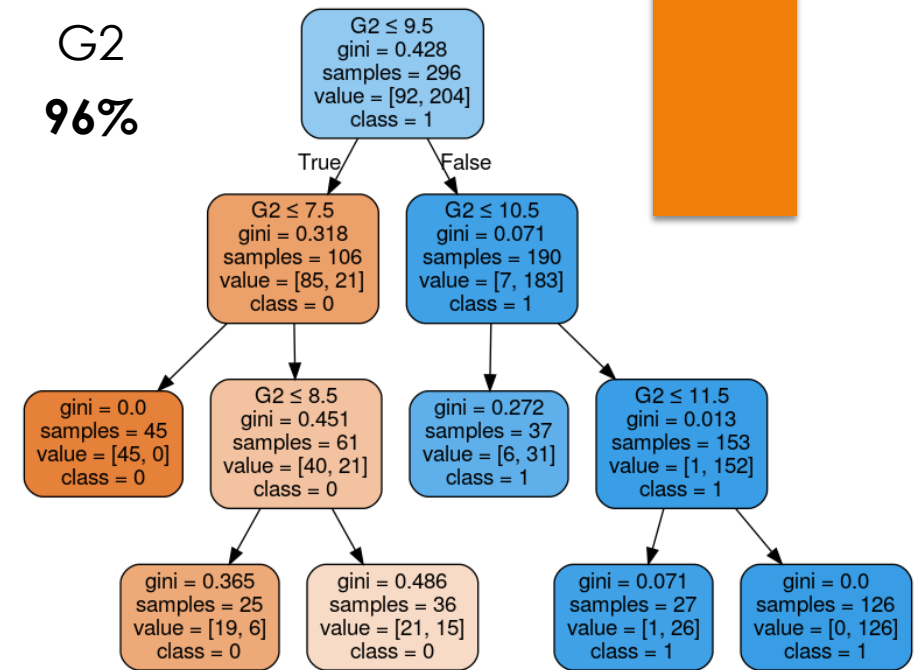
87%



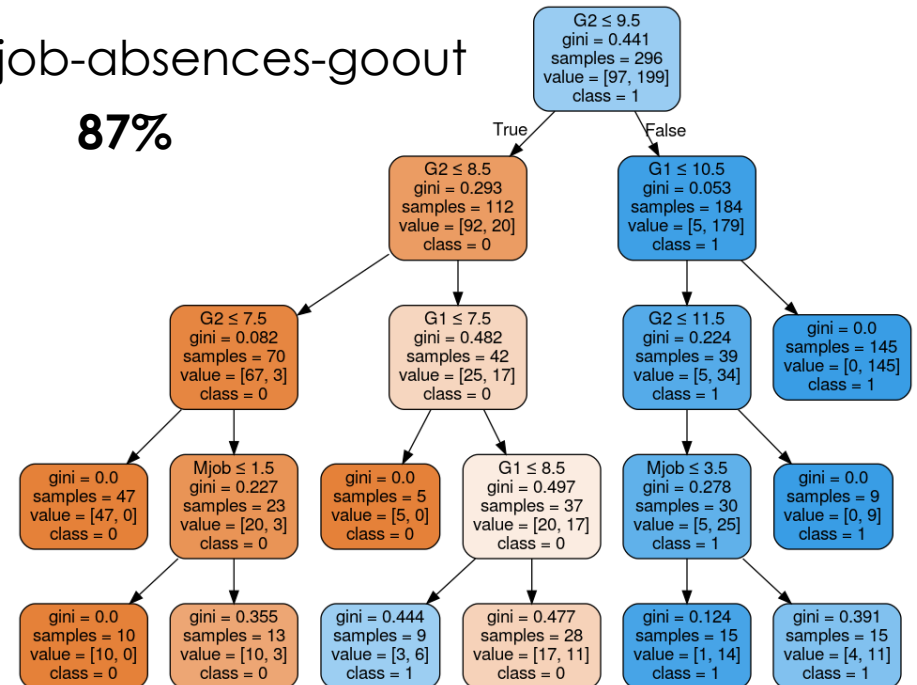
Decision Trees (Results) - Math

# Decision Trees (Results) - Math

G2  
96%



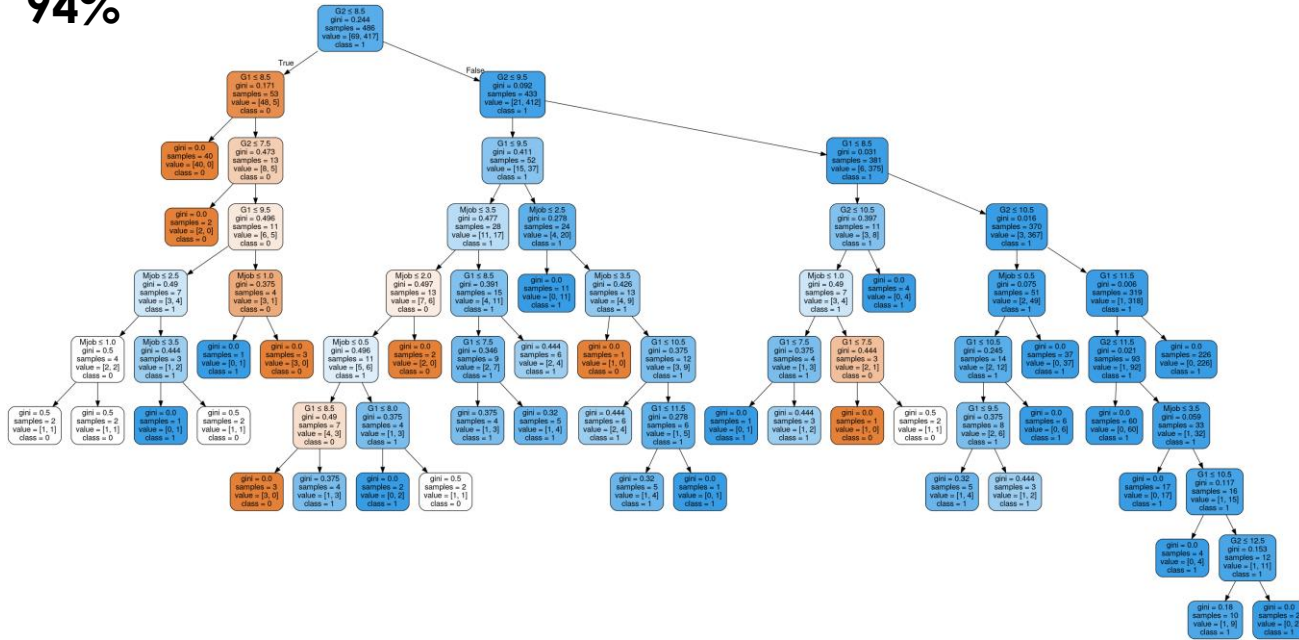
G1-G2-mjob-absences-goout  
87%





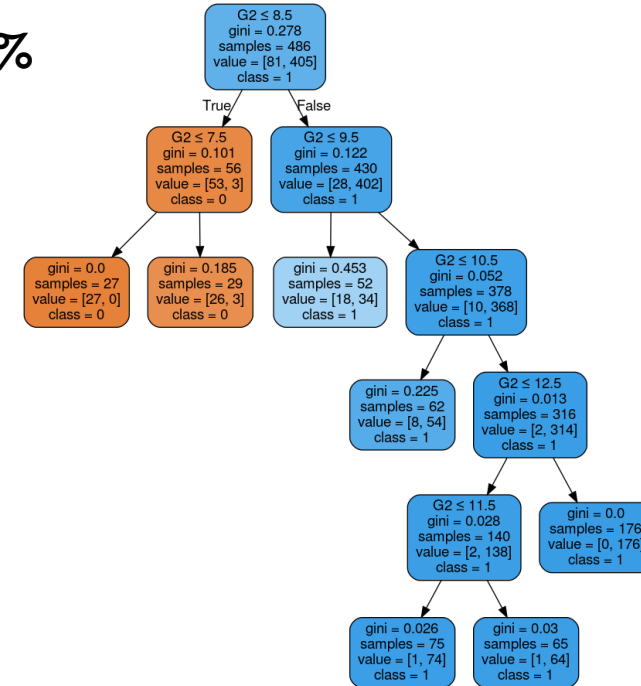
G1-G2-mjob

94%



G2

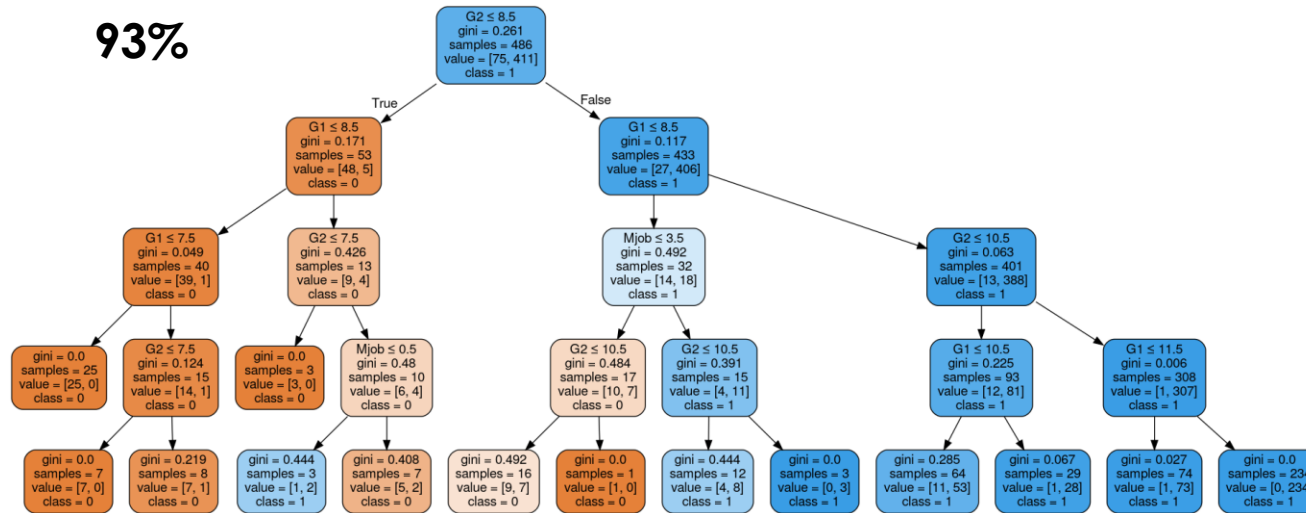
94%



Decision Trees (Results) - Por

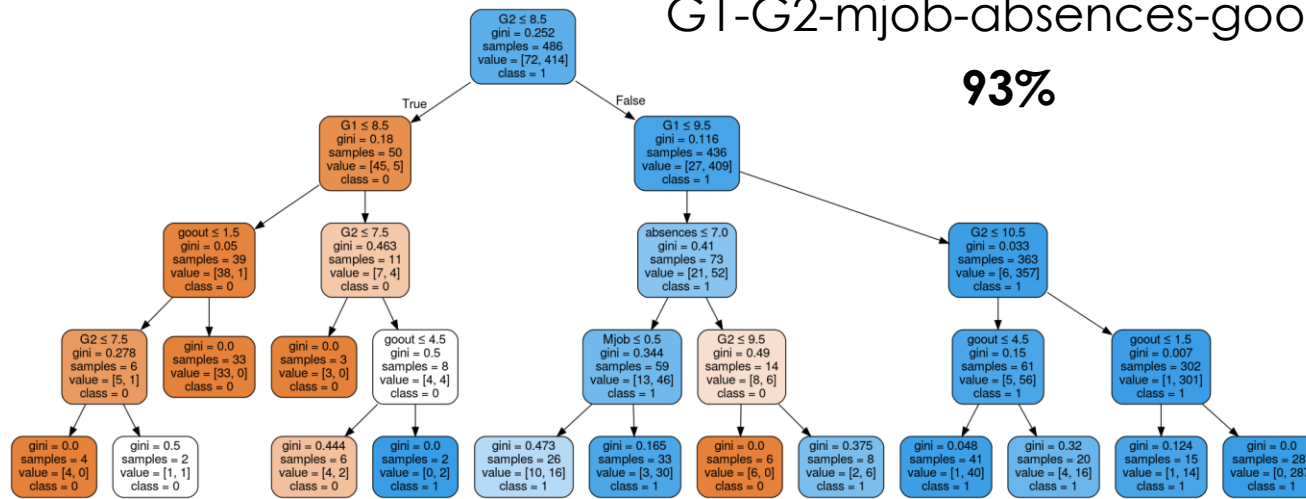
G1-G2-mjob

93%



G1-G2-mjob-absences-goout

93%



Decision  
Trees  
(Results) -  
Por

# Logistic Regression (Binary)

## ▶ **Major features:**

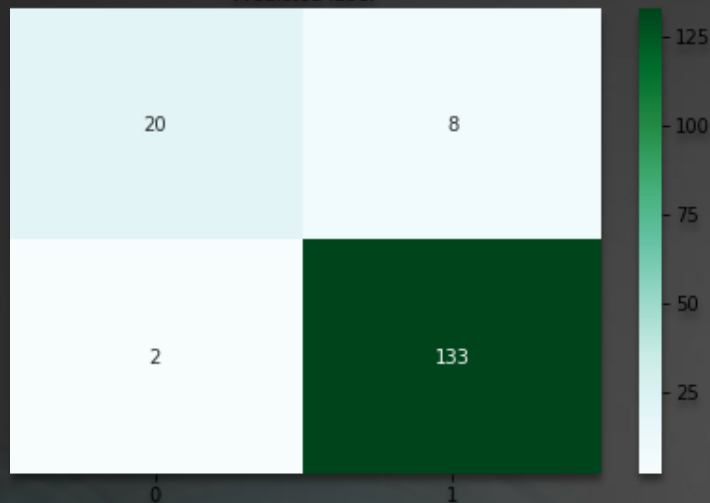
- ▶ G1 = first period grade (numeric: 0 - 20)
- ▶ G2 = second period grade (numeric: 0 - 20)
- ▶ absences = number of school absences (numeric: 0 - 93)
- ▶ mjob = mother's job (nominal)
- ▶ goout = going out with friends (numeric: 1 (very low) - 5 (very high))

## ▶ **Confusion Matrix**

- ▶ Describes the performance of our classification model
- ▶ Accuracy: How often the classifier is correct

93,8%

Confusion Matrix  
Predicted label



93,9%

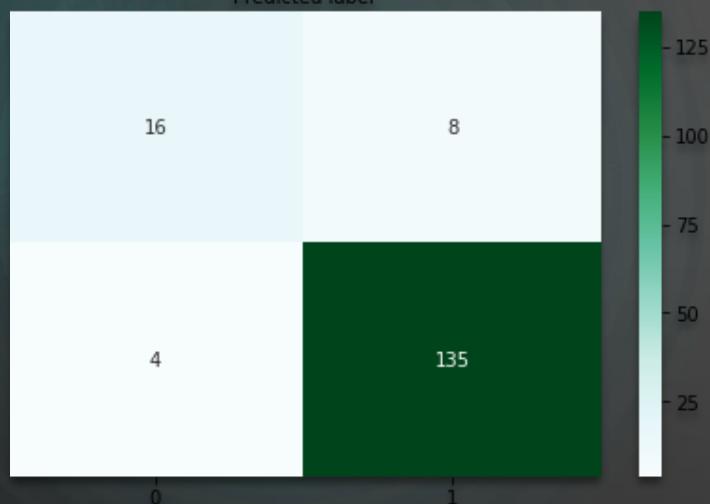
Confusion Matrix  
Predicted label



G2

92,6%

Confusion Matrix  
Predicted label



93,9%

Confusion Matrix  
Predicted label



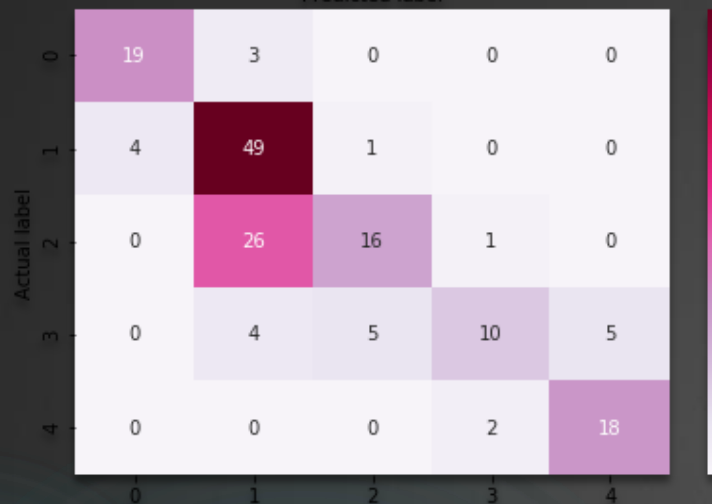
Logistic  
Regression  
(Binary –  
Results)

# 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)

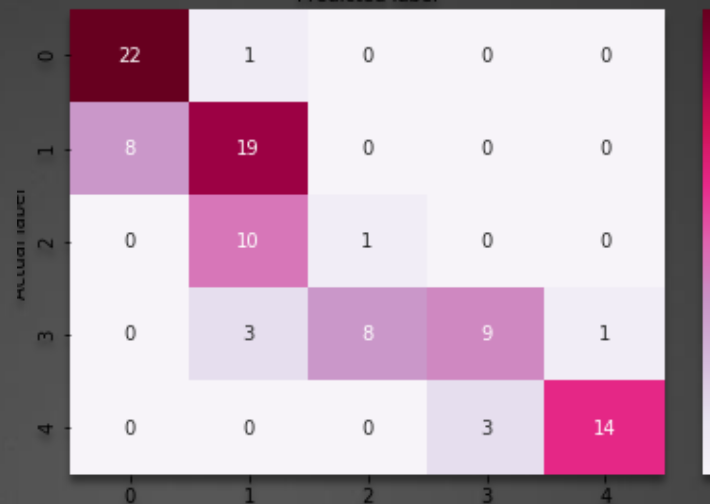
68,7%

Confusion Matrix  
Predicted label



65,6%

Confusion Matrix  
Predicted label



G2

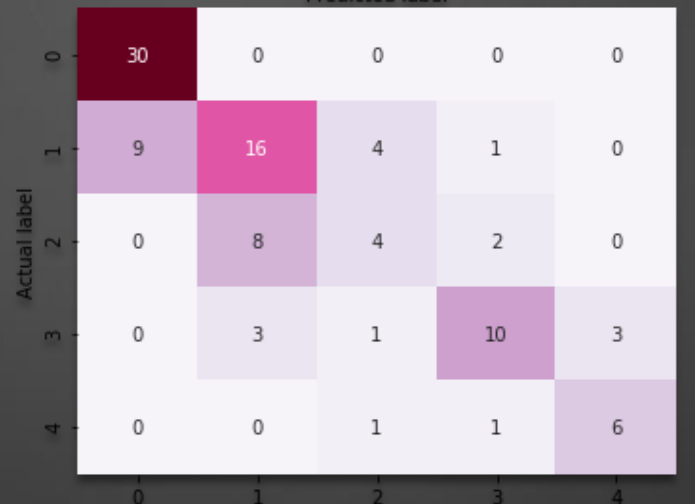
62,5%

Confusion Matrix  
Predicted label



66,7%

Confusion Matrix  
Predicted label



Logistic  
Regression  
(Ordinal) -  
Results

# Conclusions

G1 and G2 are very important for a correct prediction of G3

G2 suffices 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.

THANK YOU