Student Performance in Exams Using Machine Learning

Title and Authors

The project is titled "Student Performance in Exams Using Machine Learning”. It is an independent project conducted by Tejashwini P, a Master's student at the University of North Carolina at Charlotte.

Introduction

The project explores how machine learning techniques can address the challenges educational institutions face in evaluating and improving student performance. A common issue in education systems is the inconsistency in evaluation structures. Factors like economic, social, psychological, and cultural influences further complicate measuring students' abilities accurately. Additionally, subjective evaluation processes, such as biased grading, inappropriate assessment methods, and dishonest invigilation, often lead to inaccurate performance metrics.

This issue is critical not only for students but also for institutions and governments. For instance, in systems where education is publicly funded, such as Iraq or subsidized through loans like in the United States, student failures impose a significant financial burden. Predicting student performance using advanced computational methods can help identify at-risk students early, enabling institutions to intervene effectively and reduce such expenses.

This research focuses on creating predictive models using machine learning algorithms to determine student outcomes based on specific input factors. The primary goal is to provide institutions with actionable insights to support students who may struggle academically, thereby ensuring better resource allocation and retention rates.

Motivation

The motivation for this project stems from the need to use data science effectively to understand and predict educational outcomes. Machine learning algorithms are particularly suited to uncovering patterns in data that may be invisible to the human eye. This study introduces novel attributes such as internet usage as a learning tool and time spent on social media, alongside traditional academic indicators, to better analyze their collective impact on student performance.

To build an accurate predictive model, the project employs four machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Logistic Regression. These models are tested for accuracy, and the best-performing algorithm is selected to form the foundation of the solution.

Open Questions in the Domain

The field of predicting student performance raises several important questions:

1. Can machine learning algorithms provide reliable and consistent results when applied to diverse datasets?
2. What role do social and behavioral factors, such as internet usage or time spent on social media, play in influencing academic outcomes?
3. How can predictive models be improved to provide not only higher accuracy but also actionable insights for educational institutions?

This study attempts to address these questions through a combination of data preprocessing, feature engineering, and algorithm selection, exploring the practical feasibility of using machine learning to predict student outcomes.

Approach Overview

The project applies a systematic methodology that begins with data preprocessing and visualization, followed by feature selection and algorithm testing. A dataset containing 1,000 student records from Kaggle was analyzed, with features such as gender, race, parental education level, lunch status, and test preparation courses. The data was visualized using Python libraries like Seaborn to identify patterns and correlations.

Multiple machine learning algorithms were tested, and their performances were compared based on accuracy metrics. SVM emerged as the most effective algorithm, achieving 93% accuracy. To make the system user-friendly, a web-based interface was developed using Flask, allowing users to input details and obtain predictions in real time.

Background

Extensive research has been conducted in the domain of educational data mining to predict academic outcomes. Predictive models typically analyze various factors such as demographics, grades, and attendance to identify at-risk students. For example, classification algorithms like Decision Tree, Naive Bayes, and KNN have been widely used to develop these models.

While these studies have contributed valuable insights, they often have limitations in feature selection, dataset size, or algorithmic efficiency. This project builds on these foundations, introducing SVM as a robust algorithm to improve prediction accuracy.

One major advantage of machine learning in this domain is its ability to handle large datasets and identify hidden patterns. However, challenges like overfitting, limited data availability, and the need for extensive feature engineering remain significant hurdles.

By leveraging previous research, this project implements an improved methodology that compares the performance of multiple algorithms and validates their results against a real-world dataset.

Methods

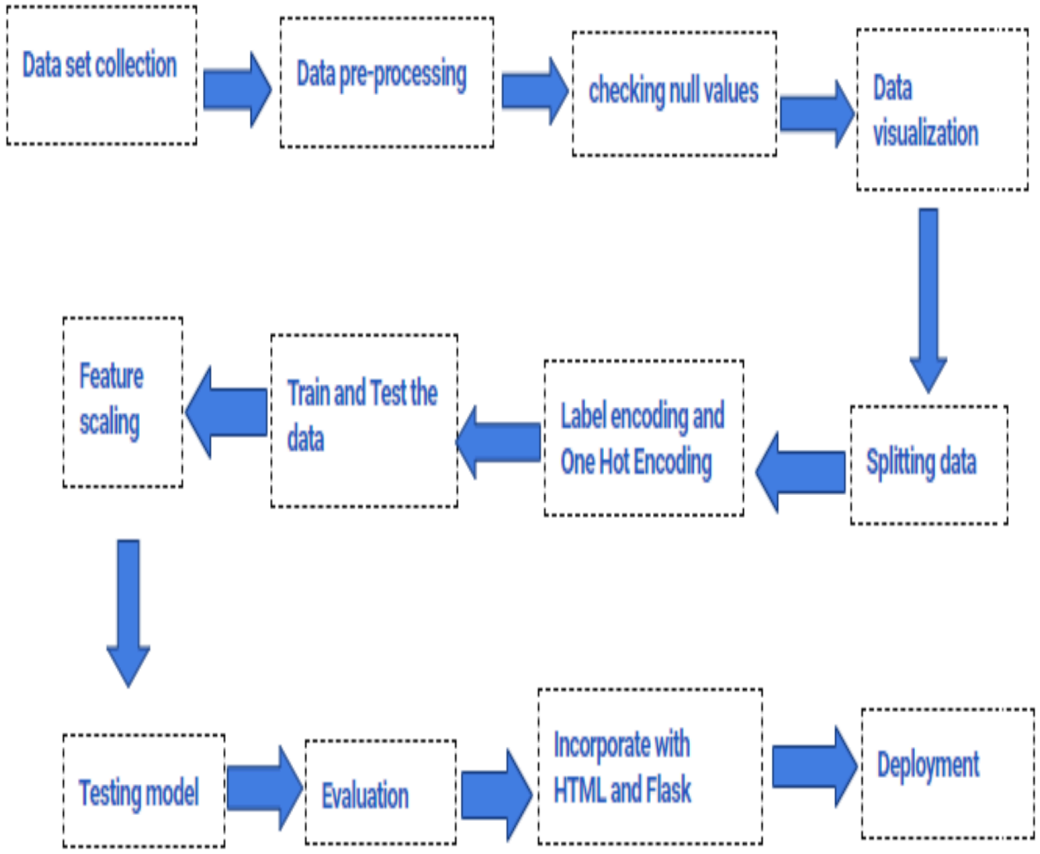
The methodology employed in this project combines comprehensive data preprocessing, algorithm testing, and system deployment to ensure a reliable and user-friendly solution.

Algorithms and Techniques Used:
The project tested six machine learning algorithms: Support Vector Machine (SVM), Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and KNN. Among these, SVM provided the highest accuracy of 93%, making it the primary choice for the final model.

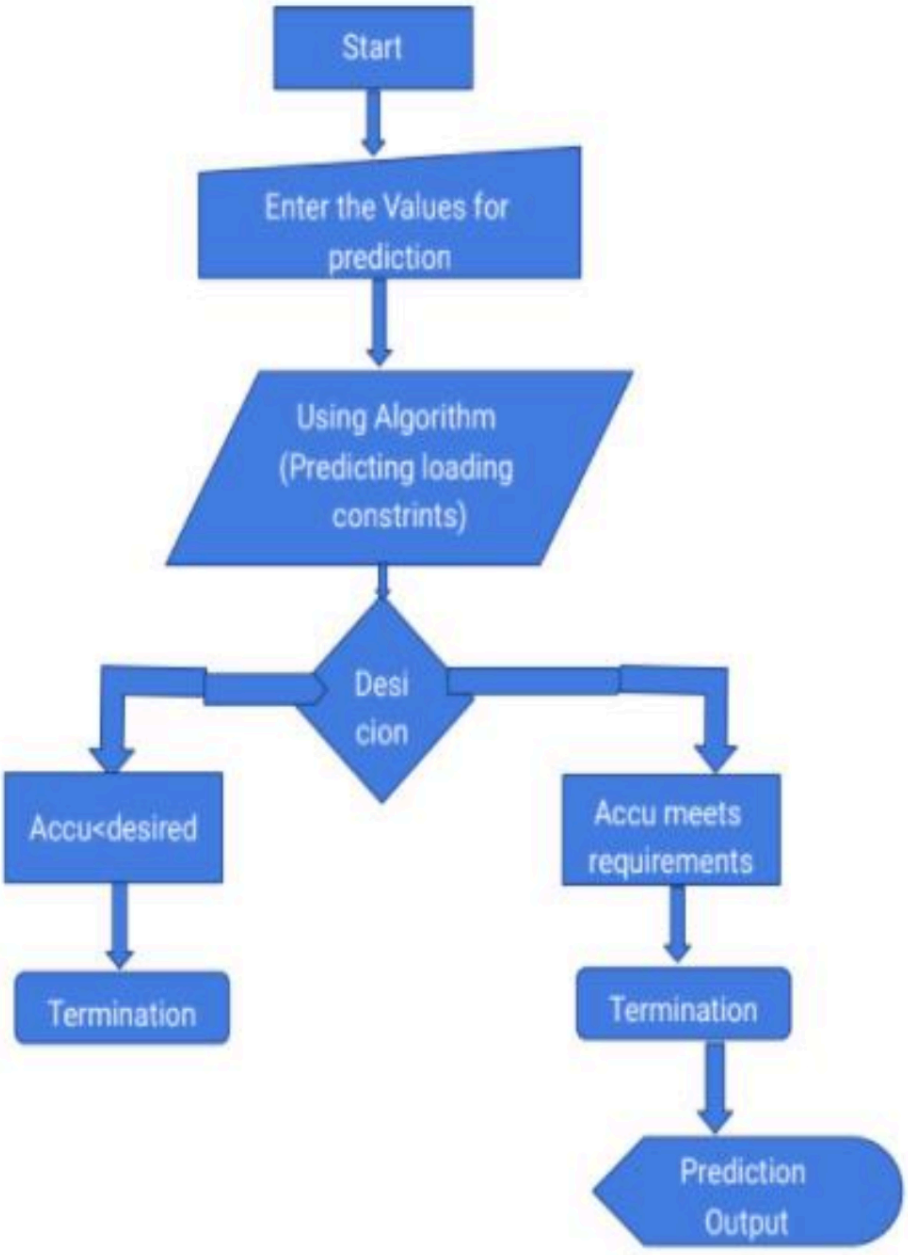
Data Preprocessing:
The dataset was preprocessed to handle missing values, eliminate redundant features, and encode categorical data into numerical formats. Techniques like label encoding and one-hot encoding were used to ensure compatibility with machine learning models. Feature scaling was applied to address data outliers, ensuring that all features contributed equally to the model's performance.

Deployment Framework:
The predictive model was deployed using Flask, a lightweight web framework in Python. Flask facilitated the creation of a user-friendly interface where users could input parameters and receive predictions in real time. HTML templates were designed to ensure an intuitive and visually appealing interface.

Block Diagram :



Flow Diagram :



Experiments

The experimental phase involved testing various machine learning algorithms to evaluate their performance. The dataset consisted of 1,000 records, which were split into training and testing sets. Performance metrics like accuracy were used to compare algorithms.

Results:

- SVM outperformed all other algorithms with a 93% accuracy rate.
- Logistic Regression followed with 90% accuracy, while Decision Tree achieved 89%.
- Random Forest, Naive Bayes, and KNN showed comparatively lower accuracies, at 84%, 87%, and 74%, respectively.

Analysis:
The results highlight the robustness of SVM in handling classification problems, particularly in datasets with complex relationships. The accuracy of the SVM model indicates its strong generalization capabilities, making it suitable for predicting student performance in varied contexts.

The research also involved creating a web-based prediction system. The user interface allows users to input key attributes and obtain a predicted grade based on the trained model. This real-time system demonstrates the practical applicability of the research.

Results and Analysis

This research aimed to evaluate the effectiveness of machine learning algorithms in predicting student performance based on academic and socio-demographic features. Six algorithms—Support Vector Machine (SVM), Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN)—were tested to determine their predictive accuracy. The evaluation metrics included accuracy, precision, and recall, with SVM emerging as the best-performing model.

Algorithm Accuracy

The experiment compared the models using a dataset of 1,000 student records. Below are the accuracy results:

Algorithm	**Accuracy (%)**
Support Vector Machine (SVM)	93
Logistic Regression	90
Decision Tree	89
Naive Bayes	87

Random Forest	84
K-Nearest Neighbors (KNN)	74

SVM outperformed all other algorithms, achieving an accuracy of 93%. This can be attributed to its ability to find the optimal hyperplane for classification, even in high-dimensional spaces. Logistic Regression and Decision Tree followed closely, with accuracies of 90% and 89%, respectively. KNN showed the lowest performance at 74%, likely due to its sensitivity to feature scaling and noise.

Performance Insights

The SVM model demonstrated robust generalization capabilities, with a reduced risk of overfitting compared to other algorithms. Its kernel trick enabled the handling of complex relationships within the data. Conversely, KNN struggled with high-dimensional data, which impacted its overall performance.

The models were evaluated using confusion matrices to assess true positives, false positives, true negatives, and false negatives. SVM showed the best balance between precision and recall, making it the most reliable choice for predicting student grades.

Web-Based Prediction System

A Flask-based user interface was developed to allow users to predict student grades based on inputs such as parental education level, lunch type, and test preparation course completion. The UI provides real-time predictions, displaying the student's grade category upon submission. Screenshots of the interface demonstrate its functionality and user-friendly design.

Challenges and Limitations

- **Dataset Size:** The dataset contained only 1,000 records, which limited the generalization of the models. A larger dataset could improve model training and validation.
- **Feature Selection:** The dataset lacked certain attributes, such as attendance and extracurricular activities, which could provide deeper insights into student performance.
- **Algorithm Comparison:** While SVM performed well, exploring advanced algorithms like Neural Networks or Gradient Boosting may further enhance accuracy.

Conclusions and Summary

The success of this project demonstrates the potential of machine learning in the field of education. By applying advanced algorithms and effective preprocessing techniques, the study achieved a high degree of accuracy in predicting student performance. The findings suggest

that SVM is particularly well-suited for this task due to its ability to classify data efficiently and minimize overfitting.

Future Enhancements:
There is significant scope for future work in this domain. The current study is limited by the dataset size and features. Future projects could incorporate additional attributes such as extracurricular involvement, attendance records, and behavioral data to improve prediction accuracy. Additionally, exploring advanced algorithms like Neural Networks or Gradient Boosting could further enhance performance.

Github Link

Github - ML Project

Anonymous Sharing Agreement

Yes, I agree to share my work as an example for future semesters.

References

[1] P.Veeramuthu Dr.R.Periasamy Application of Higher Education System for Predicting Student Using Data mining Techniques International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 5 (June 2015)

[2] Umesh Kumar Pandey , S. Pal A Data Mining view on Class Room Teaching Language IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011 ISSN (Online): 1694-0814

[3] Mrs. M.S. Mythili , Dr. A.R.Mohamed Shanavas An Analysis of students performance using classification algorithms IOSR Journal of Computer Engineering (IOSR-JCE) eISSN: 2278-0661, p- ISSN: 2278-8727 Volume 16, Issue 1, Ver. III (Jan. 2014), PP 63-69

[4] G.Paul Suthan and Lt.Dr. Santhosh Baboo Hybrid CHAID a key for MUSTAS Framework in Educational Data Mining IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011 ISSN (Online): 1694-0814

[5] S. T. Hijazi, and R. S. M. M. Naqvi, Factors Affecting Students Performance: A Case of Private Colleges, Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.