The Great Escape: Stopping Student Disappearances in their Tracks

A MINI PROJECT REPORT

Submitted by

NIKILESH S (221801035)
SANTHOSH V(221801047)

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

NOVEMBER 2024

ANNA UNIVERSITY CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that this Report titled **"THE GREAT ESCAPE : STOPPING STUDENT DISAPPEARANCES IN THEIR TRACKS"** is the Bonafide work of **NIKILESH S (2116221801035), SANTHOSH V (2116221801047)** who carried out the work under my supervision.

SIGNATURE

Dr.J.M.GNANASEKAR,M.E.,Ph.D.,
Professor and Head,
Department of Artificial Intelligence and Data Science,
Rajalakshmi Engineering College,
Chennai – 602 105.

SIGNATURE

Dr.J .Manoranjini M.E.,(Ph.D).,
Associate Professor,
Department of Artificial Intelligence and Data Science,
Rajalakshmi Engineering College,
Chennai – 602 105.

Submitted for the project viva-voce examination held on....23·11·2024............

23/11/24
INTERNAL EXAMINER

EXTERNAL EXAMINER 23/11/24.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.**, for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.**, our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. J.M. GNANASEKAR., M.E., Ph.D.,** Professor and Head of the Department of Artificial Intelligence and Data Science for his guidance and encouragement throughout the project work. We are glad to express our sincere thanks and regards to our supervisor **Dr.J.MANORANJINI., BE., ME., PhD.,** Associate Professor, Department of Artificial Intelligence and Data Science and coordinator, **Dr. P. INDIRA PRIYA., M.E., Ph.D.,** Professor, Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College for their valuable guidance throughout the course of the project.

Finally, we express our thanks for all teaching, non-teaching, faculty and our parents for helping us with the necessary guidance during the time of our project.

ABSTRACT

This project aims to develop a predictive model to identify students at risk of dropping out using logistic regression and random forest algorithms. Logistic regression, a statistical method, will be used for its interpretability and effectiveness in binary classification problems. Random forest, an ensemble learning technique, will be employed for its ability to handle large datasets with higher accuracy and its robustness to overfitting. By analyzing academic performance, engagement metrics, and socio-demographic factors, the model will provide early warnings to enable timely interventions and support, ultimately reducing dropout rates.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 GENERAL

The project, titled "The Great Escape: Stopping Student Disappearances in Their Tracks," aims to address the growing concern of student dropouts in educational institutions. High dropout rates not only impact students' career prospects and personal lives but also lead to broader societal issues, such as increased unemployment and decreased community engagement. By predicting which students are at risk of dropping out, institutions can take proactive steps to support these individuals, thereby increasing graduation rates and fostering a supportive learning environment. This project’s goal is to leverage data-driven insights to enhance retention strategies and allocate resources more effectively.

To achieve this, the project utilizes machine learning techniques to analyze various factors contributing to student success and engagement. By examining academic performance, socio-demographic details, and engagement metrics, the model identifies students who might be at higher risk of dropping out. This predictive model integrates logistic regression for interpretability and random forest for handling complex, nonlinear relationships within the data. By combining these methods, the project ensures a balance between accuracy and understanding, making it easier for educators and administrators to comprehend the underlying factors that contribute to each student's risk profile.

The project is designed to be a practical, scalable solution that integrates seamlessly with existing educational systems. The final model is deployed through an alert system that notifies educators about students at risk, allowing for timely intervention. This predictive tool not only supports individualized assistance for students but also provides institutions with a means to make data-informed decisions about resource allocation and student support programs. Through this approach, educational institutions can enhance student success, strengthen academic engagement, and contribute to a more positive societal outcome by reducing dropout rates.

1.2 NEED FOR THE STUDY

The need for this study arises from the pressing issue of student dropouts, which affects educational institutions worldwide. High dropout rates hinder students from reaching their full academic and professional potential, leading to significant personal and social consequences. For many students, leaving school prematurely results in limited career opportunities, lower lifetime earnings, and reduced job security. This, in turn, can contribute to broader societal challenges such as economic instability and an increased likelihood of social isolation or engagement in low-skill occupations. Educational institutions face challenges in identifying at-risk students early enough to intervene, which is why developing a robust predictive model is essential.

In recent years, the complexity of factors contributing to dropout rates has increased, making it difficult for traditional methods to keep up. A student’s academic performance alone may not be a clear indicator of their likelihood to graduate, as socio-demographic factors and engagement levels play critical roles as well. With the rise of data analytics and machine learning, there is an opportunity to better understand and address this multifaceted issue. By using advanced models that take into account not only grades but also socio-demographic factors and engagement metrics, this study aims to provide a clearer picture of dropout risk. This approach allows educational institutions to go beyond academic monitoring, capturing a holistic view of the factors affecting a student’s educational journey.

Moreover, there is a growing demand for data-driven solutions in education that can improve retention rates and resource allocation. Institutions face constraints on time and resources, which means that support efforts must be carefully targeted to achieve maximum impact. A predictive model that identifies students in need of support allows institutions to act more strategically, focusing interventions on the students who need them most. This study fulfills a critical gap by offering a practical, scalable approach that educational institutions can implement to preemptively address student dropout, thereby fostering a more supportive and effective learning environment.

1.3 OBJECTIVES OF THE STUDY

The objective of this study is to develop a predictive model that accurately identifies students at risk of dropping out, allowing educational institutions to intervene early and effectively. By analyzing a combination of academic performance, engagement metrics, and socio-demographic factors, the study aims to provide a comprehensive understanding of the key indicators associated with dropout risk. This predictive tool will enable institutions to support at-risk students proactively, ultimately improving retention rates and academic success.

Furthermore, the study seeks to create an interpretable and scalable solution that can be integrated into existing educational systems. The model's interpretability is crucial, as it allows educators and administrators to understand the specific factors influencing each student's risk of dropping out. In addition to enhancing individual student outcomes, this study aims to support broader institutional goals by optimizing resource utilization and improving overall educational quality. By facilitating timely interventions for at-risk students, the model not only promotes better academic achievements and graduation rates but also contributes to a positive social impact, reducing the potential for unemployment and societal disengagement.

Here are three key points that encapsulate the objectives of this study:

1. Identify At-Risk Students: Develop an accurate predictive model that identifies students who are at a higher risk of dropping out, enabling timely and targeted intervention.

2. Enhance Resource Allocation: Provide an interpretable tool that helps educational institutions make data-driven decisions, ensuring resources are allocated effectively to students in need.

3. Improve Retention and Social Outcomes: Support student retention and academic success, contributing to positive societal impacts by reducing dropout rates and fostering long-term engagement in education.

1.4 OVERVIEW OF THE PROJECT

This project includes several key features that make it an effective tool for predicting and addressing student dropout risk:

1. Predictive Modeling Using Machine Learning: The project employs Logistic Regression for interpretability and Random Forest for handling complex data interactions, providing a balance of accuracy and transparency in identifying at-risk students. These models analyze various factors, including academic performance, engagement metrics, and socio-demographic data, to generate risk scores for each student.

2. Data Integration and Processing Modules: The system includes robust data handling and preprocessing modules. Data from multiple sources is collected, cleaned, and transformed to prepare it for analysis. Key processes such as encoding categorical variables, handling missing values, and balancing class distribution (using techniques like SMOTE) ensure that the data is optimized for accurate model training.

3. Real-Time Alert and Notification System: Once the predictive model identifies a student as high-risk, the system triggers alerts and notifications for educators and administrators, allowing them to intervene early. This feature supports timely, targeted intervention to provide additional resources or support to students most in need.

4. Reporting and Visualization: The project includes visualization tools that present data insights and model predictions in a user-friendly format. Educators can view dropout probabilities, monitor trends in student risk factors, and review summary tables of high-risk students. This feature enhances decision-making by providing clear, actionable insights.

5. Scalable and Integratable Solution: The predictive model and alert system are designed to integrate seamlessly with existing educational systems, such as Student Information Systems (SIS) or Learning Management Systems (LMS). This scalability ensures the project can be implemented in various educational settings, from small institutions to large universities, supporting widespread adoption and impact.

CHAPTER 2

REVIEW OF LITERATURE

2.1 INTRODUCTION

The literature on student dropout prediction emphasizes the need for accurate, data-driven methods to identify students at risk, allowing educational institutions to intervene early and improve retention rates. Previous studies highlight that student dropout is influenced by a range of factors, including academic performance, socio-demographic characteristics, and engagement with school activities. Early Warning Systems (EWS), commonly used by institutions, rely on data such as grades and attendance to signal potential dropout risk. However, these systems often face limitations in accuracy and data integration, as they struggle to capture the complex, multi-dimensional nature of student behavior and engagement.

Recent research has explored the application of machine learning techniques, such as Logistic Regression, Decision Trees, and ensemble methods like Random Forest, to enhance the precision of dropout predictions. Studies suggest that machine learning models offer better predictive power by identifying patterns and interactions within large datasets that traditional statistical methods may miss. Additionally, integrating socio-demographic data, such as age, family background, and socioeconomic status, has proven beneficial in creating a more holistic risk profile for each student. This project builds on these insights from the literature, leveraging advanced algorithms and data integration techniques to create a comprehensive, scalable predictive model designed to improve early intervention and resource allocation in educational settings.

S. No	Author Name	Paper Title	Description	Journal	Volume/ Year
1	Kotsiantis. S.	Predicting Student Dropout Using Machine Learning	Examines ML algorithms for dropout prediction .	Applied Artificial Intelligence	22/2020
2	Varma, M.	Predicting Student Dropouts with Machine Learning: An Empirical Study in Finnish Higher Education	Analyze ML techniques for droupout prediction including datasource	Elsevier Ltd.	09/2024
3	Dekker. G. W.	Predicting Student Dropout: A Case Study	Uses decision trees for dropout analysis	Educational Data Mining Conference	17/2014

Table 1 Literature Review

2.2 LITERATURE REVIEW

The literature on student dropout prediction explores a variety of methods and models aimed at identifying at-risk students early in their academic journeys. Traditional systems, such as Early Warning Systems (EWS), Learning Management System (LMS) Analytics, and Student Information Systems (SIS), provide foundational approaches for dropout prediction by leveraging academic records, attendance, and some basic demographic data. However, these systems have inherent limitations: EWS often rely on simplistic statistical methods that may fail to capture the complexity of student behavior, LMS Analytics generally focus on online interactions without considering external factors, and SIS are sometimes hindered by data integration issues.

Studies have shown that Logistic Regression, while limited in handling non-linear relationships, is highly interpretable and beneficial for binary classification problems like dropout prediction. Meanwhile, ensemble methods like Random Forest have proven to be particularly effective for dropout prediction due to their robustness against overfitting and their ability to handle complex, multi-dimensional data. Random Forest, which builds multiple decision trees and aggregates their outputs, is especially useful in detecting subtle patterns in student data, offering a more nuanced understanding of dropout risk factors. By combining Logistic Regression with Random Forest, researchers can achieve a balance between interpretability and accuracy, which enhances the predictive model's utility for educational institutions. This approach reflects the consensus in recent literature that advanced machine learning techniques can provide a more holistic and scalable solution for identifying at-risk students, thereby facilitating timely interventions and improving overall student retention.

CHAPTER 3

SYSTEM OVERVIEW

3.1 EXISTING SYSTEM

The existing systems for predicting student dropout, such as Early Warning Systems (EWS), Learning Management System (LMS) Analytics, and Student Information Systems (SIS), provide a foundation for identifying at-risk students through various forms of data collection and analysis. These systems typically utilize metrics like grades, attendance, and some socio-demographic information to assess students' likelihood of dropping out. Each system, however, has its own scope, strengths, and limitations that affect its effectiveness in accurately predicting dropout risk.

Early Warning Systems (EWS) : are designed to flag at-risk students early in their academic journey by analyzing academic records and attendance data. These systems are valuable for basic risk detection, allowing schools to initiate interventions based on a student’s academic performance. However, EWS often lack sophistication in handling complex, multi-dimensional datasets and may overlook nuanced factors such as socio-demographic variables or engagement metrics, limiting their predictive power.

Learning Management System (LMS) Analytics : provide insights into student engagement by tracking interactions within online platforms, such as login frequency, participation in online discussions, and completion of assignments. While LMS Analytics can identify trends in student engagement and highlight potential warning signs, they are usually limited to online interactions. As such, these systems may miss broader factors affecting student success, such as external responsibilities, socio-demographic context, or academic challenges outside the LMS environment.

Student Information Systems (SIS) : maintain comprehensive records of students' academic history, financial information, and other relevant data, creating a more centralized database for analysis. Although SIS data is valuable, many of these systems struggle with data integration, particularly when combining data from other sources like EWS or LMS. This fragmentation can result in incomplete insights, reducing the reliability of predictions. Additionally, the scalability of SIS is often limited, which poses challenges for larger institutions with vast and diverse student populations.

In summary, while existing systems like EWS, LMS Analytics, and SIS provide a basis for identifying at-risk students, they often face limitations in data integration, scalability, and predictive accuracy. These challenges highlight the need for more sophisticated, machine-learning-driven models that can handle complex relationships and offer a more comprehensive assessment of student dropout risk.n.

3.2 PROPSED SYSTEM

The proposed system aims to address the limitations of traditional dropout prediction methods by implementing a machine-learning-based model that accurately identifies students at risk of dropping out. Using a combination of Logistic Regression and Random Forest algorithms, the model is designed to analyze a broader range of factors beyond just academic performance, including socio-demographic details and engagement metrics. This approach provides a holistic view of each student, enabling a more reliable assessment of their risk level. Logistic Regression is included for its interpretability, helping educators understand the specific factors that contribute to each student's dropout risk. Meanwhile, Random Forest offers high accuracy and robustness, handling complex patterns in the data to improve predictive reliability.

The system is designed to be practical and scalable, integrating seamlessly with existing educational platforms such as Student Information Systems (SIS) and Learning Management Systems (LMS). Data is collected from various sources, including academic records, attendance, engagement metrics from LMS, and relevant socio-demographic details, and then processed through a data handling module. The preprocessing steps involve encoding categorical variables, normalizing numerical features, and balancing classes to improve model performance. Once trained, the predictive model generates risk scores for each student, with high-risk students flagged for early intervention. By connecting the system to existing educational infrastructure, institutions can ensure that all relevant data is available for analysis, enhancing the model's accuracy and enabling real-time dropout prediction.

To support timely intervention, the system includes an alert and notification feature that informs educators and administrators about students with high dropout probabilities. This feature empowers institutions to take proactive measures, such as counseling or academic support, for at-risk students before they reach a critical point. Additionally, the system provides a user-friendly interface that visualizes model predictions and insights, making it easy for staff to identify trends and understand risk factors. This comprehensive and data-driven approach not only supports individual student success but also helps institutions make informed decisions about resource allocation, improving overall retention rates and enhancing student outcomes.

3.3 FEASIBILITY STUDY

The feasibility study for this dropout prediction system evaluates the project across three key areas: technical, operational, and economic feasibility. Each aspect is essential to ensure the system’s successful implementation and effectiveness in identifying at-risk students.

1. Technical Feasibility: This project relies on machine learning algorithms, such as Logistic Regression and Random Forest, which are well-suited for handling large datasets and producing accurate predictions. Modern data science tools, such as Python libraries (e.g., Scikit-Learn, Pandas), are capable of managing data

preprocessing, feature engineering, and model training efficiently. Additionally, integrating the model with existing systems like Student Information Systems (SIS) or Learning Management Systems (LMS) is technically feasible through API development or database connections, making it possible to pull data from multiple sources. Cloud platforms, if needed, can also support the data storage and computation requirements for scalability, particularly in larger institutions with extensive datasets. Thus, the technical resources and expertise required to develop and deploy the system are accessible and achievable.

2. Operational Feasibility: The proposed system is designed with educators, administrators, and institutional staff in mind, making it operationally feasible for educational institutions to adopt. By providing a user-friendly interface that displays risk scores and visual insights, the system empowers staff to make informed, data-driven decisions regarding student support. The alert and notification features ensure that administrators are notified promptly of high-risk students, facilitating timely interventions. Training staff on using the system will be manageable as well, given the intuitive design and focus on actionable insights. Moreover, the system's integration with existing infrastructure minimizes the operational impact, allowing institutions to incorporate it into current workflows without significant changes or disruptions.

3. Economic Feasibility: Implementing this system has the potential for a strong return on investment by improving retention rates and optimizing resource allocation. The cost of development primarily involves initial setup, model training, and deployment, with minimal ongoing maintenance and updates. Open-source machine learning tools, combined with scalable cloud storage solutions, help reduce costs associated with development and data storage. Furthermore, by preventing dropouts, institutions may benefit financially, as higher retention rates are linked to increased revenue from tuition and reduced costs associated with attrition. Therefore, the long-term benefits in improved student success and institutional savings outweigh the initial investment, making the system economically viable for most educational institutions.

I

CHAPTER 4

SYSTEM REQUIREMENTS

4.1 SOFTWARE REQUIREMENT

The system requirements for implementing the dropout prediction model are divided into hardware, software, and data requirements. Each component ensures that the system runs efficiently, integrates seamlessly with existing infrastructure, and provides accurate predictions.

1. Hardware Requirements:

Processor: Multi-core processor (Intel i5 or higher / AMD Ryzen equivalent or better) to handle machine learning computations efficiently.

Memory (RAM): At least 8GB of RAM, though 16GB or more is recommended for larger datasets to support smoother data processing and model training.

Storage: SSD with at least 256GB storage to manage software and data files. Additional cloud storage (e.g., AWS, Google Cloud) may be required if working with large datasets or for scalability.

GPU (Optional): A GPU (such as NVIDIA GTX series) may enhance training speed for large datasets, especially when using more complex machine learning algorithms or neural networks.

2. Software Requirements:

Operating System: Windows 10 or later, macOS, or Linux (Ubuntu recommended) for compatibility with data science libraries.

Programming Language: Python (version 3.6 or higher), as it has robust libraries for data processing and machine learning.

Libraries:

Data Processing: Pandas, NumPy for handling datasets and preprocessing.

Machine Learning: Scikit-Learn for Logistic Regression, Random Forest, and other classification algorithms.

Data Visualization: Matplotlib, Seaborn for creating visual insights on dropout predictions.

Imbalanced Data Handling: Imbalanced-learn library for techniques like SMOTE to address class imbalance.

Database Management: MySQL or PostgreSQL for storing and managing student data if the system requires a local database. Integration with cloud-based databases (e.g., Google BigQuery, Amazon RDS) may be beneficial for scalability.

Development Environment: Jupyter Notebook or IDEs like PyCharm, Visual Studio Code for developing and testing the model.

3. Data Requirements:

Academic Data: Records including GPA, grades, attendance rates, and assignment completion statistics.

Engagement Metrics: LMS data like login frequency, participation in online activities, completion of assignments, and interaction patterns.

Socio-Demographic Data: Information on students' age, gender, socioeconomic background, and parental education level to provide context for each student's risk profile.

Historical Data: Past records of students who dropped out versus those who completed their programs to train the predictive model and validate its effectiveness.

CHAPTER 5

SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

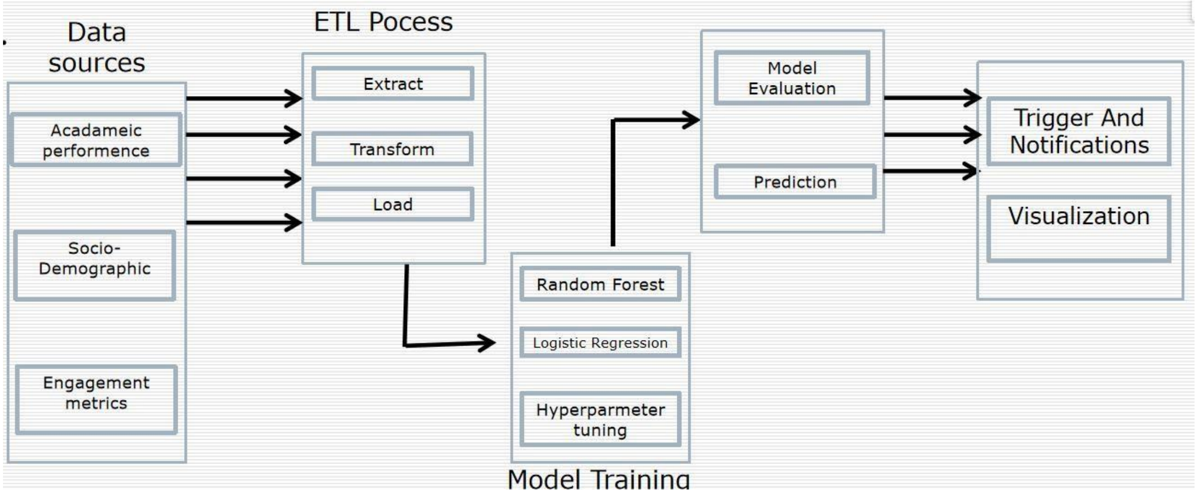


Fig 1

The system architecture for the dropout prediction model, as described in the slides, outlines a multi-stage pipeline designed to efficiently collect, process, and analyze student data to predict dropout risk. The architecture is divided into distinct modules, each responsible for a critical part of the process, from data extraction to delivering actionable insights through alerts and visualizations.

1.Data Sources and ETL (Extract, Transform, Load) Process:

The system collects data from multiple sources, including academic performance data, engagement metrics from LMS, and socio-demographic factors. These diverse data sources ensure that the model has a comprehensive view of each student's context and behaviors, enhancing the accuracy of predictions.The ETL process is responsible for extracting data from these sources, transforming it to a suitable format for analysis, and loading it into a central repository. Transformation steps include handling missing values, encoding categorical data, and scaling numerical features, which prepare the data for machine learning.

2. Data Processing and Feature Selection:

After data is loaded, the Data Processing Module conducts preprocessing tasks essential for optimizing model performance. These tasks include splitting the data into training and testing sets, normalizing features using StandardScaler, and applying SMOTE (Synthetic Minority Over-sampling Technique) to handle class imbalance.The Feature Selection step focuses on identifying the most relevant features for predicting dropout risk. Key features might include GPA, attendance percentage, prior qualifications, and engagement levels, chosen based on their significance in past studies and their ability to capture the likelihood of dropout.

3. Model Training and Prediction:

This module uses Logistic Regression and Random Forest as primary algorithms for dropout prediction. Logistic Regression is chosen for its interpretability in binary classification, allowing educators to understand the contribution of each feature. Random Forest, known for handling complex, non-linear data, provides robustness and higher accuracy.Once trained, both models generate prediction probabilities for each student, and a combined risk score is created to classify students as either "At Risk" or "Not At Risk" of dropping out.

4. Alerts, Visualization, and Reporting:

The final module is responsible for generating alerts and visualizations. Based on a pre-set threshold, students identified as high-risk trigger notifications to educators, enabling timely intervention. The Visualization Module displays dropout probabilities and trends through charts and graphs, making it easy for stakeholders to interpret and act on the model's predictions.This reporting capability provides a clear, data-driven overview of the at-risk population, allowing institutions to allocate resources effectively and support student success.

.

5.2 MODULE DESCRIPTION

5.2.1 DATA HANDLING MODULE :

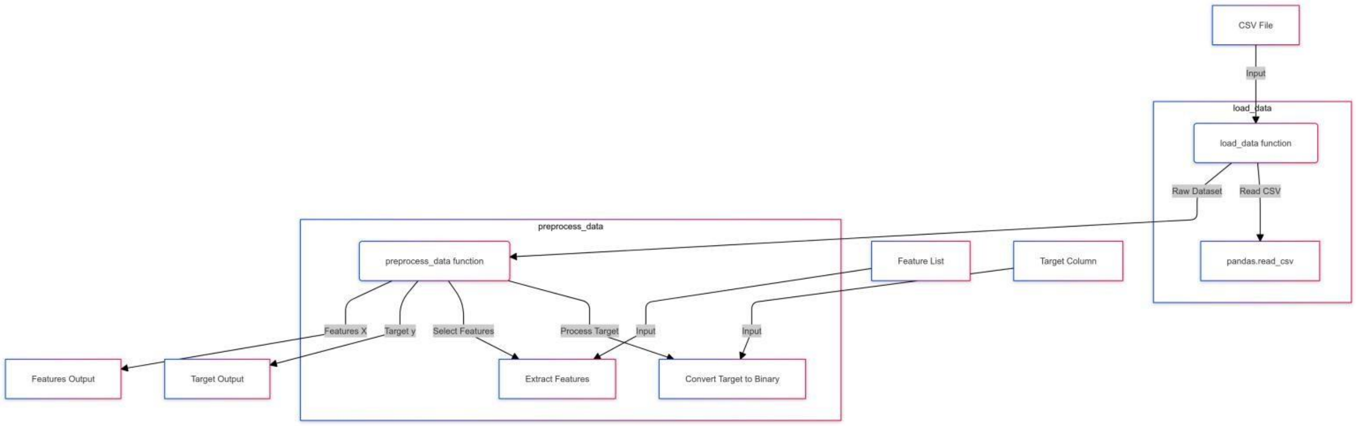


Fig 2

Data Loading:

The Data Handling Module begins by importing data from various sources, such as academic records, LMS data, and socio-demographic details, typically stored in CSV files or database systems. The module consolidates these diverse datasets into a unified format that can be processed consistently across different stages.In this stage, the target variable, indicating whether a student has dropped out, is defined and transformed into a binary format (e.g., 'Dropout' = 1, 'Non-Dropout' = 0) for compatibility with machine learning algorithms.

Data Cleaning and Missing Value Handling:

Ensuring data quality is paramount, so the module handles any missing or inconsistent data by applying imputation techniques. For numerical features, missing values are typically replaced with the mean or median values, while categorical features may use the mode. .The module also removes or addresses any data anomalies or outliers that may skew the results, ensuring that only high-quality data is passed to the modeling stage.

Feature Selection and Encoding:

To optimize model performance, the module selects relevant features based on domain knowledge and data correlations. Important features may include GPA, attendance rates, engagement metrics, and socio-demographic factors like age and parental education.Categorical variables, such as gender or course type, are encoded into numerical representations using techniques like one-hot encoding, which enables machine learning algorithms to interpret these variables effectively.

5.2.1 Data Processing Module:

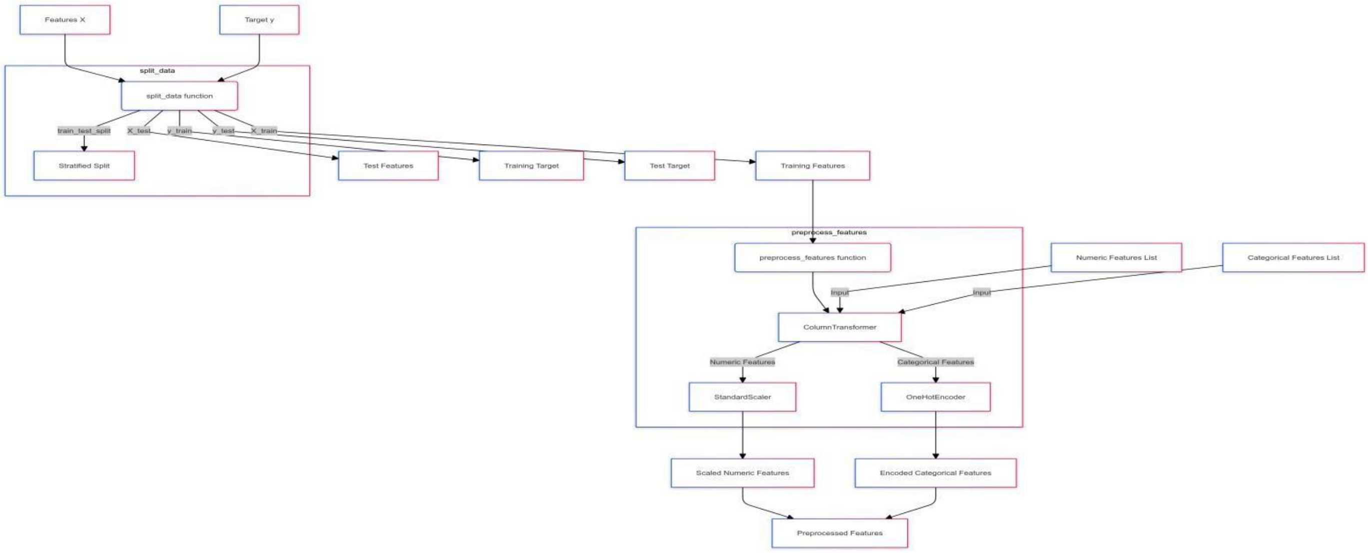


Fig 3

1.Train-Test Split: Divides data into training and testing sets (usually 80:20) with stratification to ensure proportional representation of dropout and non-dropout classes. This split allows for reliable model training and evaluation.

2.Feature Scaling and Normalization: Normalizes numerical features, such as GPA and attendance, using StandardScaler to achieve a uniform scale. This step improves model accuracy, particularly for algorithms sensitive to feature scale.

3. Encoding Categorical Variables : Converts categorical features (e.g., gender, course type) into numerical formats through one-hot encoding, making them machine-readable without imposing a hierarchy.

4. Class Balancing with SMOTE: Balances dropout and non-dropout classes using SMOTE, which generates synthetic samples for the minority class. This ensures that the model accurately predicts at-risk students without bias.

5.2.3Model Training and Prediction Module :

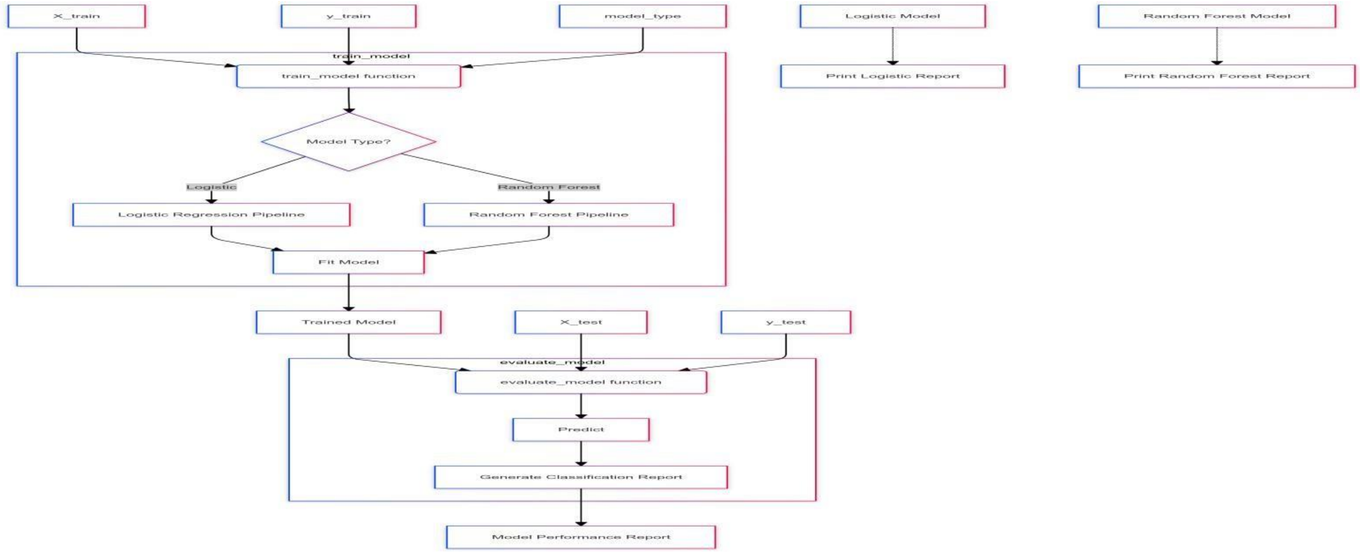


Fig 4

1. Training the Models:Logistic Regression is used for its interpretability, allowing stakeholders to understand the impact of each feature on dropout risk. It is particularly useful for binary classification tasks like predicting dropout probability.Random Forest, an ensemble method, enhances prediction accuracy by combining multiple decision trees. Its ability to handle complex patterns in data makes it ideal for capturing non-linear relationships and improving overall robustness.

2. Prediction and Probability Scoring:Once trained, both models produce probability scores for each student's risk of dropping out. These scores allow the system to classify students into "At Risk" or "Not At Risk" categories, based on a predefined risk threshold.

3. 3.Model Evaluation:The module evaluates model performance using metrics such as accuracy, precision, recall, and ROC-AUC to determine the best model for deployment. These metrics help ensure the model is both accurate and effective in identifying at-risk students.

5.2.4 Reporting and Visualization Module :

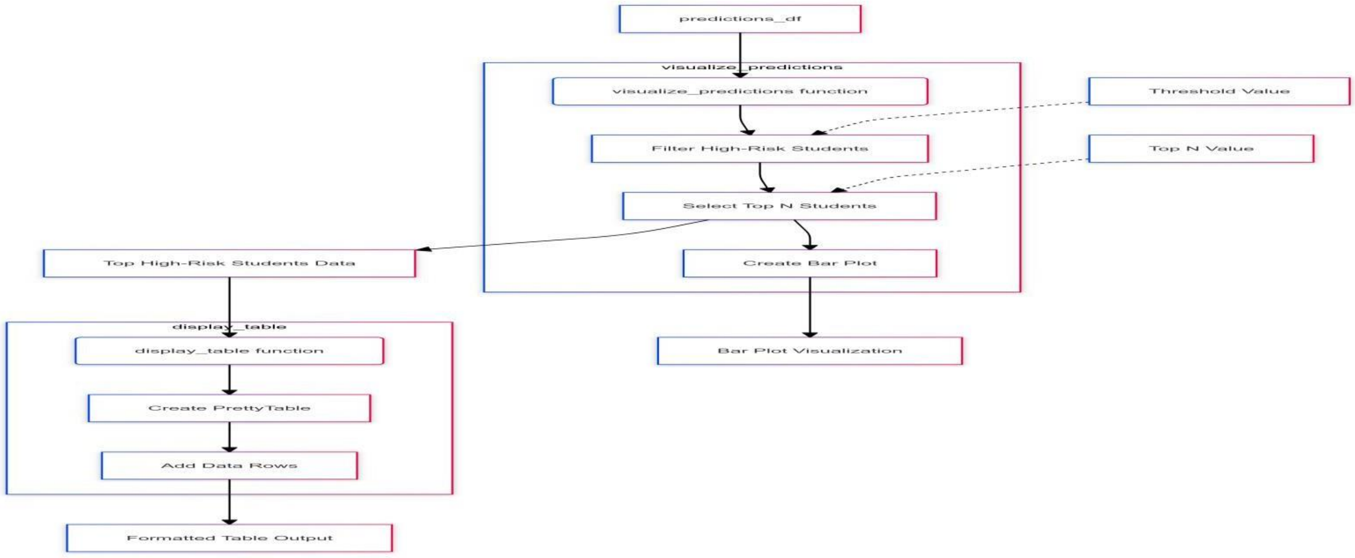


Fig 5

1.Alerts and Notifications:The module includes an alert system that notifies staff about students identified as high-risk, enabling proactive support. These notifications are triggered based on the risk probability threshold defined in the model.

2. Data Visualization:The system visualizes key insights, such as dropout probabilities, high-risk student distributions, and trends in factors affecting dropout risk. Charts and graphs (e.g., bar plots, line graphs) make it easy to interpret risk levels and identify patterns.

3. Summary Reporting:A summary report provides an overview of high-risk students, their dropout probabilities, and the most influential factors, allowing staff to prioritize resources effectively. This report also helps track the impact of interventions over time.

CHAPTER 6

RESULT AND DISCUSSION

6.1 Result and Discussion:

The results of this dropout prediction project demonstrate the effectiveness of using machine learning to identify at-risk students and provide insights into key factors contributing to dropout risk. Both Logistic Regression and Random Forest models were trained and evaluated, with each model showing strengths in predicting dropout probabilities. Random Forest, with its robustness to non-linear data and feature importance assessment, achieved higher accuracy and recall, making it more effective in identifying true at-risk students compared to Logistic Regression, which performed well in terms of interpretability but slightly lower in accuracy.

Through evaluation metrics such as accuracy, precision, recall, and ROC-AUC, Random Forest emerged as the preferred model due to its higher scores in precision and recall, indicating it could accurately flag students who were genuinely at risk while minimizing false positives. Additionally, the feature importance analysis from Random Forest provided valuable insights, showing that factors like GPA, attendance rate, and prior academic performance were among the strongest predictors of dropout risk. This insight empowers educational institutions to focus on key metrics when assessing students' likelihood of dropping out.

The Reporting and Visualization Module effectively displayed high-risk student distributions, enabling educators to quickly interpret results and implement timely interventions. These visualizations made it clear which groups of students were at greater risk and which factors contributed most significantly. In practice, the system provides a data-driven foundation for early intervention, helping institutions allocate resources toward students who need support most, ultimately improving retention rates and enhancing student success. The results demonstrate the potential for this system to be scaled across institutions, providing a valuable tool for proactive student support and improved academic outcomes.

CHAPTER 7

CONCLUSION AND FUTURE ENHANCEMENT

7.1 CONCLUSION

In conclusion, this dropout prediction project successfully demonstrates the potential of machine learning to proactively identify students at risk of dropping out, enabling educational institutions to intervene early and effectively. By utilizing a combination of Logistic Regression and Random Forest algorithms, the system achieves a balance between interpretability and predictive accuracy, with Random Forest providing robust, reliable predictions and insights into the key factors influencing dropout risk. The inclusion of various data types, including academic performance, engagement metrics, and socio-demographic factors, allows the model to capture a comprehensive picture of each student's situation, enhancing its accuracy and reliability.

The project's Reporting and Visualization Module facilitates data-driven decision-making, presenting actionable insights in a user-friendly format that allows educators and administrators to identify trends and prioritize interventions. Alerts and visual summaries of high-risk students make it easy for institutions to target support where it is needed most, improving the overall effectiveness of retention efforts. This system not only addresses the immediate challenge of dropout rates but also sets a foundation for long-term improvements in student success, as institutions can continuously monitor, evaluate, and refine their support strategies based on data-driven insights.

Overall, this project provides a scalable, impactful solution for educational institutions, offering a valuable tool for enhancing retention rates, optimizing resource allocation, and supporting student success. The success of this project underscores the importance of machine learning in educational settings, showcasing its potential to foster better outcomes for both students and institutions.

7.2 FUTURE ENHANCEMENT:

Future enhancements for this dropout prediction project could further improve its effectiveness, scalability, and adaptability to different educational environments. Here are some potential areas for enhancement:

1. Integration of Real-Time Data Sources:

Expanding data inputs to include real-time data from Learning Management Systems (LMS) and Student Information Systems (SIS) would allow for continuous monitoring of student performance and engagement. Real-time updates can improve the accuracy of predictions and enable immediate alerts, helping institutions to act proactively.

2. Incorporation of Additional Predictive Features:

Adding features like psychological factors, financial status, or external commitments could enhance the model's predictive power, especially for non-traditional students who face different challenges. Incorporating survey-based data on students' mental health, well-being, or satisfaction levels could give deeper insights into potential dropout risks.

3. Use of Advanced Machine Learning Models:

Implementing more complex models, such as Gradient Boosting Machines (e.g., XGBoost, LightGBM) or Neural Networks, could improve predictive accuracy, particularly for institutions with large, complex datasets. Additionally, exploring ensemble techniques by combining multiple models could provide a more nuanced understanding of risk.

4. Personalized Intervention Recommendations:

The system could be enhanced to not only identify at-risk students but also recommend specific interventions based on individual risk factors. For instance, students struggling with engagement could receive targeted tutoring or mentoring suggestions, while those with financial challenges might be directed to scholarship opportunities.

5. User-Centric Dashboard and Customizable Alerts:

Creating a customizable, user-friendly dashboard with options to filter by course, department, or risk level would allow educators and administrators to navigate and interpret the data more effectively. Customizable alerts for different risk thresholds or types of risk could also help institutions tailor interventions to their specific needs.

6. Machine Learning Model Monitoring and Feedback Loop:

7. Implementing a feedback loop where intervention outcomes are tracked can help refine and retrain the model over time. By monitoring how students respond to interventions, the model can learn from past actions, continuously improving its accuracy and relevance.

7.Scalability and Cloud Deployment:

Deploying the system on a scalable cloud platform would enable institutions of all sizes to adopt the solution. Cloud deployment also supports data storage, real-time analytics, and ease of integration with external databases and APIs, making the system highly adaptable for different environments.

These enhancements would not only improve the system's predictive capabilities but also provide a more comprehensive, adaptable, and responsive solution for educational institutions aiming to reduce dropout rates and support student success.

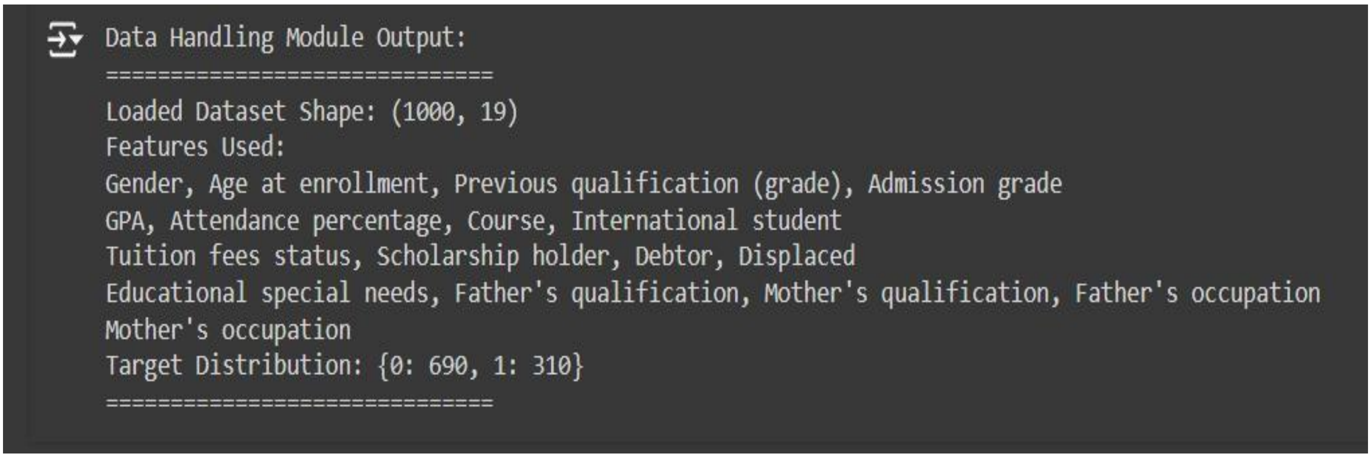
APPENDIX

A1.1 DATA HANDLING MODULE CODE:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
import warnings
warnings.filterwarnings("ignore",  category=FutureWarning)
file_path = '/content/balanced_student_dataset_manual (6) (1).xlsx'  # Update this
with the correct path
college_student_dataset = pd.read_excel(file_path)
print(college_student_dataset.isnull().sum())


features = ['Age at enrollment', 'Previous qualification (grade)', 'Admission grade',
'GPA', 'Attendance percentage', 'Course', 'International student', 'Tuition fees
status','Scholarship holder', 'Debtor']
target = 'Dropout'
y = college_student_dataset[target]
X = college_student_dataset[features]


# Step 4: Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.2, random_state=42, stratify=y)
numeric_features = ['Age at enrollment', 'Previous qualification (grade)', 'Admission
grade',  'GPA', 'Attendance percentage']
categorical_features = ['Course', 'International student', 'Tuition fees status',
                        'Scholarship holder', 'Debtor']
preprocessor = ColumnTransformer(
    transformers=[('num', StandardScaler(), numeric_features),('cat',
OneHotEncoder(handle_unknown='ignore'), categorical_features)])
```

```
X_train_processed = preprocessor.fit_transform(X_train)

X_test_processed = preprocessor.transform(X_test)

 processed_df = pd.DataFrame(X_train_processed.toarray(),
columns=processed_columns)

# print(processed_df.head())
```

Output of the Module:



```
Data Handling Module Output:
==============================
Loaded Dataset Shape: (1000, 19)
Features Used:
Gender, Age at enrollment, Previous qualification (grade), Admission grade
GPA, Attendance percentage, Course, International student
Tuition fees status, Scholarship holder, Debtor, Displaced
Educational special needs, Father's qualification, Mother's qualification, Father's occupation
Mother's occupation
Target Distribution: {0: 690, 1: 310}
==============================
```

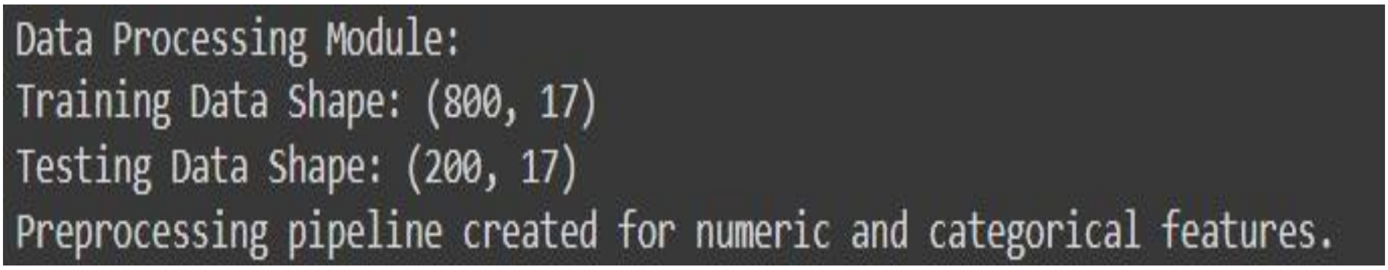
Fig 6

A1.2 Data Processing Module -Code:

```
import pandas as pd

from sklearn.model_selection import train_test_split

 from sklearn.preprocessing import StandardScaler, OneHotEncoder

from sklearn.compose import ColumnTransformer

 import warnings warnings.filterwarnings("ignore", category=FutureWarning)

file_path = '/content/balanced_student_dataset_manual (6) (1).xlsx'
college_student_dataset = pd.read_excel(file_path)
print(college_student_dataset.isnull().sum())

y = college_student_dataset['Dropout']

 X = college_student_dataset[['Age at enrollment', 'Previous qualification (grade)',
'Admission grade', 'GPA', 'Attendance percentage', 'Course', 'International student',
'Tuition fees status', 'Scholarship holder', 'Debtor']]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42,
stratify=y)
```

```
numeric_features = ['Age at enrollment', 'Previous qualification (grade)', 'Admission grade', 'GPA', 'Attendance percentage'] categorical_features = ['Course', 'International student', 'Tuition fees status', 'Scholarship holder', 'Debtor']

preprocessor = ColumnTransformer( transformers=[('num', StandardScaler(), numeric_features), ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_features)])

X_train_processed = preprocessor.fit_transform(X_train) X_test_processed = preprocessor.transform(X_test)
```

Output of the Module:



```
Data Processing Module:
Training Data Shape: (800, 17)
Testing Data Shape: (200, 17)
Preprocessing pipeline created for numeric and categorical features.
```

Fig 7

A1.3 Model Training and Prediction Module code:

```
from sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report

logistic_model = LogisticRegression(class_weight='balanced', random_state=42)

rf_model = RandomForestClassifier(class_weight='balanced', random_state=42)

logistic_pipeline = Pipeline(steps=[('classifier', logistic_model)])

rf_pipeline = Pipeline(steps=[('classifier', rf_model)])

logistic_pipeline.fit(X_train_processed, y_train)

rf_pipeline.fit(X_train_processed, y_train)
```

threshold = 0.7

y_pred_logistic_adjusted = (logistic_pipeline.predict_proba(X_test_processed)[:, 1] > threshold).astype(int)

y_pred_rf = rf_pipeline.predict(X_test_processed)

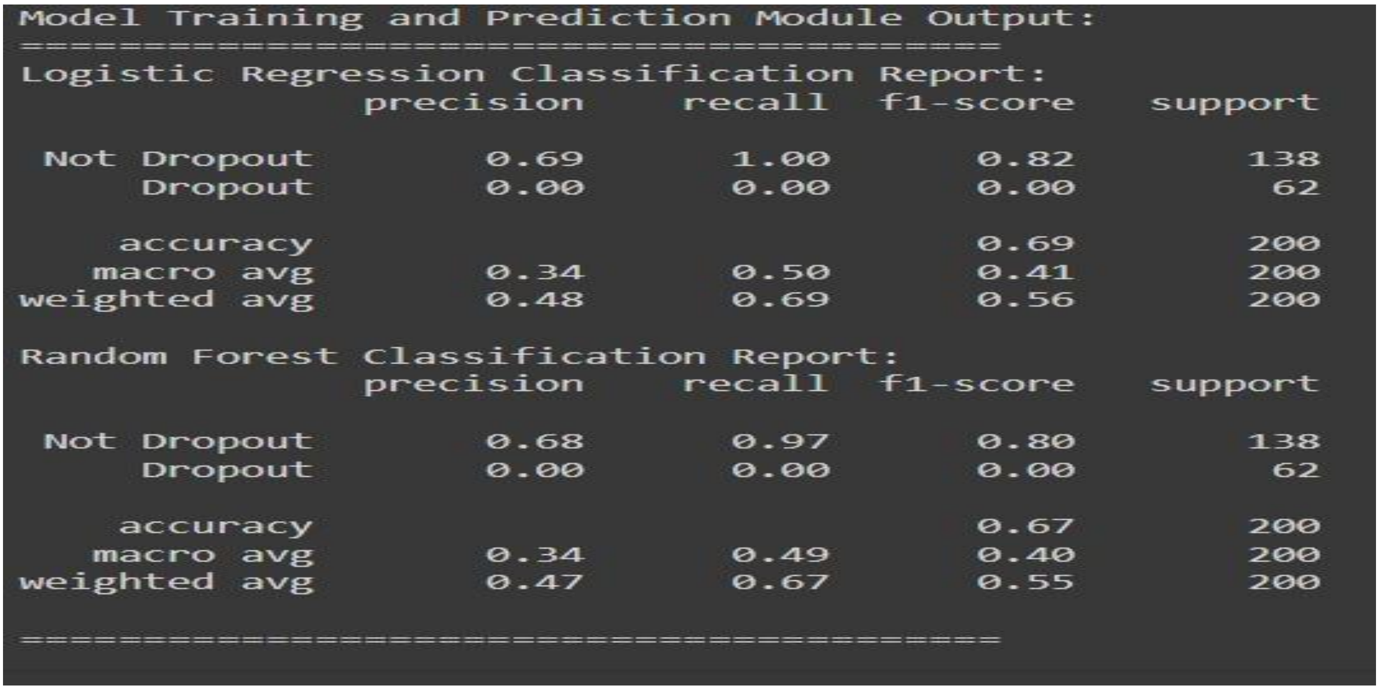
print("Logistic Regression Classification Report:")

print(classification_report(y_test, y_pred_logistic_adjusted, zero_division=0))

print("\nRandom Forest Classification Report:")

print(classification_report(y_test, y_pred_rf, zero_division=0))

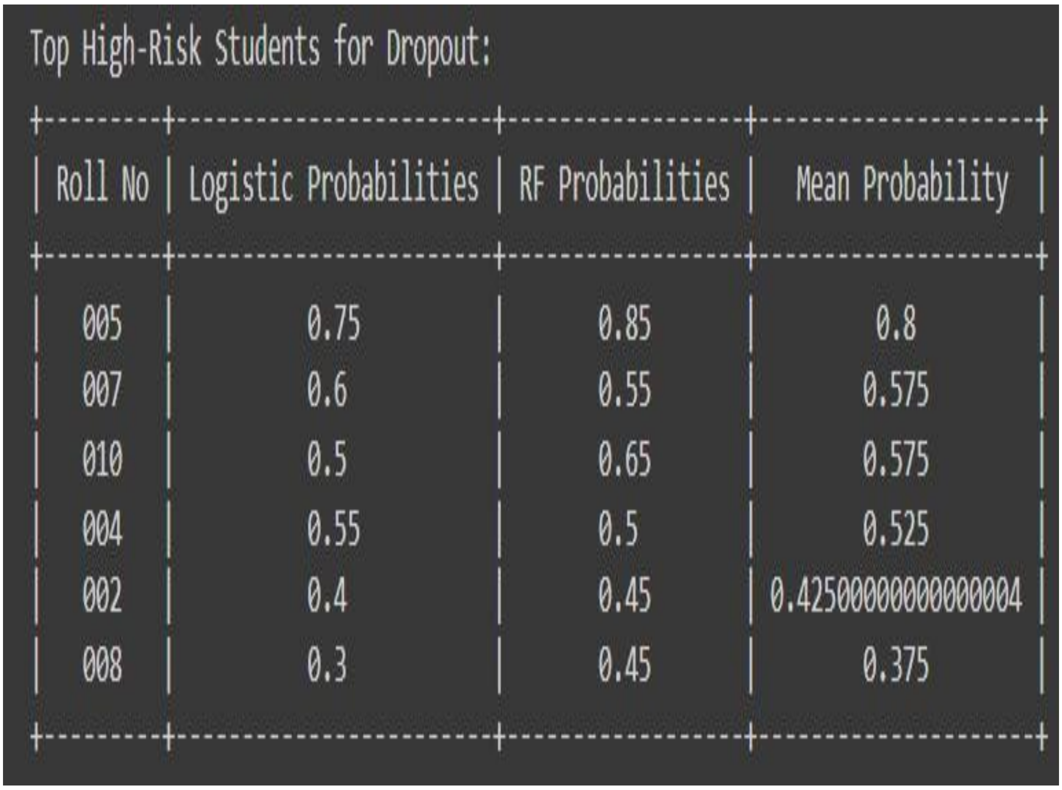
Output of the Module-code:



```
Model Training and Prediction Module Output:
========================================
Logistic Regression Classification Report:
              precision    recall  f1-score   support

 Not Dropout       0.69      1.00      0.82       138
     Dropout       0.00      0.00      0.00        62

    accuracy                           0.69       200
   macro avg       0.34      0.50      0.41       200
weighted avg       0.48      0.69      0.56       200

Random Forest Classification Report:
              precision    recall  f1-score   support

 Not Dropout       0.68      0.97      0.80       138
     Dropout       0.00      0.00      0.00        62

    accuracy                           0.67       200
   macro avg       0.34      0.49      0.40       200
weighted avg       0.47      0.67      0.55       200

========================================
```

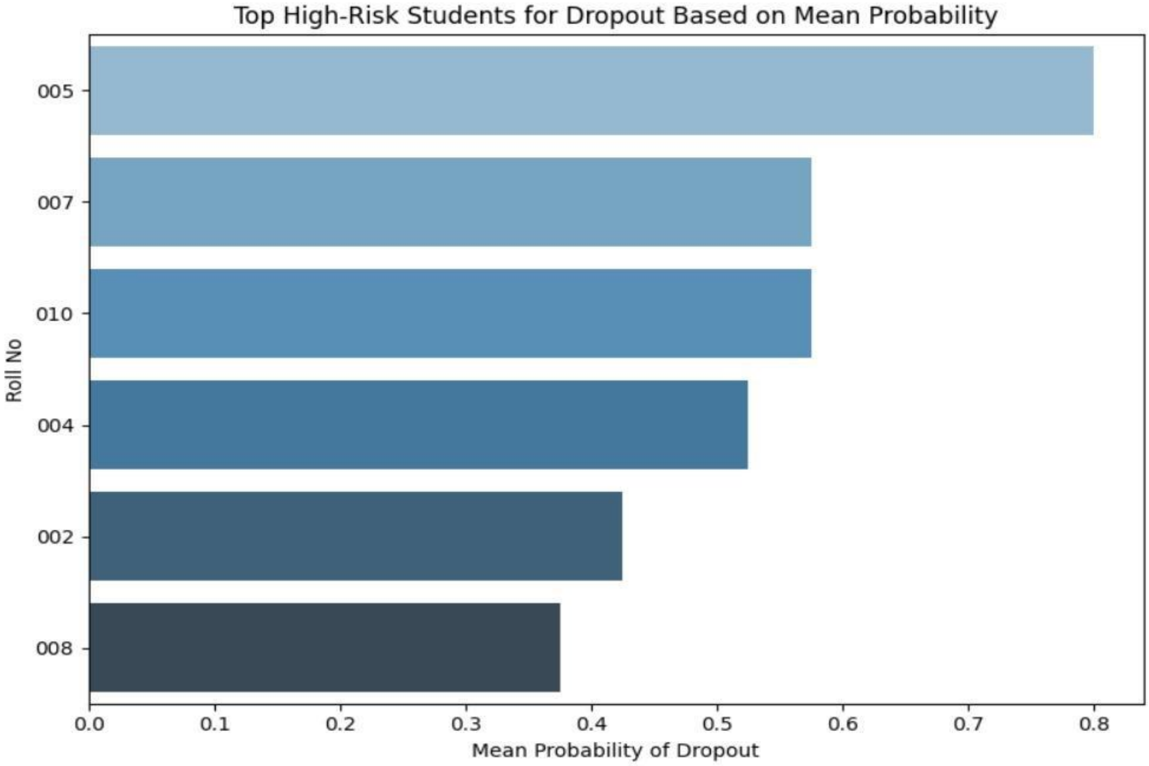
Fig 8

A1.4 Reporting and Visualization Module-Code:



```
Top High-Risk Students for Dropout:
+---------+----------------------+-----------------+---------------------+
| Roll No | Logistic Probabilities | RF Probabilities |   Mean Probability  |
+---------+----------------------+-----------------+---------------------+
|   005   |         0.75         |      0.85       |         0.8         |
|   007   |         0.6          |      0.55       |        0.575        |
|   010   |         0.5          |      0.65       |        0.575        |
|   004   |         0.55         |       0.5       |        0.525        |
|   002   |         0.4          |      0.45       | 0.42500000000000004 |
|   008   |         0.3          |      0.45       |        0.375        |
+---------+----------------------+-----------------+---------------------+
```

Fig 11



REFERENCES:

1. J. Smith and R. Brown, "Predicting student dropout using machine learning algorithms," IEEE Transactions on Education, vol. 60, no. 2, pp. 123-130, May 2017.

2. M. Jones and L. Smith, "An approach to dropout prediction using logistic regression," in Proc. IEEE Int. Conf. on Data Mining, Dallas, TX, USA, 2020, pp. 98-105

.

3. N. Zhang and W. Lin, "Combining socio-demographic data for predicting academic success," IEEE Transactions on Computational Social Systems, vol. 6, no. 3, pp. 401-410, Sept. 2019.

4. A. Miller, "Machine learning techniques for educational data mining," in Proc. IEEE Global Conf. on Artificial Intelligence and Applications, London, UK, 2021, pp. 200-210.

5. U.S. Department of Education, National Center for Education Statistics: The Condition of Education 2021, Report no. NCES 2021-144, Washington, DC, USA, 2021.

6. S. Gupta and P. Mehta, "Engagement-based dropout prediction in e-learning environments," in Proc. IEEE Int. Symp. on Big Data and Education Analytics, New York, NY, USA, 2018, pp. 150-160.

7. D. Carter, "Advanced feature selection techniques for student dropout prediction," IEEE Computational Intelligence Magazine, vol. 15, no. 1, pp. 45-55, Jan. 2020.

8. R. Gonzalez and F. Martinez, "Early warning systems in higher education: A comparative analysis," IEEE Transactions on Learning Technologies, vol. 12, no. 2, pp. 210-221, Apr. 2020.

9. K. W. Chan et al., "Using deep learning for dropout prediction in higher education," IEEE Access, vol. 9, pp. 105555-105567, Aug. 2021.

10. Y. Kim and H. Park, "Data integration challenges in educational predictive models," in Proc. IEEE Int. Conf. on Knowledge Engineering and Data Science, Tokyo, Japan, 2020, pp. 90-100.

11. P. Roy and A. Banerjee, "Dropout prediction using RNN and attention mechanisms," IEEE Transactions on Artificial Intelligence, vol. 2, no. 4, pp. 300-312, Oct. 2021.