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
mmalamat@csd.auth.gr

Papanikolaou Stefanos
School of Informatics
Aristotle University
Thessaloniki, Greece
spapanik@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 RELATED WORK

The field of educational data mining has seen various approaches to predict student performance using machine learning algorithms. A study by Yağcı (2022) proposed a model to predict final exam grades of undergraduate students based on their midterm exam grades, Department data, and Faculty data. This study utilized machine learning algorithms such as random forests, nearest neighbor, support vector machines, logistic regression, Naïve Bayes, and k-nearest neighbor algorithms, and achieved a classification accuracy of 70–75%.

In another line of work, Kuzilek et al. (2017) have developed the Open University Learning Analytics dataset (OULAD), which comprises demographic data together with aggregated clickstream data of students' interactions in the Virtual Learning Environment (VLE). This enables the analysis of student behavior, represented by their actions. The dataset includes information about 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks. The dataset can be used in various scenarios, such as the evaluation of predictive models for predicting student assessment results and final course results and comparison of models with other models developed by other researchers.

Khan et al. (2021) presented an artificial intelligence approach to monitor student performance and devise preventive measures. The study used the UC Irvine Machine Learning Repository dataset and developed a model that incorporated learning trajectories and student characteristics to identify patterns relevant to learning analytics.

Mohamed Shahiri et al. (2015) conducted a review on predicting student's performance using data mining techniques, focusing on how the prediction algorithm can be used to identify the most important attributes in a student's data.

Mishra et al. (2014) mined students' data for prediction performance, with the goal of providing timely counseling and coaching to increase student success.

Cortez (2014) provided a student performance dataset in the UCI Machine Learning Repository, which includes student grades, demographic, social and school-related features collected by using school reports and questionnaires.

3 APPROACH

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.

3.1 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)

3.2 UCI Dataset [6]

This data approaches student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).

1. School - Student's school.
2. Sex - Student's sex.
3. Age - Student's age.
4. Address - Student's home address type.
5. FamilySize - Family size.
6. ParentsStatus - Parent's cohabitation status.
7. MothersEducation - Mother's education.
8. FathersEducation - Father's education.
9. MothersJob - Mother's job.
10. FathersJob - Father's job.
11. Reason - reason to choose this school.
12. StudentsGuardian - Student's guardian.
13. TravelTime - Home to school travel time.
14. StudyTime - Weekly study time.
15. Failures - Number of past class failures.
16. SchoolSupport - Extra educational support.
17. FamilySupport - Family educational support.
18. ExtraClasses - Extra paid classes within the course subject.
19. Activities - Extracurricular activities.
20. NurserySchool - Attended nursery school.
21. HigherEducation - Students want to take higher education.
22. InternetAccess - Internet access at home.
23. Relationship - With a romantic relationship.
24. FamilyRelationships - Quality of family relationships.
25. FreeTime - Free time after school.
26. GoingOut - Going out with friends.
27. DailyAlcohol - Workday alcohol consumption.
28. WeeklyAlcohol - Weekend alcohol consumption.
29. Health - Current health status.

30. Absences - Number of school absences.

The dataset has grades for the first, second and final periods. We evaluated and interpreted the students' status and predicted the possibility of success or failure for the students. The feature to predict was the final period, and is the deciding factor of the pass or fail for the student.

Before running any machine learning algorithms we checked the correlation of the features and dropped. Then we splitted the group into 80% training set and 20% evaluation set. And then we made 2 new sets, one set, the first used the grades of the first and second period and predicted the final period grade. But because the grades had a high correlation, the experts should have a prediction from the start of the year and not at the middle of it. We made a dataset and trained algorithms that will predict the success or failure of the students without the mid year grades.

The machine learning algorithms we used are the following:

- Linear Regression
- Decision Tree
- Random Forest
- SVM

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

The Open University Learning Analytics Dataset (OULAD) Dataset, is a publicly available dataset that has been used for research in the field of learning analytics. It was made available by The Open University, which is a public distance learning and research university in the UK.

It contains anonymized student data and their interactions with online learning materials. The dataset includes demographic data about the students, their interactions with the virtual learning environment (like clicks), their academic results, and other related data.

The dataset is structured into several CSV files that include:

- Assessments: Information about course assessments.
- Courses: Information about the courses.
- StudentAssessment: Student's scores in course assessments.
- StudentInfo: Information about the students.
- StudentRegistration: Information about the registration of students.
- StudentVle: Information about student's interactions with the virtual learning environment (VLE).
- Vle: Information about the virtual learning environment.

For the purpose of this project, only StudentInfo and StudentAssesment data were incorporated into the problem as the other data were not comparable to the other datasets structure due to the online learning nature of the dataset.

Below are listed the features used from the dataset:

- code_module: Identification code for the module.
- code_presentation: Identification code of the presentation.
- gender: Student's gender.
- region: Geographic region where the student lived.
- highest_education: Highest education level on entry.
- imd_band: Index of Multiple Deprivation band. It is an index showing the relative deprivation (essentially a measure of poverty) of small areas.
- age_band: Band of the student's age.
- num_of_prev_attempts: Number of times the student has attempted the module.
- studied_credits: Total number of credits for the modules the student is studying.
- disability: Indicates whether the student has declared a disability.
- first_assessment_score: The score achieved by the student in his first assessment.
- date_submitted: The index of the date the student submitted his first assessment.
- is_banked: A status flag indicating that the assessment result has been transferred from a previous presentation.

Initially, an exploratory data analysis was performed to extract useful insights that could be used at a later stage during experimentation.

Moreover NaN values were identified into the dataset and more specifically in the imd_band feature. Those values were imputed by performing KMeans clustering and placing the mode of each cluster.

Additionally, all categorical values were either transformed by setting them with a sequential order number (from categorical to numerical), or by performing One Hot Encoding to the ones that the previous technique could not be applied to.

After the preprocessing was completed, a baseline benchmarking was performed with all preprocessed features (without any techniques applied) and with default hyperparameters for 10 model classifiers. Those were:

1. Logistic Regression
2. Random Forest
3. Gradient Boosting
4. Decision Tree
5. K-Nearest Neighbors
6. Support Vector Machine
7. Gaussian Naive Bayes
8. Linear Discriminant Analysis
9. Multi-Layer Perceptron
10. Light GBM

Those models were evaluated with Accuracy, Precision, Recall and F1 metrics and the 5 best of them were selected according to their F1 Accuracy, to proceed to the upcoming experimentations.

The next experiments were incorporating cost into the problem. A cost matrix was defined that penalized with a ratio 1:3 more the misclassified students that actually failed. This aligns with the problem definition that we want as many failed students as possible to be found in order to focus more on them and reduce failure rates in general.

Expected cost minimization was performed along with experimentation in instance sampling by using both undersampling and oversampling methodologies. In terms of undersampling, random undersampling was performed and for oversampling, the SMOTE (Synthetic Minority Oversampling TEchnique) was used.

Finally, the best model in terms of cost was selected to perform shapley values extraction in order to find useful model interpretations, as to how the model is affected by each feature and its various possible values.

3.4 **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.

4 RESULTS

4.1 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::

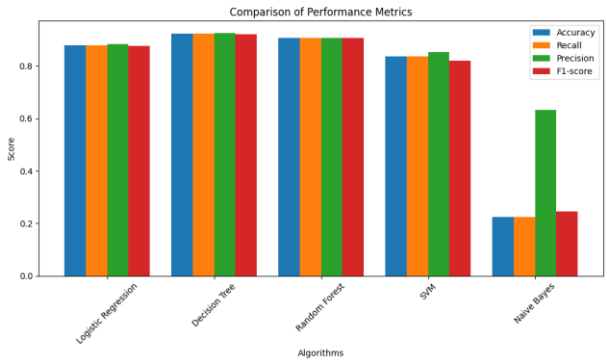

Figure 1: Performance metrics for the KAEU dataset

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.

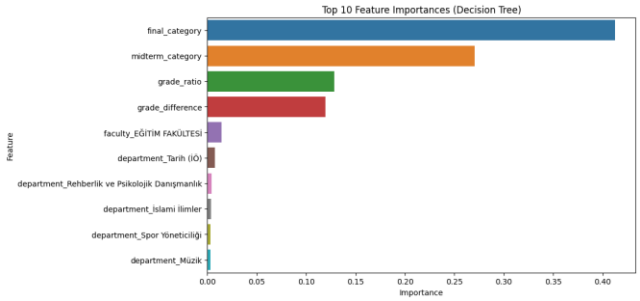

Figure 2: Top-10 important features (Decision Tree)

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.

4.2 UCI Dataset

The dataset contained a large number of features and a relative small number of features. We dropped features that had a high correlation and were not able to be transformed from a nominal to a numerical set. We then used the classification algorithms of Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM). The results are shown on the next Table.

Course 'Math'				
Model	Precision	Recall	F1-Score	Support
Linear Regression	0.87	0.89	0.9	42
Decision Tree	0.84	0.91	0.87	36
Random Forest	0.96	0.95	0.96	34
SVM	0.94	0.93	0.94	40

Table 1: Model Performance on 'Math' course.

Course 'Portuguese'				
Model	Precision	Recall	F1-Score	Support
Linear Regression	0.93	0.88	0.90	39
Decision Tree	0.86	0.85	0.86	44
Random Forest	0.92	0.86	0.89	42
SVM	0.92	0.92	0.91	50

Table 2: Model Performance on 'Portuguese' course.

Used the linear regression and the random tree model to determine and interpret the importance of the features as shown on the next figures.

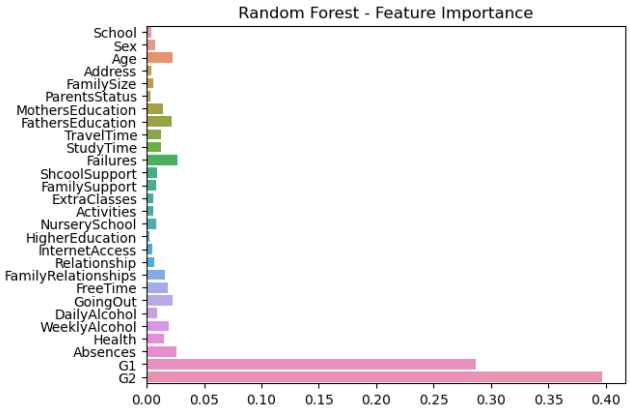

Figure 3: Feature Importance of Random Forest Math class

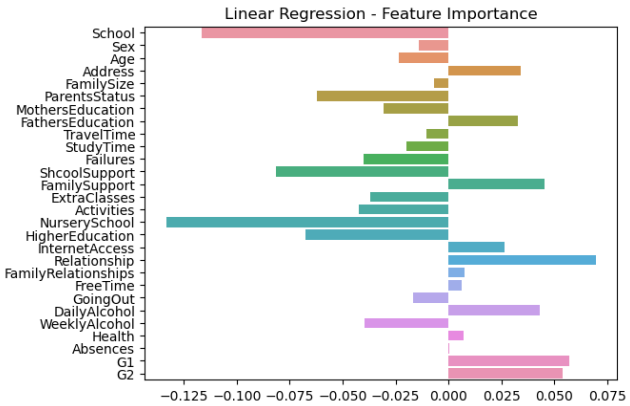

Figure 4: Feature Importance of Linear Regression Math class

The most important feature is the G1 (first period) and G2 (mid period) grades for the students. We observe that those two features are the most important ones for the forest algorithm. Also those are the only features that are not known at the start of the school year. So we then produced some data and checked the importance of the features except the G1 and G2. A model that can be used to predict the results of the students at the start of the year. The results are shown on the next tables.

Course 'Math'				
Model	Precision	Recall	F1-Score	Support
Linear Regression	0.56	0.57	0.56	42
Decision Tree	0.52	0.52	0.52	40
Random Forest	0.59	0.59	0.58	40
SVM	0.60	0.61	0.60	41

Table 3: Model Performance on 'Math' course.

Course 'Portuguese'				
Model	Precision	Recall	F1-Score	Support
Linear Regression	0.68	0.82	0.69	65
Decision Tree	0.67	0.67	0.67	65
Random Forest	0.67	0.76	0.66	67
SVM	0.69	0.75	0.68	66

Table 4: Model Performance on 'Portuguese' course.

Used the linear regression and the random tree model to determine and interpret the importance of the features as shown on the next figures.

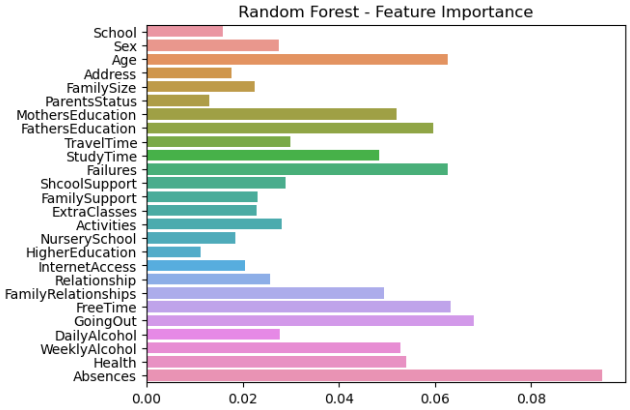

Figure 5: Feature Importance of Random Forest Math class

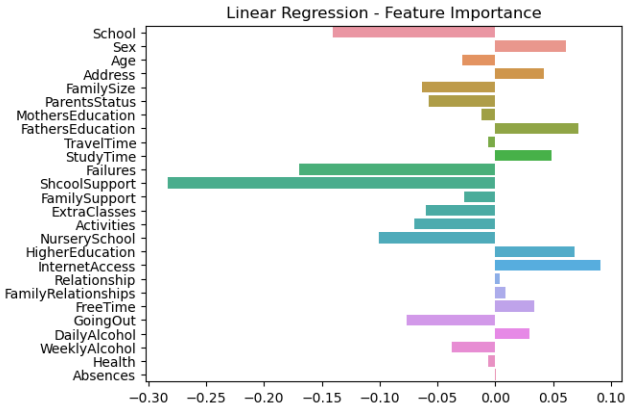

Figure 6: Feature Importance of Linear Regression Math class

Without the G1 and G2 features, the most important feature for the random forest is the Absences and for the linear regression the SchoolSupport. The algorithms overall have a reduced performance.

4.3 OULAD Dataset

After performing benchmarking with the default preprocessed dataset (without any other techniques applied), the five best performing classifiers were :

- Gradient Boosting Trees
- Light GBM
- Multi Layer Perceptron
- Logistic Regression
- Linear Discriminant Analysis

Baseline Benchmarking				
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.69	0.70	0.81	0.75
Random Forest	0.69	0.72	0.77	0.74
Gradient Boosting	0.70	0.71	0.83	0.76
Decision Tree	0.61	0.67	0.64	0.66
K-Nearest Neighbors	0.64	0.68	0.71	0.69
SVM	0.64	0.65	0.85	0.73
Gaussian Naive Bayes	0.63	0.66	0.76	0.71
Linear Discriminant Analysis	0.68	0.69	0.82	0.75
MLP	0.70	0.72	0.79	0.75
LightGBM	0.70	0.71	0.81	0.76

Table 5: Model Performance with default hyperparameters and no other techniques applied

After the model selection, cost was incorporated into the problem and the evaluation is cost centric. The cost matrix is set with a ratio 1:3 in order to penalize more misclassified students that failed. The best baseline performing model in terms of cost was Multi Layer Perceptron.

Baseline Cost Benchmarking						
Model	TN	FN	FP	TP	Missed Fails %	Cost
Gradient Boosting	1200	505	1050	2588	46.6	3655
LightGBM	1257	573	993	2520	44.13	3552
MLP	1293	631	957	2462	42.53	3502
Logistic Regression	1169	570	1081	2523	48.04	3813
Linear Discriminant Analysis	1131	555	1119	2538	49.73	3912

Table 6: Baseline Models evaluation in terms of cost

After applying expected cost minimization, all misclassifications of failed students were radically reduced but this also affected the FN predictions, which means the model tends to predict more students that will fail which won't. But the main focus here is to not miss students who are going to fail.

Expected Cost Minimization						
Model	TN	FN	FP	TP	Missed Fails %	Cost
Gradient Boosting	2053	2111	197	982	8.75	2702
LightGBM	2008	1888	242	1205	10.75	2614
MLP	2059	2152	191	941	8.48	2725
Logistic Regression	2054	2222	196	871	8.71	2810
Linear Discriminant Analysis	2044	2176	206	917	9.15	2794

Table 7: Expected Cost Minimization model evaluation

As shown in Table 7 the best model by applying expected cost minimization was the Light GBM model achieving 2614 cost and it missed about 10% of the students that actually failed.

The last experiment performed was to handle the class imbalance present in the dataset, by applying oversampling and undersampling techniques (SMOTE and Random Undersampling).

Expected Cost Minimization with UnderSampling						
Model	TN	FN	FP	TP	Missed Fails %	Cost
Gradient Boosting	2184	2646	66	447	2.93	2844
LightGBM	2126	2366	124	727	5.51	2738
MLP	2210	2804	40	289	1.77	2924
Logistic Regression	2151	2641	99	452	4.39	2938

Linear Discriminant Analysis	2149	2627	101	466	4.48	2930

Table 8: Expected Cost Minimization with UnderSampling model evaluation

Expected Cost Minimization with SMOTE						
Model	TN	FN	FP	TP	Missed Fails %	Cost
Gradient Boosting	2175	2593	75	500	3.33	2818
LightGBM	2053	2069	197	1024	8.75	2660
MLP	1540	963	710	2130	31.55	3093
Logistic Regression	2076	2283	174	810	7.73	2805
Linear Discriminant Analysis	2057	2215	193	878	8.57	2794

Table 9: Expected Cost Minimization with SMOTE model evaluation

For both approaches Light GBM model performed best in terms of cost. In terms of Missed Fails percentages the undersampling method seemed to perform better than the oversampling one but in LightGBM the cost was better with SMOTE, achieving 2660 cost.

The best cost (2614) in general was achieved again by Light GBM without any sampling applied. This best model was used to extract the shapley values and tried to interpret why and how these predictions are affected by the input features. A global interpretation of the model showed that the first exam score played the most significant role in the output prediction, which is reasonable. The second most impactful feature was the date the exam was submitted in the virtual learning environment. Other significant factors were the educational background, the specific module the student was attending, the imd_band feature which shows the deprivation in the area the student is in and the age.

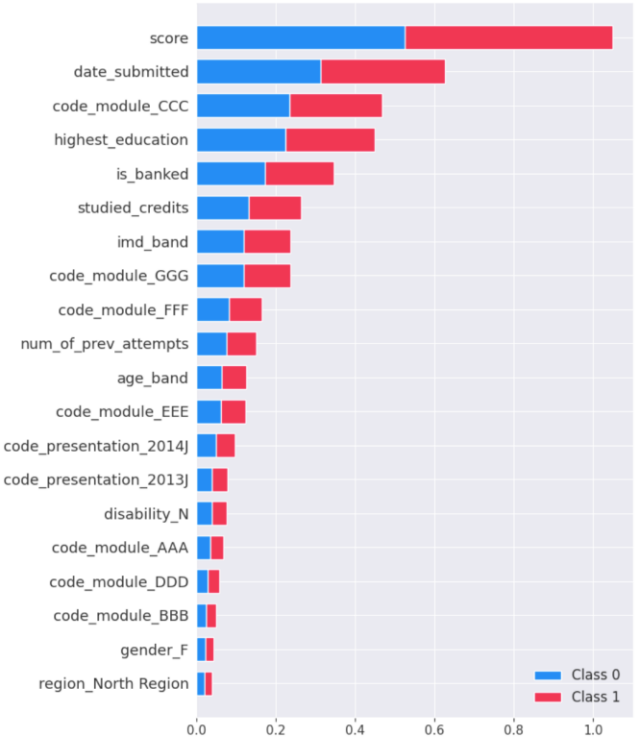

Figure 7: Mean absolute SHAP value (shows average impact on model output magnitude)

4.4 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 10: 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 11: 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 the first exam and the prerequisite subject grade.

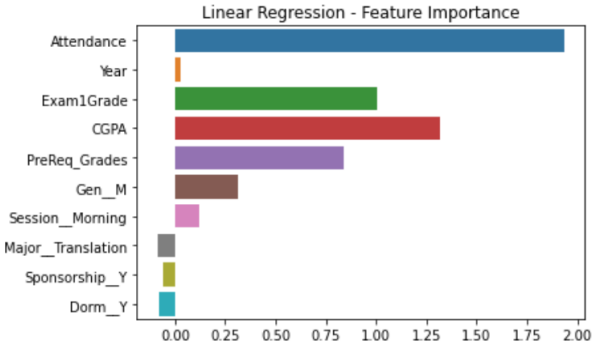

Figure 8: Feature Importance of Linear Regression

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.

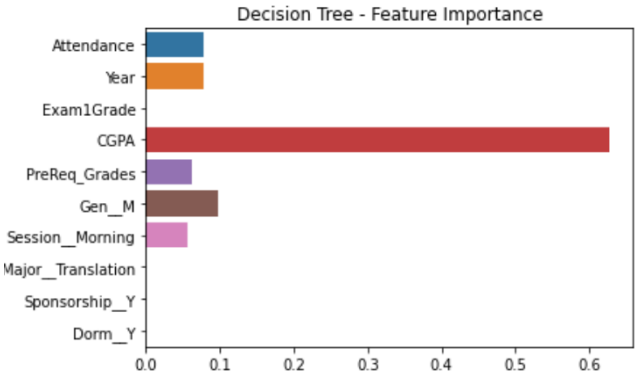

Figure 9: Feature Importance of Decision Tree

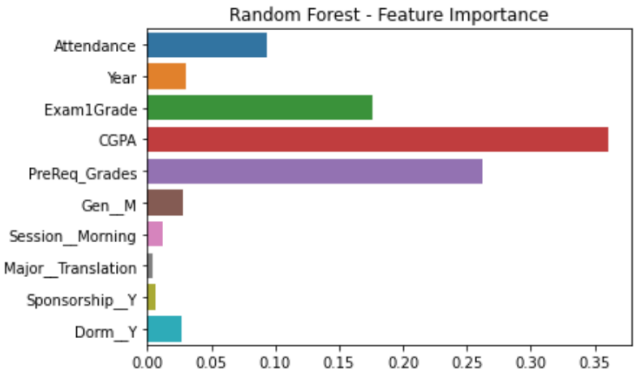

Figure 10: Feature Importance of Random Forest

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.

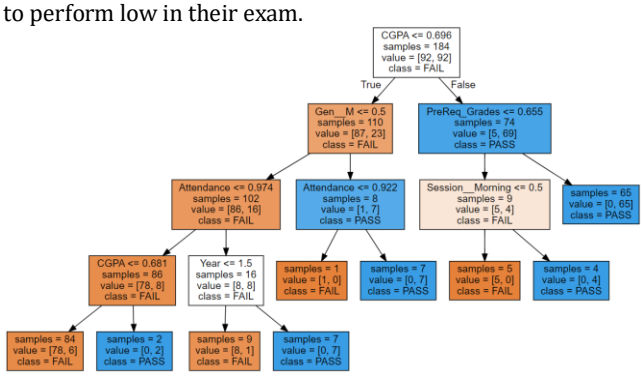

Figure 11: Graphical representation of decision tree

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.

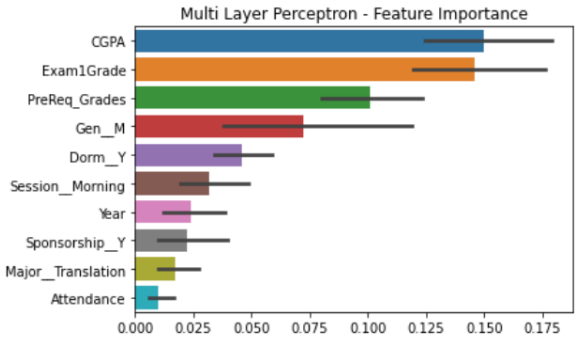

Figure 12: Shapley values of MLP

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.

5 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.

Moreover, a cost focused approach seems to better tackle the problem at hand which is identifying students that have a tendency to fail. Educators and institutes could adjust this matrix to fit their available resources and optimize the number of students they should focus on.

The shortage of comprehensive student performance data limits the depth of analysis and hinders the exploration of additional relevant features. It is worth noting that data availability also poses a challenge in this domain. 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. All related work can be found at: https://github.com/MnAppsNet/ML_on_student_performance_assignment

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

[6] Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.