PREDICTING ACADEMIC PERFORMANCE OF STUDENTS (A MODIFIED VERSION OF WAHEED ET. AL. 2020)

Hamidreza Hashemi
University of florence (Computer Engineering)

***Abstract* – The objective of this document is to make a summary about the experiment**

***Keywords* – NTRODUCTION, MPLEMENTATION SUMMARY, MODULE STRUCTURE, TESTS AND CONCLUSIONS, REFERENCES**

I. INTRODUCTION

This is a university project that creates a python project that modifies the table 3 of the Waheed et. al. 2020 [1] replacing ANN, LR with ADA Boost, Perceptron and ignores SVM Algorithm.

II. IMPLEMENTATION SUMMARY

This project uses open univesity dataset [2] that has extracted database's tables in the *.csv* format. It uses **pandas** and **scikit learn** in order to extract, preprocess the data and then implement the machine learning techniques on it. It uses less features than the main article [1] as it's described to be 30 optimized features.

Features
assessment-type
highest-education
num-of-prev-attempts
studied-credits
disability
sum-click

TABLE I
Selected features

You can see the selected features in the table I. It's important to say that in the program it's selected more columns; But they are ID's and weights that will be dropped before fitting the model.

III. MODULE STRUCTURE

This module has 2 main python scripts.

A. Preprocess script

This script prepares a ready dataframe for the algorithms. It reads the *.csv* files, changes data types for optimizing the space occupation and it merges the tables with specific ID's. It also encodes the categorical features in order to fit well in the algorithms since those algorithms accept only numerical values. It divides the data into four categories which are :

- Pass/Fail
- Distinction/Fail
- Distinction/Pass
- Withdrawn/Pass

It drops the rows which contain missing variables. Then it saves the merged data with the encoded categorical features and label into a pickle file.

B. Analysis script

This script reads the produced pickle file. Then it splits the data into feature and label variables. It also uses the weight feature as the sample weight for the ADA Boost algorithm. Finally it analyses the divided data with ADA Boost and Perceptron algorithm (on a 75 percent train and 25 percent test data) and it gives the model's important scores as an output.

IV. TESTS AND CONCLUSIONS

This program is been tested on a random sample of 1000 students.

Results					
Categories	Techniques	Accuracy %	Loss	Precision	Recall
Pass/Fail	ADA-Boost	86.95	0.1304	0.88	0.98
	MLP	91.32	0.0867	0.93	0.97
Distinction/Fail	ADA-Boost	88.30	0.1169	0.88	0.76
	MLP	92.91	0.0708	0.92	0.86
Distinction/Pass	ADA-Boost	76.45	0.2354	0.82	0.87
	MLP	82.17	0.1782	0.83	0.95
Withdrawn/Pass	ADA-Boost	83.83	0.1616	0.49	0.08
	MLP	88.68	0.1132	0.85	0.36

TABLE II
Results on 1000 random students

On the table II you can see the important scores of this experiment. As you can see we have low precision and recall scores on the Withdraw/Pass section but other results seems to be satisfiable.

REFERENCES

[1] H. Waheed, "Predicting academic performance of students from VLE big data using deep learning models", *Science Direct*, 2020.

[2] "Open university dataset", URL: `https://analyse.kmi.open.ac.uk/open-dataset`.