**Exploring Feature Importance for Predicting Students' Academic Performance with Machine Learning Algorithms**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ntaoulas Vasileios**  School of Informatics  Aristotle University  Thessaloniki, Greece  ntaoulasv@csd.auth.gr | **Malamatinos Marios-Christos**  School of Informatics  Aristotle University  Thessaloniki, Greece  XXXX@csd.auth.gr | **Papanikolaou Stefanos**  School of Informatics  Aristotle University  Thessaloniki, Greece  XXXX@csd.auth.gr | **Kalyvas Emmanouil**  School of Informatics  Aristotle University  Thessaloniki, Greece  kalyvase@csd.auth.gr |

**ABSTRACT**

The integration of technology and data analysis in the field of education has opened up new possibilities for understanding and improving students' academic performance. Researchers and educators can now leverage the power of machine learning algorithms to gain valuable insights about students’ performance. In this study we investigate the importance of features in predicting students' failure or success using machine learning algorithms. The findings of this study have practical implications for educators, policymakers, and researchers, enabling the development of targeted interventions and personalized strategies to enhance students' learning experiences and overall academic achievements.

**1 INTRODUCTION**

In this project, we present a comprehensive analysis of four diverse datasets related to student performance, namely the Kırşehir Ahi Evran University dataset, an UCI Dataset, an Open University Learning Analytics Dataset, and a Buraimi University College Dataset. The primary objective of our study is to explore the important features that contribute significantly to student performance across four different datasets. By applying various machine learning techniques and evaluating the feature importance, we aim to uncover the underlying factors that have the most significant impact on student outcomes. This knowledge can provide valuable insights to educational institutions, policymakers, and stakeholders, aiding them in making informed decisions to improve educational practices and support student success.

**2 DATASET**

We utilized four distinct datasets in order to strengthen the validity of our work. By encompassing datasets with varying feature dimensions and record sizes, we have obtained a robust understanding of our model's performance under different conditions. This thorough analysis enhances the reliability and generalizability of our findings, allowing us to draw meaningful conclusions and make informed decisions based on the outcomes.

2.1**Kırşehir Ahi Evran University (KAEU)** [1]

The dataset consisted of the academic achievement grades of 1854 students who took the Turkish Language-I course in a state University in Turkey (Kırşehir Ahi Evran University) during the fall semester of 2019–2020. The features are:

|  |  |
| --- | --- |
| **Feature** | **Comment** |
| stdID | Student’s ID |
| mid-term | Student’s mid-term exam grade |
| final | Student’s final exam grade |
| faculty | Department’s faculty |
| department | University’s department |

Midterm and final exam grades are ranging from 0 to 100. The end-of-semester achievement grade is calculated by taking 40% of the midterm exam and 60% of the final exam.

We categorized the mid-term, final and achievement grades as below:

* **1** grade < 32.5
* **2** 32.5 <= grade < 55
* **3** 55 <= grade <77.5
* **4** grade >= 77.5

Before we use any machine learning model to predict the student final achievement grade, we did the following steps:

1. Drop unnecessary columns.
2. Calculate achievement\_grade as we showed previously.
3. Calculate the absolute difference between the 'mid-term' and 'final' grades (grade\_difference).
4. Calculate the ratio of the 'mid-term' grade to the 'final' grade (grade\_ration).
5. Make a binary feature, called grade\_change, which is 1 when final grade is greater than mid-term grade, unless it’s 0.
6. One-hot encoding on faculties and departments.
7. Categorize the 3 grades based on the above categories (named final\_category, midterm\_category, achievement\_grade\_category)

2.2**UCI Dataset**

. . .

2.3**Open University Learning Analytics Dataset** [2]

. . .

2.4**Buraimi University College Dataset (BUC)** [3]

The data of this dataset was collected from the Buraimi University College of Oman and consists of students’ performance, for 3 semesters, on the course ‘Phonetics and Phonology’. All in all it contains 151 student records with 10 features and 1 prediction class. The prediction class classified the students’ performance as ‘LOW’ or ‘HIGH’. The features mentioned are the following:

* Attendance - The percentage of student attendance in the lectures
* Exam1\_Grade - The students’ grade on the first exam out of three exams in total
* CGPA - Cumulative Grade Points Average of student
* GEN - The students’ gender
* Major - Major of the student. All the students have a Major either in Translation or in Literature so we have two possible values.
* Session - Whether the course was performed in the evening or the morning so we have two possible values.
* PreReq\_Grades - The grade, in percentage, for the prerequisite subject
* Final\_Grade - The classification of the student as ‘LOW’ or ‘HIGH’ depending on their performance
* Year - The students’ year of study

We splitted the dataset in two parts, one with 80% of the data that was used for training and one with 20% of data used for the evaluation of our models.

Out of the 151 students in this dataset, 115 have been classified as ‘HIGH’ performing with only 36 of them characterized as ‘LOW’. This is a very significant class imbalance with almost 70% of the classes being ‘HIGH’. To resolve the imbalance, we applied oversampling with replacement on our training data to achieve the same number of records for each of our classes. In this way we also increase the importance of the minor class which is also the most important of the two as we would like to understand which students are about to fail.

The machine learning algorithms we used are the following:

* Linear Regression
* Decision Tree
* Random Forest
* MultiLayer Perceptron (MLP)

We used linear regression and tree models because they are easily interpretable and give us an immediate view on the importance of our features.

We also tried MLP that is not easily interpretable to see how it will perform compared to the others. The interpretability of this model was made through Shapley Values.

**3 RESULTS**

3.1**KAEU Dataset**

Given the limited number of features in this dataset, consisting of only five variables, the potential for overfitting was anticipated. To mitigate this issue, we employed feature engineering techniques, including one-hot encoding, to enhance the predictive performance of our models. In our analysis, we employed five well-established classification algorithms: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Naive Bayes, to generate predictions based on the dataset. The results obtained from these algorithms are provided in the next figure::

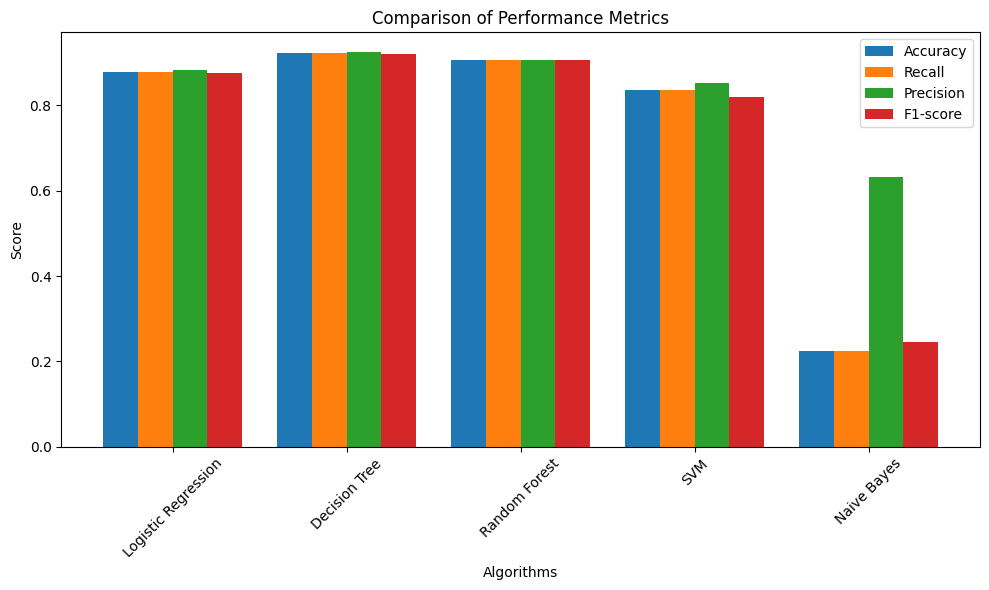


Figure : Performance metrics for the KAEU dataset

The performance of the model was anticipated to be relatively high given the limited availability of data and informative features. However, the Naive Bayes algorithm exhibited unsatisfactory results with an accuracy of 22% and an F1-score of 24%. This can be attributed to the insufficiency of features as well as the underlying assumption of independence made by Naive Bayes.

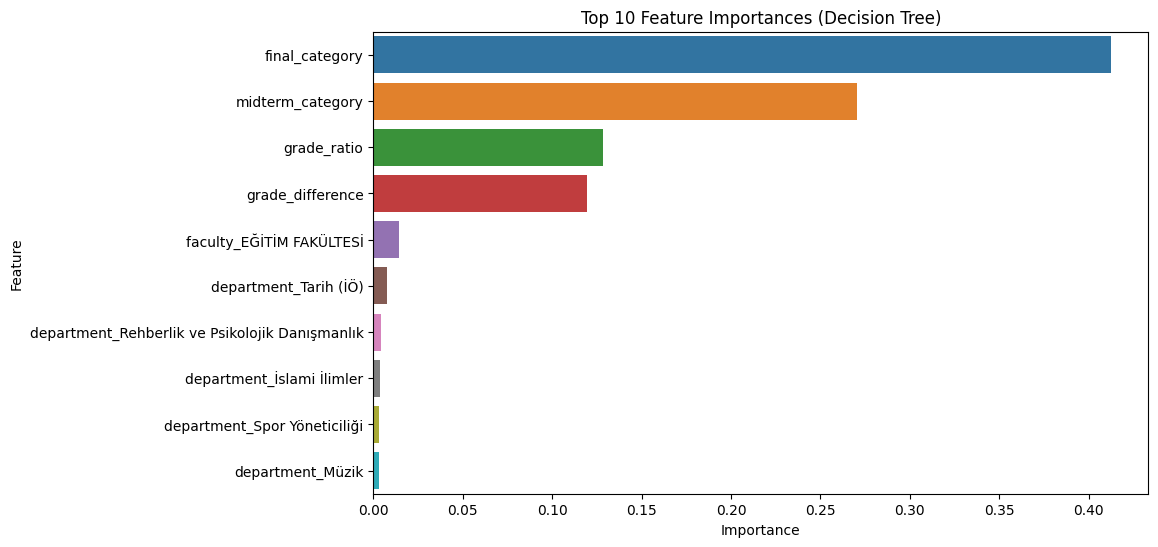


Figure : Top-10 important features (Decision Tree)

Based on the findings presented in Figure 2, it can be observed that the features "final\_category" and "mdterm\_category" have a significant impact on the prediction accuracy of the Decision Tree model. This outcome was anticipated due to the scarcity of informative data. Furthermore, these two features exhibit a strong correlation, highlighting their importance in the prediction process.

3.2**BUC Dataset**

All the models used on this dataset seem to perform well on predicting the students’ performance. In the tables below we can see the performance of the models used for each of our classes. The worst performing one is the MLP.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class ‘HIGH’** | | | | |
| Model | Precision | Recall | F1-Score | Support |
| Linear Regression | 0.87 | 1.00 | 0.93 | 20 |
| Decision Tree | 0.83 | 0.95 | 0.88 | 20 |
| Random Forest | 0.87 | 1.00 | 0.93 | 20 |
| MLP | 0.83 | 0.9 | 0.86 | 21 |

Table : Model Performance on 'HIGH' class

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class ‘LOW’** | | | | |
| Model | Precision | Recall | F1-Score | Support |
| Linear Regression | 1.00 | 0.73 | 0.84 | 11 |
| Decision Tree | 0.88 | 0.64 | 0.74 | 11 |
| Random Forest | 1.00 | 0.73 | 0.84 | 11 |
| MLP | 0.75 | 0.60 | 0.67 | 10 |

Table : Model Performance on 'LOW' class

For the linear regression we used the coefficients to determine the importance of the features. Based on that, the most significant features for the linear regression are the Attendance, the CGPA, the grade of first exam and the prerequisite subject grade.

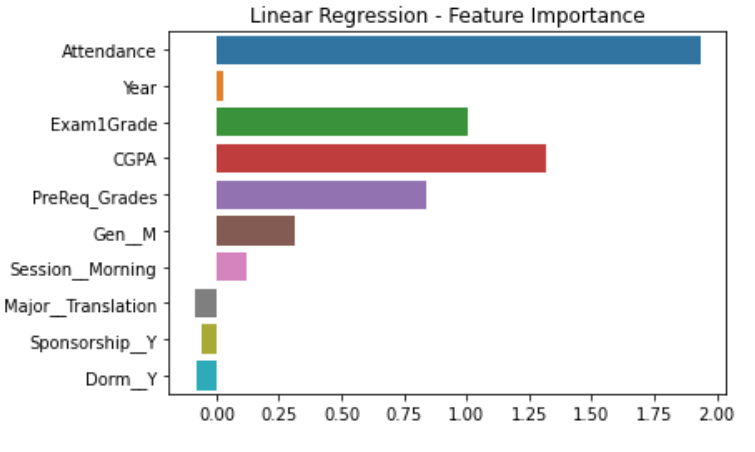


Figure : Feature Importance of Linear Regression

For the tree models, the random forest and the single decision tree, we determine the importance of the features based on the reduction in the gini criterion.

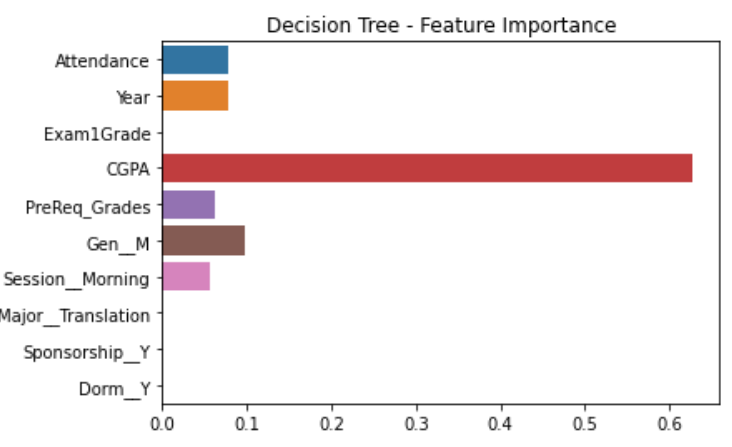


Figure : Feature Importance of Decision Tree

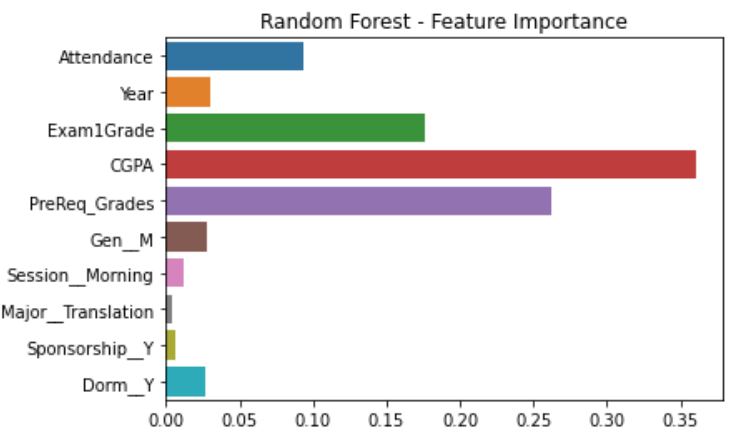


Figure : Feature Importance of Random Forest

The tree models seem to give the highest importance on CGPA and less on Attendance in contrast to what we see on linear regression where attendance was the most important feature.

The single decision tree can also give us a set of rules that can be leveraged by tutors to understand which students are going to perform low in their exam.

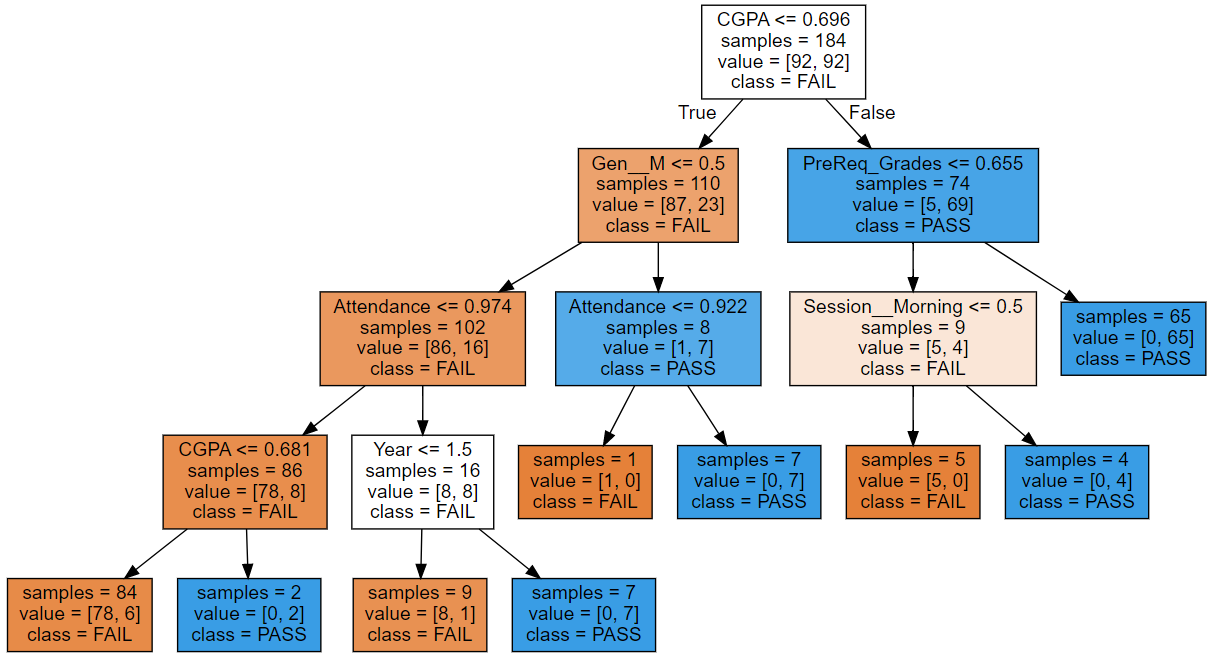


Figure : Graphical representation of decision tree

Finally using Shapley values, we also got the most important features for the MLP model, like the tree models, give a high significance on GCPA while considering attendance as the least significant.

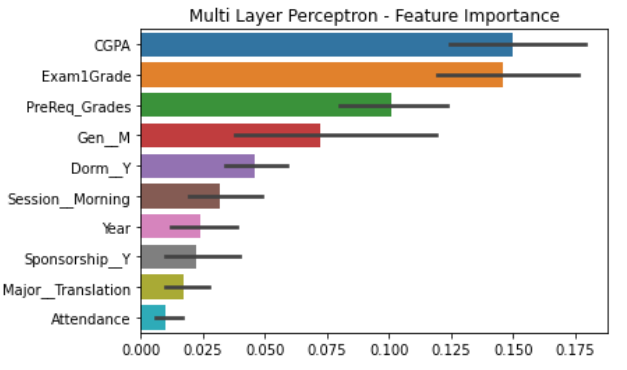


Figure : Shapley values of MLP

All in all, the four models seem to agree on the importance of three of the features, the CGPA score, the grade of the first exam and the grade of the prerequisite subjects. Attendance was the most on the linear model but it seems that the rest of the models don’t give it such a high importance.

**CONCLUSIONS**

Our analysis revealed that several factors significantly influence educational outcomes. Among the most important features are First/Mid-Term Exam Scores, CGPA, Date of Submission, Attendance, Department, Educational Background, Wealth/Relationship Status and Age. It is worth noting that data availability poses a challenge in this domain. The shortage of comprehensive student performance data limits the depth of analysis and hinders the exploration of additional relevant features. To address this limitation, efforts should be made to collect and curate rich datasets with a wide range of variables.

**ACKNOWLEDGMENTS**

This paper was built for educational purposes for the lesson “Advanced Machine Learning” of the MSc program “Data & Web Science” of Department of Informatics at Aristotle University of Thessaloniki.

**REFERENCES**

[1] Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. 9, 11 (2022). <https://doi.org/10.1186/s40561-022-00192-z>

[2] Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset (2017) Sci. Data 4:170171 doi: 10.1038/sdata.2017.171

[3] Khan, I., Ahmad, A.R., Jabeur, N. et al. An artificial intelligence approach to monitor student performance and devise preventive measures. Smart Learn. Environ. 8, 17 (2021). <https://doi.org/10.1186/s40561-021-00161-y>

[4] Amirah Mohamed Shahiri, Wahidah Husain, Nur’aini Abdul Rashid, A Review on Predicting Student's Performance Using Data Mining Techniques, Procedia Computer Science, Volume 72, 2015, Pages 414-422, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2015.12.157>

[5] T. Mishra, D. Kumar and S. Gupta, "Mining Students' Data for Prediction Performance," 2014 Fourth International Conference on Advanced Computing & Communication Technologies, Rohtak, India, 2014, pp. 255-262, doi: 10.1109/ACCT.2014.105