Student Performance Analysis Report

Team Members 1: Vinayak Kumar Singh (23MCA1030)

Team Members 2: Neha Singh (23MCA1049)

Subject: Machine Learning project (PMCA507L)

Guided By: Dr. Saleena B

Abstract

This Project study aims to analyze student performance across various demographic and academic factors, with the goal of developing reliable machine learning models that can predict student outcomes and identify the key drivers of academic success.

By leveraging a diverse range of algorithms, from classical statistical methods to advanced ensemble and deep learning techniques, this project provides valuable insights to educators, policymakers, and school administrators on how to enhance student achievement and address the educational challenges faced by diverse student populations.

Problem Statement

The primary objective of this study is to analyze a comprehensive student performance dataset and construct predictive models that can reliably forecast academic outcomes. The dataset includes information about students' demographic characteristics, such as gender, race, parental education, and socioeconomic status, as well as their test scores in math, reading, and writing. The challenge lies in building accurate and interpretable models that can not only predict student performance but also provide insights into the factors that contribute to academic success, enabling targeted interventions and support.

Dataset

The dataset used in this project is the "StudentsPerformance.csv" file, which contains the following features:

Dataset URL

https://github.com/CodeVinayak/Student-Performance-Analysis-Machine-Learning/blob/main/StudentsPerformance.csv

- **Gender**: the gender of the student (female or male)
- **Race**: the race of the student (group A, B, C, D, E)
- **Parental Education:** the highest level of education achieved by the student's parents
- **Lunch**: whether the student received a standard or free/reduced lunch
- **Test Preparation Course**: whether the student completed a test preparation course
- **Math Score**: the student's math test score
- **Reading Score**: the student's reading test score
- **Writing Score**: the student's writing test score

Collab Code URL

https://colab.research.google.com/drive/1nCRQSJOYZ0d1_mSibGnYfW1q_MtWSv8q?usp=sharing

Tools Used

The following tools and libraries were used in this project:

Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn and CatBoost

Algorithms Used

The following machine learning algorithms were employed in this project:

Classification Algorithms

1. Logistic Regression
2. CatBoost Classifier
3. Decision Tree
4. K-Nearest Neighbors

Clustering Algorithms

5. K-Means Clustering
6. Hierarchical Clustering
7. Mean Shift
8. DBSCAN clustering

Ensemble Classifiers/Deep Learning Techniques

9.Random Forest

10. Artificial Neural Network (ANN)

Methodology

1. **Data Preprocessing**: The dataset was carefully inspected for missing values, and categorical features were encoded using label encoding. Feature selection was performed to identify the most relevant variables for the predictive models.
2. **Model Training and Evaluation**:
    - **Classification Algorithms:** Logistic Regression, CatBoost Classifier, Decision Tree, and K-Nearest Neighbors were used to predict the gender of students based on their academic performance.
    - **Clustering Algorithms:** K-Means, Hierarchical Clustering, Mean Shift, and DBSCAN were employed to identify distinct groups of students with similar performance patterns.
    - **Ensemble and Deep Learning Techniques:** Random Forest Classifier and Artificial Neural Network were implemented to forecast student performance across various subjects.
3. **Performance Evaluation**: The accuracy, precision, recall, and F1-score were calculated for each classification model, while the Silhouette Score and Davies-Bouldin Index were used to assess the quality of the clustering algorithms.
4. **Visualization and Interpretation**:
    - Confusion matrices were plotted to analyze the misclassification patterns of the models.
    - Feature importance plots were generated to identify the most influential factors in predicting student performance.
    - Cluster visualizations were created to understand the grouping of students based on their academic scores.

Inferences

1. The dataset has 1000 entries with 8 columns, including numerical features like math, reading, and writing scores, and categorical features like gender, race, parental education, etc.
2. Female students on average have higher scores in reading and writing, while male students have higher scores in math
3. K-Means clustering grouped the students into 3 clusters based on their math, reading, and writing scores.
4. The Random Forest Classifier, KNN Classifier, and Artificial Neural Network models achieved high accuracy (87%, 91%, and 90% respectively) in predicting student gender based on their academic performance.
5. The feature importance plot from the Random Forest Classifier showed that math score is the most important feature for predicting gender.

Novelty

1. The notebook demonstrates the application of a wide range of machine learning techniques, including regression, classification, and clustering algorithms, to analyze and model the student performance dataset.
2. It provides a comprehensive analysis of the dataset, including exploring gender differences in academic performance, applying various clustering algorithms, and evaluating the performance of different classification models.
3. The use of ensemble methods like Random Forest and deep learning techniques like Artificial Neural Networks adds to the novelty of the approach.
4. The visualization of the clustering results and feature importance plots provide valuable insights into the underlying patterns and relationships in the dataset.
5. The code is well-structured and documented, making it easy for others to understand and build upon.

Results

The key findings and performance metrics of the various models are as follows:

Classification Algorithms	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression	89.5	92.71	86.41	89.45
CatBoost	88.5	91.67	85.44	88.44
Decision Tree	84.0	85.71	80.41	82.98
K-Nearest Neighbors	91.0	88.35	93.81	91.00

Clustering Algorithms	**Silhouette Score**
K-Means	40.54
Hierarchical	35.24
Mean Shift	47.71
DBSCAN	54.07

Ensemble Classifiers/ Deep Learning	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Random Forest	87.0	91.40	82.52	86.73
Artificial Neural Network	90.0	93.68	86.41	89.90

Conclusion

This project has successfully demonstrated the application of various machine learning algorithms to analyze and predict student performance. The results suggest that the Artificial Neural Network and Logistic Regression models are the most effective in accurately classifying students' gender based on their academic scores.

Among the classification algorithms, the K-Nearest Neighbors (KNN) Classifier stood out as the best-performing model, achieving an accuracy of 91.0%, precision of 88.35%, recall of 93.81%, and an F1-Score of 91.00%. The KNN Classifier's strong performance indicates that it is the most effective in predicting the gender of students based on their math, reading, and writing scores.

The findings of this study can be valuable for educators, policymakers, and school administrators in developing strategies to enhance student success and address the educational challenges faced by diverse student populations. By leveraging the insights gained from this analysis, stakeholders can implement targeted interventions, allocate resources more effectively, and create personalized learning experiences to unlock the full academic potential of students.

Moving forward, it is recommended to further explore the application of advanced techniques, such as feature engineering and hyperparameter optimization, to improve the performance of the models. Additionally, incorporating other relevant data sources, such as student attendance, extracurricular activities, and socioeconomic factors, may provide a more comprehensive understanding of the drivers of academic success.