Student Performance, Analysis, Visualization and Prediction

Ed Cruz
ITCS 5156 – 081
Professor: Minwoo Lee
April 23, 2024

Abstract

This project analyzes and predicts student performance using socioeconomic factors. Data Analysis and Machine Learning algorithms were the primary techniques implemented to achieve this goal. Some of this effort's challenges and limitations include defining academic success and balancing model performance and interpretability. The end goal of this project is to provide insights on student performance to help students and faculty gather enough information and act promptly to reduce the dropout rate. This project employs KNN, Logistic Regression, Naïve Bayes, and Decision trees for analysis of the proposed goal.

1. Introduction

The primary objective of this project is to predict the academic student through the analysis of different socio-economic factors such as gender, ethnicity, parental level of education, nutrition, and engagement in test preparation courses. Academic performance is a key component in a student's path to joining the workforce, and predicting student outcomes could be a critical component for educational institutions as it allows them to identify students facing challenges at an early stage and plan strategies to improve student outcomes. Academic performance is a multidimensional concept that depends on multiples aspects and requires an integration of different techniques and methodologies for its prediction (Contreras-Bravo et al., 2016). Through data analysis and Machine Learning techniques, this project aims to gain a better understanding of how social factors influence student success in education.

1.1 Challenges

Predicting student performance can have multiple benefits for students and faculty staff. However, it poses some challenges and limitations. Even though academic performance is usually measured

using grades as the performance metric, limiting student academic success to a number reduces the problem and limits the possible solutions. Academic success encompasses the complex relationship of various factors that influence academic performance, including socioeconomic conditions and study habits.

1.2 Motivation

Predicting student performance can offer numerous benefits for students and faculty. By gaining a deeper understanding of student performance, educational systems can adapt their approach to individual needs, potentially improving overall student outcomes. Predicting student performance enable educators to refine teaching techniques and methodologies, taking into consideration socio-economic factors and implementing strategies that better fit the needs of different groups. Moreover, early intervention can reduce dropout rates and improve education access and quality.

1.3 Open Questions in the Domain

Before we proceed to the analysis, it is critical to consider two factors: First, interpretability and predictive performance. How can classification algorithms be further improved to enhance the model performance while maintaining a balance with interpretability? Another critical factor is feature selection. What are the most important factors that impact student performance? How can we effectively select relevant features to incorporate into the classifiers?

1.4 Overview of the Approach

The approach to achieving the proposed objective is composed of different stages. Initially, we have dataset selection, data preprocessing, and visualization. Subsequently, we have model selection. In this project, KNN, Logistic Regression, Naïve Bayes, and Decision tree algorithms were implemented. These models are trained on the data and evaluated using metrics such as accuracy, precision, and F1 score to ensure the quality of the classification (Contreras-Bravo, et al. 2022).

2. Background

2.1 Prediction of University-level Academic Performance through Machine Learning Mechanisms and Supervised Methods

This approach utilizes data analytics and Machine Learning techniques to predict student performance. In the last decade, different studies have established the variables that impact academic performance (Contreras-Bravo et al., 2022). The dataset used includes socio-demographic attributes. The paper focuses on identifying the most relevant variables in students' academic journey using methods such as wrapping, filters, and assembly techniques. This approach utilizes Decision Trees, KNN, SVC, and Naïve Bayes to classify and predict academic performance for each semester. The values showed an 80% and 78% accuracy result, and this analysis suggested that academic average may not necessarily determine student performance in subsequent semesters.

2.2 Performance Prediction for Higher Education Students Using Deep Learning

The paper highlights the importance of predicting student performance and introduces the concept of Educational Data Mining to obtain insights from education data, emphasizing the critical role it plays in improving educational systems (Li, et al., 2021). This research article proposes the implementation of deep learning techniques such as convolutional neural networks and long short-term memory to predict student performance.

2.3 Pros and Cons

It is crucial to consider the advantages and drawbacks of these approaches to understand the method selected for this project. Decision trees and KNN classifiers offer interpretability, enabling users to understand the decision-making process. Additionally, they are efficient, as these algorithms are fast and simple, making them suitable for handling large datasets with categorical features. However, it is critical to consider that decision trees may have overfitting issues, especially with complex datasets, which could lead to poor performance.

On the other side, CNNs can automatically learn hierarchical representations of data and capture complex patterns and structures. However, training these models requires a significant number of computational resources and can be time-consuming, especially with large datasets. CNNs

typically require large amounts of labeled data to learn meaningful representations effectively, and this may not always be available. Finally, CNNs are also prone to overfitting, especially with small datasets.

2.4 In Relation

Both approaches aim to predict student performance using machine learning techniques. In the first research article, the author focuses on identifying relevant features and then implements decision trees and KNN, among other classifiers, to achieve the main objective. Since these supervised methods require fewer computational resources, they have been chosen for this project as they allow easier interpretability and understanding of the decision-making process.

Similarly, the second approach aims to predict student performance through a different technique. Although the main method for this project is different from this approach, both share the goal of predicting student performance. Having two different approaches can allow us to compare the obtained results and not only improve model performance but also gain more insights and have a deeper understanding of student performance.

3. Method

Two different approaches for performance prediction were implemented. For the first approach, three classifiers were implemented to predict test preparation from socioeconomic factors. For the second approach a decision tree was implemented to predict the average score from different socioeconomic factors.

Logistic Regression is an effective tool to determine the probability of a binary outcome. In the context of education, predicting if a student is likely to take a test preparation course based on various factors could provide insights into how socioeconomic factors impact this student's decision.

KNN presents an effective method for predicting the class of a new instance by considering the classes of its nearest neighbors. Similarly to logistic regression, KNN allows the prediction of

student participation in test preparation courses based on features such as study habits, lunch type, or parental level of education.

Naive Bayes serves as another viable approach for predicting student participation in test preparation courses by leveraging factors such as gender or ethnicity. This technique assumes that features are conditionally independent given the class label.

Finally, the decision tree is a versatile tool capable of handling both numerical and categorical data, characterized by its simplicity of interpretation. The decision tree was trained to predict the average score for students. This prediction technique incorporated different demographic features such as gender, ethnicity, and parental level of education among others. The decision tree learns the patterns and relationships between these features and the target variable to make predictions on unseen data based on learned patterns.

3.1 Framework

3.1.1 Data preprocessing:

- Load the dataset.
- Encode categorical variables.
- Split the dataset into features (X) and target variable (Y)
- Split Data into Training and Testing Sets:Use train_test_split to split X and Y into X_train, X_test, y_train, y_test

3.1.2 Model Training:

- Initialize the Classifiers
- Fit the classifier to the training data (X_train, y_train)

3.1.3 Model Evaluation:

- Define an evaluation function (evaluation_report) to calculate accuracy.
- Call the evaluation function with the trained classifier and test data.
- Print the training and testing accuracy scores.

3.2 Data Preprocessing

The data preprocessing stage involved loading the dataset and adding two new attributes, "total score" and "average," to consolidate student performance metrics. Checking for null or missing values ensured data integrity while encoding categorical data facilitated numerical representation for machine learning algorithms. Additionally, the dataset was split into training and testing sets to evaluate the model performance.

4. Experiment

The experiment began by selecting the "StudentPerformance.csv" file, containing eight attributes: four categorical and four numerical values. Initially, data preprocessing was conducted to ensure there are no missing values. Additionally, two extra attributes, "total score" and "average," were added, calculated from the math, writing, and reading scores. Categorical variables were encoded using LabelEncoder for ease of use in the machine learning models.

Following this, the dataset was split into features and the target variable. This step involved dividing the data into training and testing sets using the train_test_split function with an 80-20 ratio, ensuring sufficient data for both training and evaluation. After preparing the data, the next phase involved selecting and training various classifiers: Logistic Regression, KNN, Naïve Bayes, and Decision trees.

The chosen models were trained on the training set and evaluated using the testing set. Performance metrics such as accuracy and F1 score were used to measure the models' effectiveness in predicting student performance. The results obtained indicated that Logistic Regression, KNN, and Naive Bayes exhibited similar accuracies, with train accuracies around 65% and test accuracies hovering around 60%. However, the Decision Tree model showed a striking difference, achieving perfect accuracy on the training data but slightly lower accuracy (97%) on the test data, suggesting potential overfitting. These results highlight the need for careful analysis and interpretation of model performance, especially when there is a significant gap between training and testing accuracies.

Furthermore, cross-validation was implemented to learn the generalization error of the models. This step provided insights into the models' performance and their ability to generalize to unseen data. Cross-validation results indicated consistent performance for Logistic Regression and Naïve Bayes, with slightly higher error for KNN. The experiment concluded with an analysis of the results, identifying areas for improvement and setting the ground for future steps.

5. Conclusion

This project provided valuable insights into the performance of various machine learning models in predicting student performance based on a given dataset. Through experimentation and analysis, we observed that while logistic regression, KNN, and Naive Bayes models demonstrated similar accuracies, the Decision Tree model exhibited signs of overfitting, emphasizing the importance of model selection and evaluation. Throughout the project, the project showed the importance of careful data preprocessing, model training, and evaluation in machine learning workflows. Challenges such as understanding model intricacies, interpreting results, and addressing potential overfitting were encountered. In terms of contributions, the project benefited from existing knowledge and methodologies in data preprocessing, model selection, and evaluation techniques. Specifically, techniques such as cross-validation and train-test splitting, as well as algorithms like Logistic Regression, KNN, and Naive Bayes, were drawn upon from existing machine learning literature. However, the unique contribution of this work lies in its application and analysis within the context of student performance prediction, providing valuable insights and informing future research directions in educational data mining.

References

Contreras-Bravo, L., Nieves-Pimiento, N., & Gonzales Guerrrero, K. (2022). Prediction of University-Level Academic Performance through Machine Learning Mechanisms and Supervised Methods. Ingenieria. Ing, vol 28, no. 1. https://doi.org/10.14483/23448393.19514

Li, S. & Liu, T. (2021). Performance Prediction for Higer Education Students Using Deep Learning. Hindawi Complexity, Vol 2021. https://doi.org/10.1155/2021/9958203

Student Performance Dataset. https://www.kaggle.com/datasets/spscientist/students-performance-inexams?resource=download

Supplementary materials: https://github.com/eacruzs18/ITCS-5156-Project-ML