Pooja Shinde

DeVos Graduate School, Northwood University

115537-MGT-665-NW

Solv Probs W/ Machine Learning Graduate Midland Summer 2024-2025

Dr.Itauma

15/06/2025

Student Performance Classification Using Machine Learning

Abstract

The goal of this study is to use supervised machine learning models with demographic and academic characteristics to predict which level students will perform at . Both categories of student information are taken into consideration by the dataset attributes which include gender, parental education, lunch, and multiple examinations or tests. Creating and testing three different classification methods - Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree - was achieved using different evaluation metrics, including accuracy, precision, recall, and F1-score. The Decision Tree method's model had the highest performance when predicting student performance levels, showcasing its potential for ensuring educational appraisal. Overall, the study presents nuances of discussing the utility of classification models to identify students approaching critical stages of risk, leading to data-driven educational support.

Introduction

Education systems can benefit tremendously from data analytics using predictive models of student performance in a timely manner. In this research, we are attempting to use a range of machine learning algorithms to predict student performance based on different attributes. The purpose of this research is to assesses the predictive capability of Logistic Regression, k-NN and Decision Tree to classify student results into performance-related categories. Predictive models such as this can provide valuable input into strategies educators can develop to assist students to achieve academic success through tailored interventions.

Related Work

Previous studies have used classification algorithms in educational datasets to a great extent. Cortez and Silva (2008), for instance, used Decision Trees and Neural Networks on data collected from responses from students in Portuguese secondary school to predict student grades. Kotsiantis et al. (2004) used Naive Bayes and k-NN classifiers to examine student performance prediction. While these studies represent just a fraction of the work on using machine learning in education, they demonstrate the importance of preprocessing and feature selection gained from their practical experience in ensuring good accuracy within their model.

Methodology

Dataset

The dataset under consideration contained 1000 students with features including their gender, race/ethnicity, parental education level, lunch type, and preparation for the test. The target variable was categorized based on the average scores of student performance.

Preprocessing Steps

• Utilized pandas to load the data set.

• Created an average_score column and a performance label:

o Poor (<40), Average (40–69), Excellent (≥70)

• Encoded categorical features using LabelEncoder.

• Split to training and testing set (80/20).

• Standardized features using StandardScaler.

Model Development

There were three classification models that were trained:

• Logistic Regression

• k-Nearest Neighbors (k=5)

• Decision Tree Classifier

Each model evaluated with:

• Accuracy

•Precision

• Recall

• F1-Score

All metrics were generated using the classification_report method from scikit-learn.

Results

Model Accuracy Precision Recall F1-Score

Logistic Regression = 0.86 0.86 0.86 0.86

k-NN (k=5) = 0.88 0.88 0.88 0.88

Decision Tree = 0.92 0.92 0.92 0.92

• The Decision Tree was stronger than the other models with measures of output performance on every metric.

• Confusion matrix and classification report evidence further support these findings.

• Visuals: Confusion matrix and performance comparisons by bar chart were included in the notebook.

Discussion

The Decision Tree classifier's better performance implies its ability to capture more complex patterns in a categorical education dataset. Alternatively, k-NN had similar performance, but is sensitive to scaling and distance metrics. Logistic Regression is interpretable, but a lower level of performance likely because of linear biases. Model evaluation reinforces the notion that model selection should be based on the dataset and education dataset.

Limitations

- The limited dataset size may impact how generalizable results are to a larger population.
- The dataset lacks consideration of socio-economic or psychological constructs that would factor into the performance in study habits on the platform.

Conclusion

This study showed that machine learning models, especially Decision Trees, are good at predicting student performance based on demographic and academic indicators. Future research could utilize larger datasets and additional features (e.g. behavioral data), and more sophisticated methods including ensemble models (e.g. Random Forest, XGBoost).

References

- Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *EUROSIS*.
- Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. (2004). Predicting students' performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, *18*(5), 411-426.
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Github Link: Student_/Student_Performance-Jupyter.pdf at main · Shinde2-creator/Student_