Machine learning methods to predict student performance

José Miguel Hernández Cabrera[1], Jaume Pladevall Morros[1]

[1]Facultat Informàtica de Barcelona, Universitat Politècnica de Catalunya

Introduction

- Educational Data Mining (EDM) is a disciplinary research area, concerned with developing methods for exploring the unique and increasingly large-scale data that can be obtained from different educational sources.
- EDM works towards the improvement of educational processes by introducing better, and effective learning practices for students.

Source: Pixabay by FelixMittermeier

Dataset

- We worked with the data provided by Cortez and Silva and retrieved from the UCI Machine Learning Repository.
- The data features can be classified into two categories: mark reports and questionnaires.
- The data is integrated into two different data sets. One for the subject of Mathematics and the other one for the Portuguese language subject.
- Each data set has 33 features and a different number of students (649 for Portuguese and 395 for Mathematics).

Objectives

- Use machine learning methods to predict student success **pass/fail** (0/1) according to the information retrieved about them.
- Find which are the most important features that affect the student performance.
- Try to see if with a model only trained to predict the success of students in the Portuguese subject could be generalized into predicting the success of students for the Mathematics subject.

Source: Pixabay by Fathromi Ramdlon

Grades

- From 0 to 20.
- pass: Grade >= 10
- fail: Grade < 10

Acknowledgements

This work was possible because of the knowledge obtained from the ML subject.

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Methodology

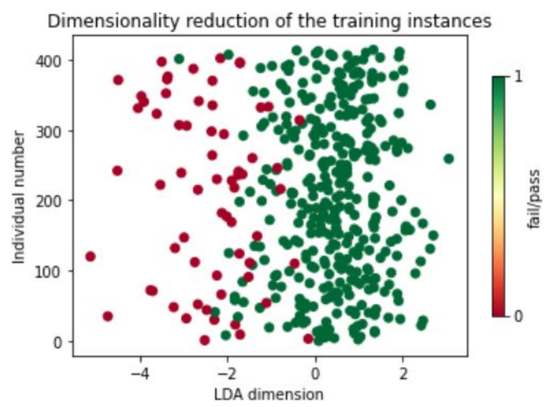
- Data exploration: Detection of outliers.
- Transformations:
  - One hot encoding for the questionnaire variables.
  - Normalization of numerical variables.
- Feature importance/selection.

Modelling

- Data set partition into Train / Validation / Test balanced sets.
- Metric used when selecting models: F1-score (macro avg)
- Models used
  - Logistic Regression
  - Linear Discriminant Analysis
  - Quadratic Discriminant Analysis
  - Naive Bayes
  - Decision Trees
  - Random Forest
  - Support Vector Machines
  - Multi Layered Perceptron

DT considers that you still can pass with the two previous grades failed!

```
if G2 <= 8.5: return 0
if G2 > 8.5: if G1 <= 8.5: if absences > 5: return 0
if G2 > 8.5: if G1 <= 8.5: if absences <= 5: return 1
if G2 > 8.5: if G1 > 8.5: return 1
```



Conclusions

- The most important variables are mainly Second period and First period grades. Other less important: failures, absences, higher, school and age.
- Best model for Portuguese subject an SVM with feature selection.
  - **91.14%** (F1-score macro avg)
  - **95.38%** (accuracy)
- Improvement of **more than 2%** for the Portuguese data set binary classification from original paper.
- Portuguese generalization to Math. SVM with feature selection.
  - **87.82%** (F1-score macro avg)
  - **89.62%** (accuracy)

References

- EDM website, https://educationaldatamining.org/
- A. Aleem and M. M. Gore et al., 2020, https://doi.org/10.1109/CSNT48778.2020.9115734
- P. Cortez and A. Silva et al., 2008, http://hdl.handle.net/1822/8024
- D. Dua and C. Graff et al., 2017, https://archive.ics.uci.edu/ml/index.php