Using Machine Learning With Student Performance Predictions

Raunak Chitre
Virginia Tech
raunakc24@vt.edu

Abstract

In order to improve academic results for students, it is important to predict student performance to give instructors more insight on how to improve their teachings. The data was investigated to find any factors that would lead to high final grades. To predict student performance, machine learning models were used based on features from the Student Performance in Exams dataset such as gender, race/ethnicity, and parental level of education. Machine learning models such as logistic regression and decision trees were used. To evaluate the model, metrics such as accuracy, recall, precision, and F1 score were used. These results would help educators find which students might need additional potential help in the future. This can also be used to prevent students from dropping out from class and/or failing.

1. Introduction

In the field of educational research, student performance prediction has become an important aspect of it. Machine learning can provide a way to understand what factors lead to a student passing or failing a class. By looking into the features from the dataset, the goal of the study is to predict the student's final grades.

1.1. Context/Motivation

Future opportunities, for students, can often times be earned through education. In many classes, there are often different levels of students, which means some students do great, while other students struggle. Typically, there are factors that can be investigated to see if there are any trends. Addressing the discrepancies in the factors can improve the chances that a student who is struggling, can end up succeeding.

In the present, many instructors rely on observations on the student and test grades to assess whether a student is struggling. These methods are solid, however they require a test to be given to a student and miss external reasons as to why a student could be struggling. These external influences being missed can lead to an incorrect evaluation of a student.

As an educator, relying on test grades of students can be reactive meaning that that help would be given to the student after they have had a rough start. Assessments also depend teacher by teacher. Some teachers intentionally make assessments more difficult, which lead to inconsistencies to identify students whom are struggling.

1.2. Objective

The objective of this project was to create a machine learning model that could predict a students performance or final grade based on different academic and non-academic factors. This problem was both a regression and a classification problem. By predicting a student's numeric grade, it falls under a regression problem, and by ordering students into different categories, it falls under a classification problem. Another problem that was addressed was to identify students who are having difficulties at an early time.

1.3. Related Works

There were a few studies that are similar to what this project aims to solve. The researchers used different techniques such as machine learning algorithms and data processes to identify students who were at-risk at failing class. They used a variety of factors such as previous academic records and community and home features. The following studies align with the objective of this project and will give important insights on the strengths of the studies.

The first paper was conducted by a member, named Mustafa Yağci of the Faculty of Engineering and Architecture at the Kirşehir Ahi Evran University located in Turkey [6]. This study focused on predicting the final grades of students using machine learning algorithms. Some of the factors that were taken into consideration were the faculty and department data and midterm exam scores. The dataset was from the Turkish Language-I course at a university with 1,854 students in the class. Some of the machine learning algorithms that he used were random forest, logistic regression, Naïve Bayes, support vector machines, and k-nearest neighbor. The classification accuracy from the models had a range from about 70% to 75%. An aspect of the study that is similar to my motivation for the project is that there was a 9 week gap between the midterm and final exam. This meant that the educators had this long time period to help students who would potentially fail the final exam.

The next paper also talks about using machine learning to predict student performance in the form of a Grade Point Average. A group of researchers used a dataset that had 222 students from different universities [3]. A difference between the dataset from this study and the previous one is that this study's dataset had five groups and 18 attributes that were in the groups. The different categories of the groups were academic, lifestyle, supporting, socioeconomic factors and student information. They used feature selection to find which features were important and had the biggest impact on the final grades. The machine learning models that were used were linear regression, support vector machines, random forest, and Gradient Boosted. As part of the feature selection process, they decided to lower the amount of features that they used from 18 to 12. To investigate if feature selection would make a difference, they ran their different models with and without feature selection, and it was seen that linear regression was the best model because of its MAE being very low at 0.226941. The next closest models were the Random Forest and the Gradient Boosted.

The paper from Applied Sciences went into predicting student performance based on machine learning and other techniques [4]. They separated the students by the different levels of school or colleges that they were in. From the study, it was seen from the researchers that models like decision trees, random forest, Naïve Bayes, and support vector machines were good to evaluate which student performance. The researchers noted that support vector machine was the best model, while decision tree and random forest were close to being the best. These models fall under the supervised learning category, and it was seen that unsupervised learning models were avoided, since they had a lower accuracy. Something to look forward to in the future is the use of neural networks since the researchers were aware that neural networks were not covered.

1.4. Limitations or Contrast of Current Practice

Some studies matched the objective of this project, as well as had some similarities, but there were still some important limitations. Some examples of these limitations are the datasets being too small, some features having too much of an impact on the final grade prediction, and a restricted amount of machine learning algorithms being explored.

The study by Esmael Ahmed, whom works at the Wollo University, had similarities to this project. However, there were some limitations that were seen and acknowledged by the researcher [1]. While the dataset contained 32,005 students, it was from Wollo University and the Kombolcha Institute of Technology, which could cause some biases because it doesn't take into account of other schools. As part of their feature selection, they did pick a variety of features from different categories such as region, number of previous attempts, studied credits, and entrance result. Although these are good features, it does not take into consideration of factors such as student background and out-of-school activities. The study went over precision, recall, and accuracy as metrics to identify whether a model was accurate or not. Throughout the paper, the researcher amplified the accuracy metric for the different models but did not talk about precision and recall in the same amount. It is necessary to talk about recall and precision in the case that the dataset is not balanced.

While this study had some positives, the limitations are important enough to be addressed. Due to the limitations, this project aims to include features from different categories and emphasize the different metrics. Features that relate to being external to school are just as important as features relating to academic. For example, the paper from the Novel Intelligent and Leading Emerging Sciences Conference had features relating to the lifestyle and socioeconomic factors. Following this format for my dataset would be beneficial for my project, as well as incorporating different aspects from the studies.

1.5. Impact

This project can have a huge impact on educators because of the ability to identify a student who might potentially fail a class. Being able to predict student grades would allow students to receive help at an earlier time. If successful, this would also lead to lower dropout rates and lower failing grades which positively impacts instructors, since their validity would increase. A difference is that this project looks at features from different areas.

2. Approach

In the project, a model was created to determine if students would potentially have a grade above a C or below a C. Different machine learning algorithms were used such as logistic regression, K-Means, and decision trees.

2.1. Implementation Information

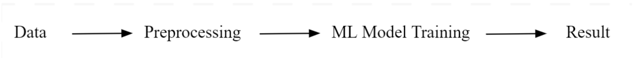

Figure 1. Overview of the Model

To begin, the data was preprocessed, so that categories like parental level of education and race/ethnicity used label encoding, so that those categories could have numerical values. By using feature engineering, a new feature was created called ”gender prep”. This feature combined the gender and test preparation course categories. A new column called ”average score” was created that took the average score from the three subject’s test grades.

To solve the problem, the first machine learning model that was used was logistic regression. The model was used to identify if a student would have a grade above or below a C. Feature importance was also done with logistic regression to see which features had a higher impact. Another model that was used was a decision tree. This was used because of its easy ability to create rules about which way it should go down the tree. It is able to pick the best available feature and continue. A third model was used which is K-Means clustering. It was used to group students based on their test scores, for example their math and reading scores.

The reason these models were chosen and believed to be successful is because of many reasons. Logistic regression was seen to be successful because its strengths in predicting whether one value is true or not. Decision Trees were seen to be successful for the way it makes predictions and K-Means would group data together. Additionally, the dataset had strong features like parental level of education and status of test preparation course. Basic data findings were done before implementing the models and that helped with understanding the data in a better way.

With the approach that was taken in this project, it combines both unsupervised and supervised learning. K-Means was also used to see if there were any patterns with the students’ performance on exams. Creating a new feature from two other features was also a point of difference. These were some new aspects of this project, however this project followed some of the main parts from the other studies.

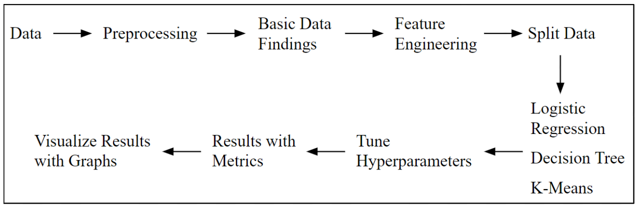

Figure 2. Visualization of Workflow

2.2. Challenges Faced

A challenge that I anticipated was deciding how to predict the students’ final grades because there were many options on how it could have been done. Another challenge that required thought was improving the accuracy. Overfitting was something that was thought about extensively to avoid with models like a decision tree. The dataset also had a good number of features, so understanding which features could impact the model was something else that was kept in mind. It was important to make sure that some of the features did not outshine the other features by a large margin.

An initial challenge that I faced was with the original dataset that was used to begin with. That dataset did not have enough data for it to be a viable option to use with the machine learning models. After running the model with the dataset, metrics such as the accuracy and precision always came out to be 1.0. This showed that the model was overfitting instead of learning a pattern that could be used to make predictions. A second challenge that I faced was with predicting the students’ final grades. To start, a thought was to predict a numeric value as the grade for each student. Since this project used logistic regression, that was not the best option. A solution was to designate whether a student was above or below the range of the C letter grade. This was decided since the letter grade ”C” is usually the passing mark for important classes. Another reason this was a viable option is because it is more important to see if a student is struggling or succeeding rather than their actual predicted grade.

The first thing I tried with the features was create some graphs that could be used to see if there were any relations between the features. Furthermore, I did feature engineering where I created a new feature based on two features. This slightly helped the accuracy improve. With the initial dataset that was used, I went through many iterations of feature engineering, but it did not make an effect on the final accuracy and other metrics. From this, it was clear that the dataset was not a sustainable option. I also changed the splitting of the data to a 70-30 split, but it did not make a difference with the metrics with logistic regression. Next, with the accuracy, the first thing that I did to improve it was to look at the hyperparameters in the decision tree. I adjusted the max depth hyperparameter to see if the decision tree would perform better if it had more or less decisions to make.

3. Experiments and Results

There were results from the data findings and the machine learning models that were noteworthy. A variety of metrics were used to classify if the models were successful or not and will be looked at in this section.

3.1. Basic Data Findings

In order to understand the dataset better, I decided to visualize the data through graphs. The first visualization is a bar graph that calculated the average score for each of the subjects which was compared with females and males. It is clear that females had a higher average score on reading and writing by about 5-7 points. However, the males had a higher average math score by about 3-4 points.

The second visualization I created was a heatmap to show the correlation between the different test scores from

Notation	Description of Notation
X	Features used during training
y	Predicting if above_C
average_of_score	Average scores across math, reading, and writing
gender_prep	New feature combining gender and prep course
Accuracy	Metric for percentage of current predictions
Precision	Metric for correct positive predictions
Recall	Metric for percentage of correct positive
F1-Score	Metric for mean between recall and precision

Figure 3. Notation Table

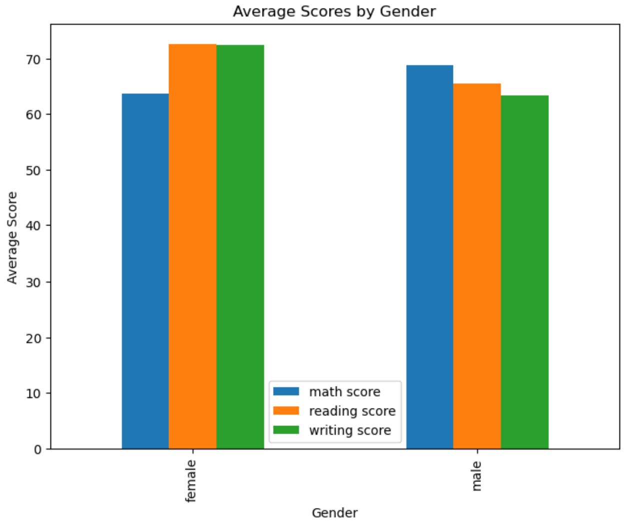

Figure 4. Average Scores by Gender For Each Subject

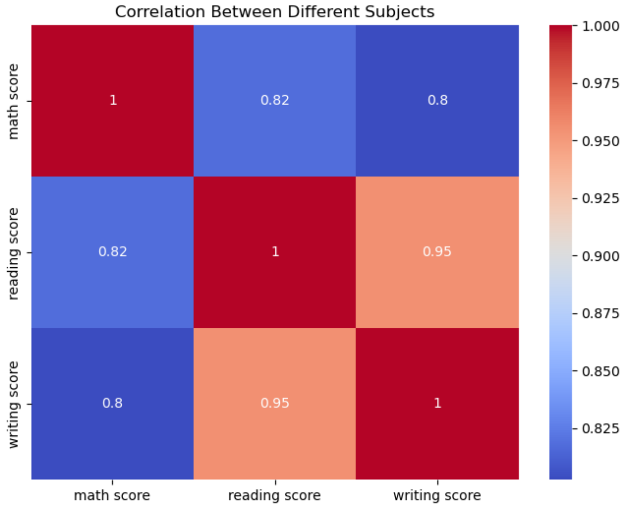

Figure 5. Heatmap of Correlation Between Test Scores

the subjects such as math, writing, and reading. Reading and writing had the highest correlation which was 0.95. The correlation between the reading and math score was also strong at 0.82 and after that came the correlation between the writing and math score at 0.8. It is seen that all three subjects have a strong correlation between each other.

Following that, the third data visualization was a stacked bar graph that looked at the student's parental level of education and the students' completion for the test preparation course [2]. From the graph, it is seen that the proportions are relatively equal, which shows there is no clear pattern. I thought that the higher the level of parental education, the test preparation course would be completed. Nonetheless, there was no relation between these two features.

3.2. Logistic Regression

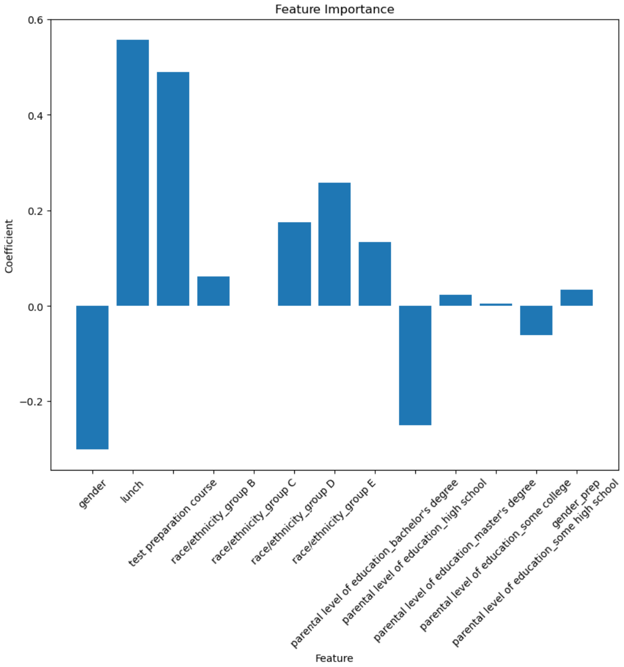

Figure 6. Feature Importance For Logistic Regression

For the Logistic Regression model, I used feature importance to see which feature had a high impact. From the graph it is clear lunch and the test preparation course had the highest coefficient. Gender and the parental level of education with high school had the lowest coefficients.

To see if a student was below or above a C, I used a binary output that indicated that a 1 was above a C and a 0 was below a C. The model had an accuracy of 68%. For students who got below a C, the precision was 70%, recall was 83%, and f1-score was 76%. This shows the model was solid at identify students who did not get a C. However, for students who got above a C, the precision was 60%, recall was 41%, and f1-score was 49%. Based on these results, the logistic regression model struggled to identify students who got above a C.

3.3. Decision Tree

The decision tree model had an accuracy similar to the logistic regression model at 66%. The precision for students below a C was 69%, recall was 82%, and f1-score was 69%. Similarly to the logistic regression model, the decision tree had a difficulty predicting students who were above a C. This was because the precision was 56%, recall was 39%, and f1-score was 46%.

The decision tree made its splitting decisions based on the features from the dataset, which is similar to other decision trees. If a student was greater than the node's boundary then it would go to the right and if the opposite was true, it would go to the left. The max depth of the decision tree was 4, so the diagram is quite big.

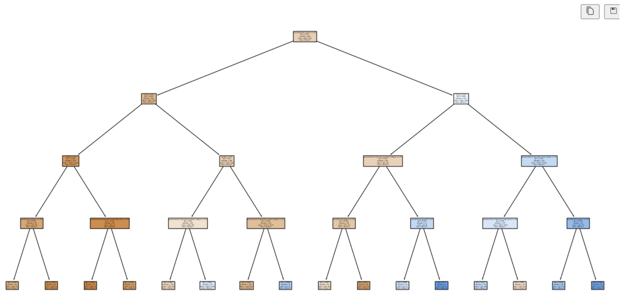

Figure 7. Decision Tree Diagram

3.4. K-Means Clustering

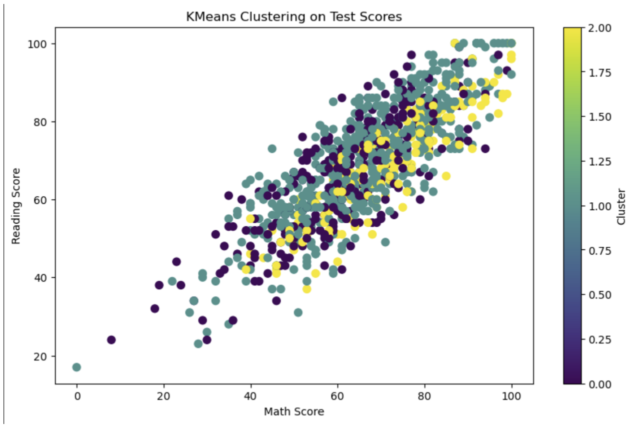

Figure 8. Decision Tree Diagram

K-Means was used to see if clusters could be created based on the subject scores for students. Specifically, the math and reading scores were used. In the graph, it is seen that the yellow cluster performs better on both of the exams than the other clusters. However, it is clear that there is a overlap, which makes the K-Means model not the best to use. Three clusters were used, since that provided the best visual. Another hyperparameter that was used was the n init, and the amount of times the centroid positions are changed is what the hyperparameter controls.

3.5. Results Overview

Model	Weighted Precision	Weighted Recall	Weighted F1-Score	Accuracy
Logistic Regression	66%	68%	66%	68%
Decision Tree	64%	66%	64%	66%

Figure 9. Results Table

To measure success, the main metric that was used was accuracy. Other metrics such as precision, recall, and f1-score were also used to determine success. The metrics' formula was put in the study by Esmael Ahmed, and that was used similarly in this project [1]. The data was split into training and testing data with a 80%-20% split. This allowed for the models to learn enough while also having enough data to get accurate results. The results from this project were quantitative because of metrics that were used and they can be seen in the table. The result of the project was a mix of succeeding and a failure because the models were not as accurate as I would have liked them to be. Also, the recall was low for students who got above a C was disappointing but an area to improve on. The logistic regression model performed better than the decision tree model as accuracy was 68% for the regression model and 66% for the decision tree model. The logistic regression model also had better numbers for the other metrics.

4. Availability

4.1. Code Usability

My code is available on a GitHub repository [2]. Open-source license was used to release my code, and to be specific, the MIT license was used. My method will be disseminated through GitHub, and the findings are also available on GitHub.

5. Reproducibility

5.1. Recreation Ability

Others can reproduce my results by downloading the code from GitHub and running it [2]. The dataset was from Kaggle, so it was publicly available and that was used for the training and testing data [5]. The model parameters are fully reproducible and a random seed of 42 was used during the splitting of the data and for the models.

6. Machine Learning Analysis

6.1. Problem Structure

The structure of my problem was to predict final students' grades. This was done by predicting whether they would be above or below the C letter grade. The structure of the models reflects the structure of the problem because logistic regression is able to make the binary classification. Decision Trees are able to split and make decisions based on the features to classify the students. In the logistic regression model, it had the coefficients for features which was a learned parameter. Next, the decision tree had the thresholds that it would use to make decisions on whether to split left or right, and the K-Means clustering had the centroids. However, the decision tree had the max depth hyperparameter which wasn't a learned parameter.

6.2. Model Preprocessing and Inputs

Some inputs for the model were the features such as lunch, parental level of education, and gender. The Standard Scaler was used to ensure that the features had an equal scale since that is a necessity for logistic regression. The output was a binary result on whether the student was above a score of 73. The data was preprocessed by making the gender feature a binary column. The parental level of education column and race/ethnicity features were changed to have a column for each type of value. Also, feature engineering was used to create the gender prep feature. The post processing for the model was predicting whether the student was above a C. The loss functions that were used were entropy for the decision tree and the log loss for the logistic regression model. With the fit function, it uses log loss. The models did not overfit because of the accuracy being at a solid number and with the decision tree, the max depth hyperparameter allowed for the chances of overfitting to be reduced.

6.3. Framework and Hyperparameters

The logistic regression model used the default hyperparameters such as the default solver, C value, and penalty. For the decision tree, however, the max depth was a hyperparameter that was used, and it was set to 4. This made it so overfitting was not an issue. The number of clusters and n init were two hyperparameters used for the K-Means model. The default optimizers were used. I used the scikit-learn machine learning framework because it was a simple and reliable library to use with logistic regression, K-Means, and decision trees. By using the built-in models from scikit-learn for the three algorithms used in this project, I was able to focus on visualizing the results. These starting points allowed me to save time, since the need to implement the algorithms from scratch was no longer necessary. I was able to adjust the hyperparameters more to get better results.

7. Conclusion

In this project, the goal was to predict students' grades using machine learning algorithms. I used models such as logistic regression, decision trees, and K-means clustering. From the results, the logistic regression model had a better accuracy than the decision tree by two percent. Future work for this dataset could utilize machine learning algorithms such as neural networks.

References

[1] Esmael Ahmed. Student performance prediction using machine learning algorithms. *Applied Computational Intelligence and Soft Computing*, 2024(1), 2024.

[2] Raunak Chitre. https://github.com/RaunakC24/ML-Project-CS4824.

[3] Mohsen M. Khoudier, E. Prediction of student performance using machine learning techniques. *5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 40:333–338, 2023.

[4] Durán-Domínguez A Rastrollo-Guerrero JL, Gómez-Pulido JA. Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, 10(3), 2020.

[5] Jakki Seshapanpu. Students performance in exams. https://www.kaggle.com/datasets/spscientist/students-performance-in-exams, 2018.

[6] Mustafa Yağci. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), 2022.