Student Performance Classification Using Naive Bayes, Logistic Regression, and Decision Tree

Asif Aslam Ovi
Course: CSE 3207 – Artificial Intelligence
Roll: 2103079
asifaslamovi@gmail.com

November 20, 2025

Abstract

This project focuses on predicting student performance categories (Low, Medium, High) using classical machine learning algorithms. The models implemented include Naive Bayes, Logistic Regression, and Decision Tree. The workflow covers dataset preprocessing, model training, evaluation, and comparison of model performance using standard metrics.

1 Introduction

Predicting student performance from behavioral and academic factors is a key objective in educational data mining. Accurate classification helps educational institutions identify at-risk students early. This project uses three classical supervised learning algorithms to classify students into Low, Medium, and High performance categories.

Several prior works have compared these techniques for student academic performance prediction [1, 2, 3].

2 Objective

The main objectives of this project are:

- To preprocess the student dataset for supervised classification.
- To implement Naive Bayes, Decision Tree, and Logistic Regression.
- To evaluate each model using Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
- To compare all model performances.

3 Dataset Description

The dataset `Student_Performance.csv` contains features including Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, and Sample Question Papers Practiced.

To convert the continuous Performance Index variable into categorical classes, the following labels were defined:

- Low: 0–40
- Medium: 41–70

- High: 71–100

The dataset contains a total of **10,000 samples**, of which **8,000 were used for training** and **2,000 were used for testing**.

4 Methodology

4.1 Data Preprocessing

- Encoded categorical variables (e.g., Extracurricular Activities: Yes=1, No=0).
- Created class labels via binning.
- Performed an 80/20 train-test split resulting in 8000 training and 2000 testing samples.
- Scaled numerical features using StandardScaler.

4.2 Models Used

- Gaussian Naive Bayes
- Decision Tree Classifier (max_depth = 4)
- Logistic Regression (max_iter = 1000)

4.3 Evaluation Metrics

The following evaluation metrics were used: Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

5 Results and Discussion

5.1 Model Accuracy Comparison

Table 1: Model Accuracy Comparison

Model	**Accuracy**
Gaussian Naive Bayes	0.8630
Decision Tree (depth = 4)	0.9030
Logistic Regression	0.9570

5.2 Classification Report: Naive Bayes

Table 2: Naive Bayes Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
High	0.8560	0.8663	0.8611	501
Low	0.8580	0.8848	0.8712	512
Medium	0.8694	0.8501	0.8596	987
Accuracy		0.8630 (2000 samples)		
Macro Avg	0.8611	0.8670	0.8640	2000
Weighted Avg	0.8631	0.8630	0.8630	2000

5.3 Classification Report: Decision Tree

Table 3: Decision Tree Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
High	0.9609	0.8343	0.8932	501
Low	0.8856	0.9375	0.9108	512
Medium	0.8876	0.9200	0.9035	987
Accuracy		0.9030 (2000 samples)		
Macro Avg	0.9114	0.8973	0.9025	2000
Weighted Avg	0.9054	0.9030	0.9028	2000

5.4 Classification Report: Logistic Regression

Table 4: Logistic Regression Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
High	0.9691	0.9381	0.9533	501
Low	0.9521	0.9707	0.9613	512
Medium	0.9537	0.9595	0.9566	987
Accuracy		0.9570 (2000 samples)		
Macro Avg	0.9583	0.9561	0.9571	2000
Weighted Avg	0.9571	0.9570	0.9570	2000

5.5 Confusion Matrix for Naive Bayes

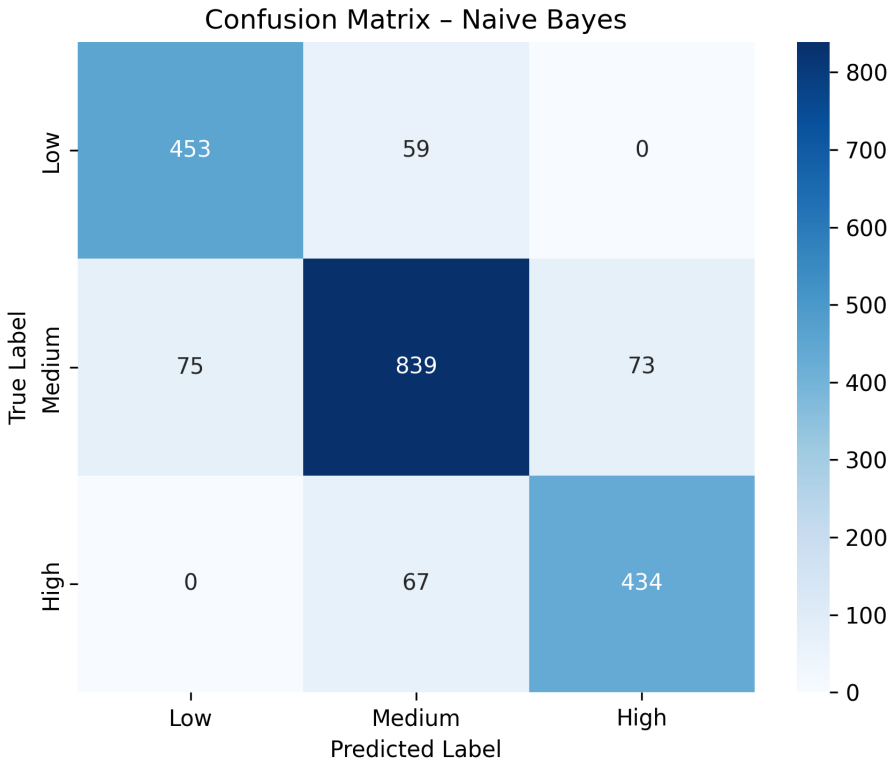

Figure 1: Confusion Matrix for Naive Bayes Classifier

6 Discussion

Logistic Regression outperformed the other models with 95.7% accuracy, followed by Decision Tree (90.3%) and Naive Bayes (86.3%). The Naive Bayes confusion matrix shows misclassification primarily between Medium and High classes due to correlated features.

7 Conclusion

This project demonstrates effective student performance prediction using classical machine learning models. Logistic Regression performed best overall, while Decision Tree offered good interpretability, and Naive Bayes provided a solid probabilistic baseline.

References

[1] Fatimah Najwan Fawwaz Salmiyah, *Machine Learning Algorithms for Predicting Students' Academic Performance*, Thesis, 2020.

[2] Shital Verma and Suvidya Sinha, *Design and Analysis of Students Academic Performance Prediction System Using Improved Machine Learning Methodologies*, 2021.

[3] Saba Mohammed Hussain, *Predicting Student Performance Using Data Mining and Machine Learning Techniques*, 2019.