Dr.B.R.AMBEDKAR UNIVERSITY, ETCHERLA

COLLEGE OF ENGINEERING

Early Predicting of Students Performance in Higher Education

A Mini project report submitted in partial fulfilment of the requirement for the award of degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

2012 - 2025

Presented by

CH.Navya Bharathi-2181951011
G.Ravi Teja-2181951017
G.Harish-2181951016

UNDER THE GUIDANCE OF

R . Sridhar M.TECH(Ph.D)
Assistant Professor

ABSTRACT

Students' learning performance is one of the core components for assessing any educational system, offering insights into the effectiveness of learning processes. Understanding students' performance is crucial in addressing issues related to the learning process and is a key factor in measuring learning outcomes. This project applies machine learning techniques to predict students' performance in higher education, examining the influence of various academic and social factors. By utilizing data mining methods, this project aims to improve educational outcomes and provide early interventions for students at risk.

INTRODUCTION :

Education plays a significant role in society, and technological advancements have revolutionized research fields, including education. Higher education institutions consider students' academic performance a critical aspect of quality education. This project aims to predict students' performance using academic and social data, providing early interventions and enhancing educational outcomes. By employing machine learning techniques, this project seeks to identify the impact of various factors on students' GPA at an early stage. The integration of technology and education can offer valuable insights and improve overall student success.

OBJECTIVES

Framework Proposal : Develop a comprehensive framework for predicting students' performance using their academic records and social relationships.

Utilizing Admission Scores: Utilize admission scores, first-level course scores, academic achievement tests (AAT), and general aptitude tests (GAT) to predict student performance.

Dimensionality Reduction: Explore the use of t-SNE dimensionality reduction to analyze the relationship between various factors (such as admission scores, first-level course scores, AAT, and GAT) and GPA.

Assessment Metrics: Use a variety of assessment metrics (such as precision, recall, F1 score, and accuracy) to evaluate the quality and performance of the machine learning models.

Educational Insights:Provide insights and recommendations to educational institutions for mitigating the risks of student failures at early stages.

SYSTEM DESIGN:

1. Hardware Components:

Computer/Server: Processes the dataset, runs the recognition algorithms, and manages the database.

2.Software Components:

Operating System: Windows 10, macOS, or Linux (Ubuntu recommended)
Development Environment: Jupyter Notebook, PyCharm, or VS Code
Database Management: MySQL, PostgreSQL, or MongoDB for storing and managing student data

3. Libraries and Frameworks:

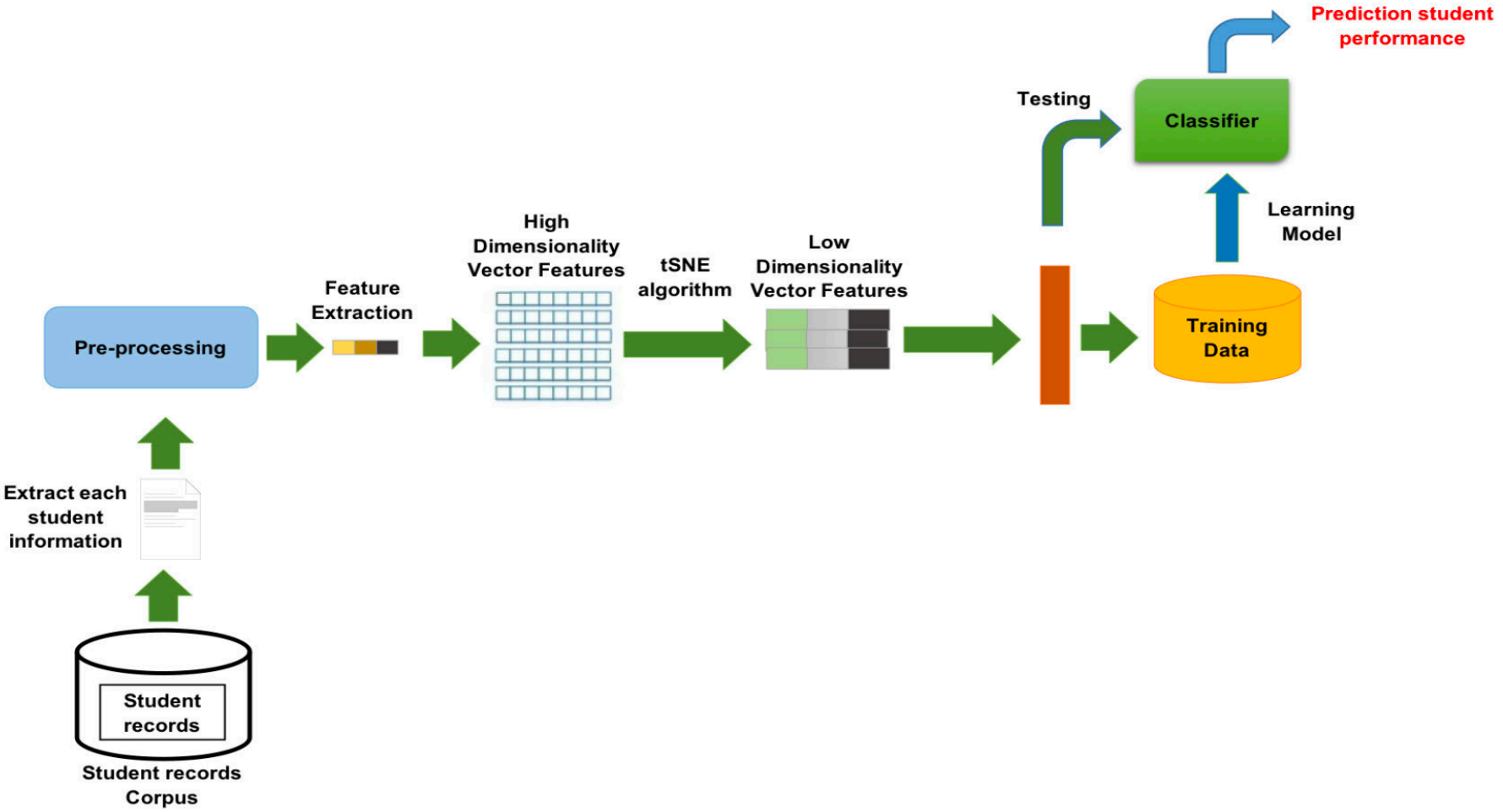
Pandas: For data manipulation and analysis
Numpy: For numerical computations
Matplotlib: For creating static, animated, and interactive
Scikit-learn: For implementing machine learning algorithms and tools

ANALYSIS :

A systematic approach to predicting student performance using machine learning techniques. The project begins by generating a synthetic dataset with various student attributes such as GPA, grades, and test scores.

Pre-processing involves data cleaning to handle missing values, encoding categorical features and scaling numerical features using StandardScaler. **Dimensionality reduction** is performed using the t-SNE algorithm to explore relationships between factors like admission scores, first-level course scores, academic achievement tests (AAT), and general aptitude tests (GAT). **Machine learning models**, including **Random Forest, SVM, and Naive Bayes**, are employed for prediction. These models are trained and evaluated using metrics like precision, recall, F1 score, and accuracy. Results with and without preprocessing techniques are compared to highlight their impact on classification accuracy, providing valuable insights for early intervention and support for students at risk.

BLOCK DIAGRAM:



ALGORITHMS

Random Forest : Random Forest is an ensemble learning method for both classification and regression tasks. It constructs multiple decision trees and merges their outputs to improve prediction accuracy and stability. This approach reduces the risk of overfitting.

Support Vector Machine (SVM) : Support Vector Machine (SVM) is used for classification and regression analysis. It finds an optimal hyperplane to separate data into different classes, making it effective in high-dimensional spaces and with clear margin separations.

Naive Bayes : Naive Bayes is a probabilistic classifier based on Bayes' theorem. It assumes independence between predictors, which simplifies computation and is particularly effective for large datasets, such as in text classification tasks.

Logistic Regression: A statistical model used for binary classification tasks, predicting the probability of a binary outcome based on one or more predictor variables.

IMPLEMENTATION

Model Selection:
- Random Forest
- SVM
- Naive Bayes

Preprocessing Pipeline:
- Automate preprocessing and modeling steps

Training Process:
- Split dataset into training and testing sets
- Train models on the training data

Model Evaluation:
- Evaluate using Precision, Recall, F1 Score, Accuracy

Model Comparison:
- Compare performance metrics
- Identify best-performing model

ADVANTAGES & APPLICATIONS:

- **Enhanced Academic Support:**By identifying at-risk students early, educational institutions can provide targeted academic support and resources, helping students improve their performance

- .**Data-Driven Decision Making:**The project leverages data mining and machine learning techniques to make informed decisions based on data analysis, improving the accuracy and effectiveness of interventions.

- **Improved Student Outcomes:**Predicting student performance allows for timely interventions, which can lead to better academic outcomes, reduced dropout rates, and higher overall student success.

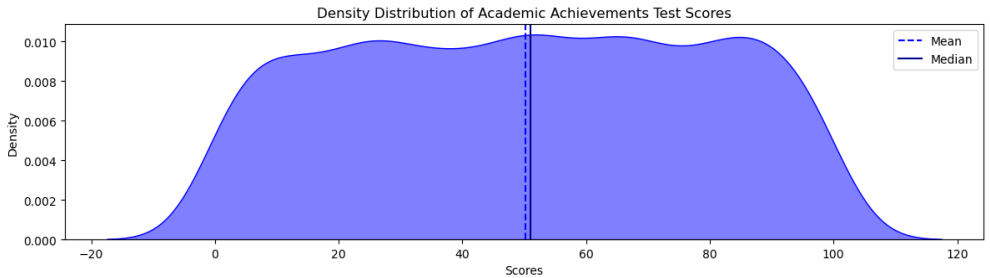
- **Resource Optimization:**Early predictions enable institutions to allocate resources more efficiently, focusing on areas where students need the most support and enhancing the overall learning experience.

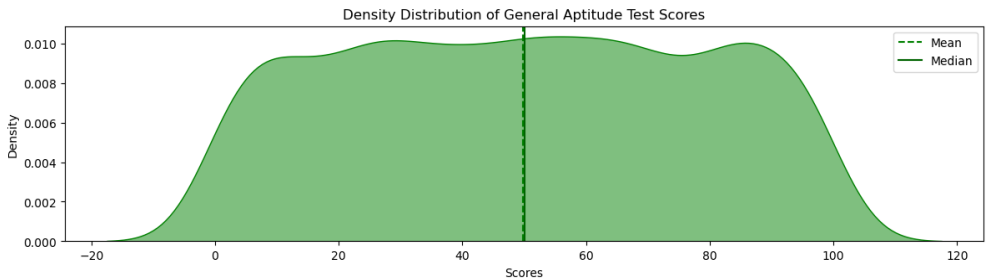
- **Insightful Educational Strategies:**The project provides valuable insights into the factors influencing student performance, allowing educators to develop more effective teaching strategies and curricula tailored to students' needs.

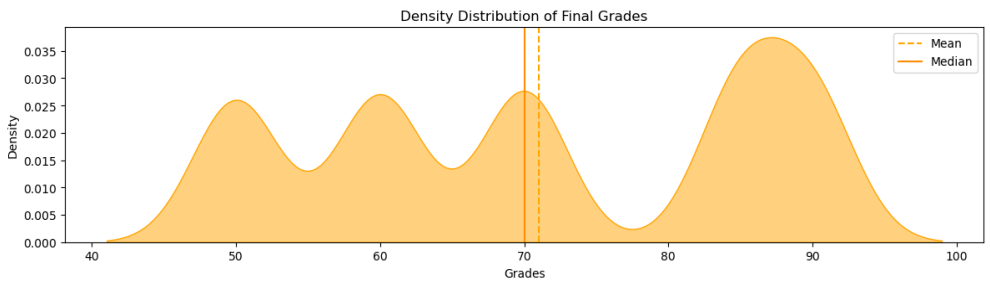
RESULT:

The results indicate that the Random Forest model outperformed SVM and Naive Bayes, demonstrating the highest accuracy and reliability. Random Forest's ensemble approach, which combines the predictions of multiple decision trees, contributes to its superior performance. Key insights from the analysis include trends in student performance, gender comparisons, and failure rates in courses. Visualizations such as density distributions, gender comparisons, box plots, and network graphs provide a comprehensive understanding of the data. These insights highlight the effectiveness of machine learning in predicting student performance and suggest that Random Forest is the most suitable model for this task.

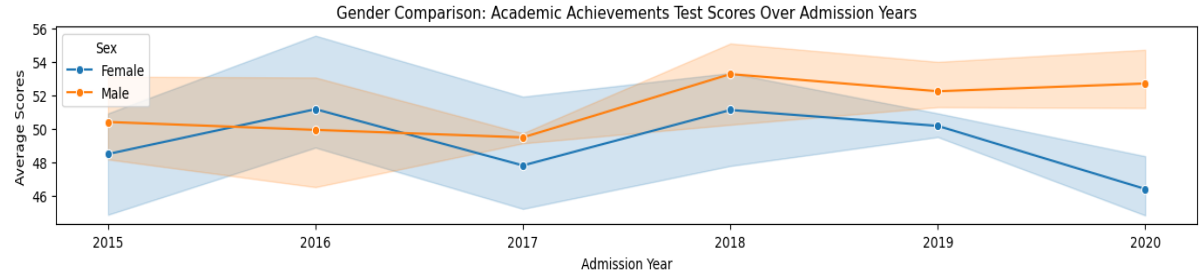
Density Distribution :

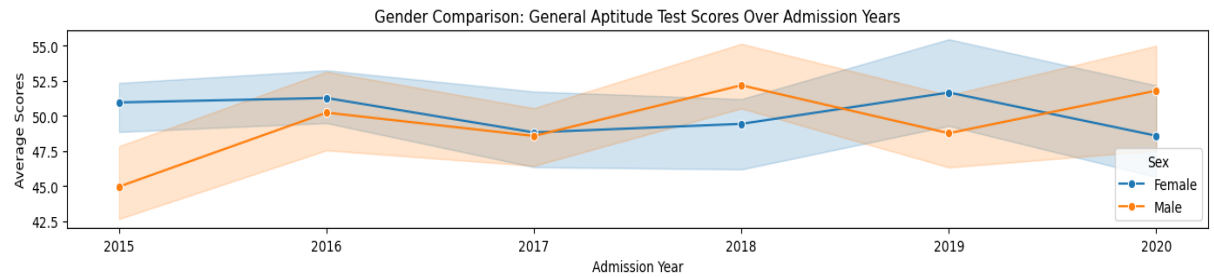


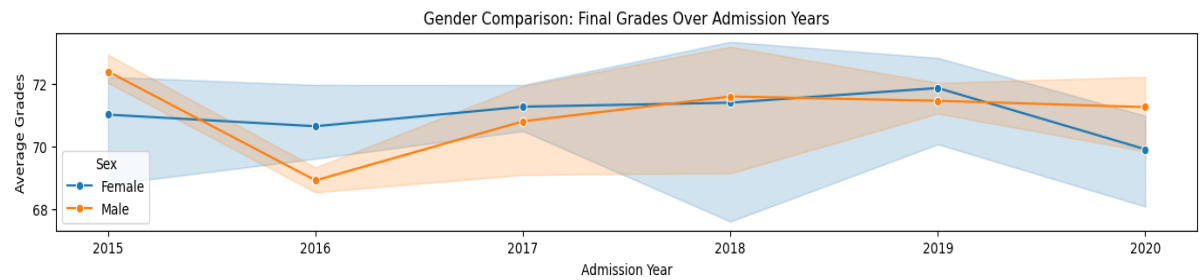




Gender Comparison :

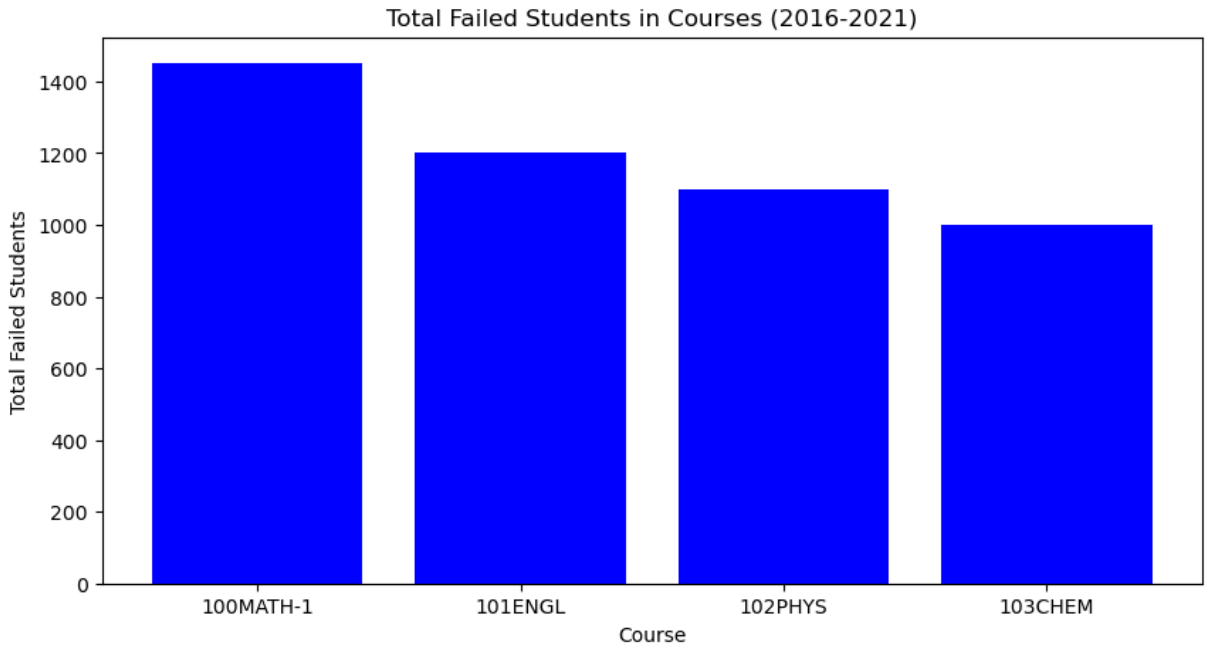






Model Performance Metrics:

Total Failed Students Analysis :



CONCLUSION:

The project successfully demonstrates the use of machine learning techniques in predicting student performance in higher education. The Random Forest model proved to be the most effective, providing accurate predictions and enabling early interventions. The analysis also revealed trends and patterns in student performance, offering valuable insights for educational institutions. Future work includes exploring additional features and advanced models to further improve predictive accuracy and generalizability.

REFERENCES:

[1] F. Giannakas, C. Troussas, I. Voyiatzis, and C. Sgouropoulou, ‘‘A deep learning classification framework for early prediction of team-based academic performance,’’ Appl. Soft Comput., vol. 106, Jul. 2021,
Art. no. 107355.
[2] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, ‘‘Student academic performance prediction model using decision tree and fuzzy genetic algorithm,’’ Proc. Technol., vol. 25, pp. 326–332, Jan. 2016.
[3] B. K. Francis and S. S. Babu, ‘‘Predicting academic performance of students using a hybrid data mining approach,’’ J. Med. Syst., vol. 43, no. 6, pp. 1–15, Jun. 2019.
[4] M. Yağcı, ‘‘Educational data mining: Prediction of students’ academicperformance using machine learning algorithms,’’ Smart Learn. Environ., vol. 9, no. 1, pp. 1–19, Dec. 2022.
[5] T. Le Quy, T. H. Nguyen, G. Friege, and E. Ntoutsi, ‘‘Evaluation of group fairness measures in Student performance prediction problems,’’ 2022,arXiv:2208.10625.
[6] X. Liu and L. Niu, ‘‘A student performance predication approach based on multi-agent system and deep learning,’’ in Proc. IEEE Int. Conf. Eng., Technol. Educ. (TALE), Dec. 2021, pp. 681–688.
[7] C. Romero and S. Ventura, ‘‘Educational data mining: A survey from 1995to 2005,’’ Expert Syst. Appl., vol. 33, pp. 135–146, Jul. 2007.
[8] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, ‘‘Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS,’’ IEEE Trans. Learn. Technol., vol. 10, no. 1, pp. 17–29,Jan./ Mar. 2017.

THANK YOU