STUDENT DROPOUT PREDICTION IN ONLINE PLATFORM USING XG BOOST AND DECISION TREE

SECA3028-MACHINE LEARNING

TECHNIQUES

Submitted in partial fulfilment of the requirements for the award of

Bachelor of engineering degree in Electronics and Communication Engineering

by

ASWIN VISHAL (42130055)

YAZIN H (42130553)

YELLEPALLI CHAITANYA NADH (42130554)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
SCHOOL OF ELECTRICAL AND ELECTRONICS

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A++" by NAAC

JEPPIAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600119

AUGUST – 2025

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A++" by NAAC

Jeppiar Nagar, Rajiv Gandhi Salai, Chennai – 600119

www.sathyabama.ac.in

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the Bonafide work of **ASWIN VISHAL A (42130055), YAZIN H(42130553) and YELLEPALLI CHAITANYA NADH (42130554)** who carried out the project Entitled "**STUDENT DROPOUT PREDICTION IN ONLINE PLATFORM USING XG BOOST AND DECISION TREE**" under our supervision from June 2025 to August 2025.

Internal Guide

Dr. P. KAVIPRIYA, M. E., Ph.D.,

Head of the Department

Dr. T. RAVI, M.E., Ph.D.,

Submitted for Viva voce examination held on _____

Internal Examiner **External Examiner**

DECLARATION

We **ASWIN VISHAL A (42130055), YAZIN H (42130553) and YELLEPALLI CHAITANYA NADH (42130554)** hereby declare that the Project Report entitled **"STUDENT DROPOUT PREDICTION IN ONLINE PLATFORM USING XG BOOST AND DECISION TREE"** is done by us under the guidance of **Dr. P. KAVIPRIYA M.E., Ph.D.,** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering Degree in Electronics and Communication Engineering.

PLACE: Chennai

DATE:

SIGNATURE OF THE CANDIDATES

1.

2.

3.

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey our thanks **to Dr. N. M. NANDHITHA, M.E., Ph.D., Professor & Dean, School of Electrical and Electronics and Dr. T. RAVI, M.E., Ph.D., Professor & Head, Department of Electronics and Communication Engineering** for providing us necessary support and details at the right time during the progressive reviews.

We would like to express my sincere and deep sense of gratitude to our Mentor **Dr. P. KAVIPRIYA, M.E., Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all Teaching and Non-teaching staff members of the Department of Electronics and Communication Engineering who were helpful in many ways for the completion of the project.

ABSTRACT

The rapid growth of online learning platforms has transformed the educational landscape by providing flexible, accessible, and personalized learning opportunities to a wide range of students. However, this expansion has also highlighted a major challenge—high student dropout rates. Identifying and predicting student dropouts is a critical concern for educators and institutions, as it directly affects academic performance, resource utilization, and the overall credibility of e-learning systems. The present project focuses on developing a predictive framework for student dropout detection using advanced machine learning algorithms such as XGBoost and Decision Tree classifiers. By analyzing student-related attributes such as engagement levels, attendance records, performance indicators, and participation behavior, the proposed system aims to generate early warnings that enable timely intervention and guidance.

The methodology involves collecting and preprocessing data, followed by applying machine learning techniques to identify key patterns that contribute to dropout tendencies. Ensemble learning models, particularly XGBoost, have been utilized to improve prediction accuracy and minimize overfitting, while Decision Trees provide interpretability for educators to understand the underlying reasons for dropout risks. The architecture integrates multiple stages including feature extraction, model training, evaluation, and validation, ensuring the reliability and robustness of the system. Experimental results demonstrate that the proposed model achieves high accuracy and efficiency in predicting dropout-prone students, significantly outperforming traditional methods. The findings validate that predictive modeling can be an effective tool to support proactive decision-making in online education.The proposed solution not only addresses an existing challenge in the educational domain but also paves the way for future innovations. By providing actionable insights, the system empowers institutions to design intervention strategies, improve curriculum delivery, and enhance student retention. The scalability of the model ensures adaptability across various online platforms, making it a versatile and practical contribution to the field of educational data mining. In conclusion, this project highlights the potential of machine learning-driven analytics in reshaping online education by reducing dropout rates, supporting student success, and fostering a data-driven academic ecosystem.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE .NO
	ABSTRACT	**v**
	LIST OF FIGURES	**vii**
1	**INTRODUCTION**	8
	1.1 Background of Study	8
	1.2 Problem Definition	8
	1.3 Motivation for the Project	9
	1.4 Importance of Dropout Prediction	9
	1.5 Overview of Current status Quo	9
	1.6 Organization of the Report	10
2	**LITERATURE SURVEY**	13
	2.1 Overview of Existing Systems	13
	2.2 National and International Research Studies	13
	2.3 Comparative Study of Approaches	14
	2.4 Open Problems in Existing Systems	15
3	**REQUIREMENTS ANALYSIS**	16
	3.1 Aim of the Project	16
	3.2 Scope of the Project	17
	3.3 Feasibility Study	17
	3.3.1 Technical Feasibility	17
	3.3.2 Economic Feasibility	17
	3.3.3 Operational Feasibility	17
	3.4 Risk Analysis	18
	3.5 Requirement Specification	18
4	**SYSTEM DESIGN AND METHODOLOGY**	21
	4.1 Data Description	21
	4.2 Data Preprocessing and Cleaning	21
	4.3 System Architecture and Overall Design	23

	4.4 Machine Learning Algorithms used	26
	4.5 Implementation Plan	29
	4.6 Testing Strategy and Schemes	29
5	**RESULTS AND PERFORMANCE ANALYSIS**	31
	5.1 Experimental Setup	31
	5.2 Model Training and Testing	32
	5.3 Performance Metrics	32
	5.4 Comparative Analysis of Algorithm	34
	5.5 Discussion of Results	35
6	**SUMMARY AND CONCLUSIONS**	36
	6.1 Summary of the Work done	36
	6.2 Key Findings	37
	6.3 Contributions of the Project	37
	6.4 Limitations of the Current Work	37
	6.5 Future Enhancements	37
	REFERENCES	40

LIST OF FIGURES

Figure No.	Figure Title	Page No.
4.1	Sample data	22
4.2	Working of Machine Learning model	25
4.3	Data Processing	25
4.4	Decision tree	26
4.5	XG Boost Algorithm	27
4.6	Bagging Algorithm	28
5.1	Ada Boost-Confusion Matrix	32
5.2	Bagging-Confusion Matrix	33
5.3	Accuracy and Precision	33

CHAPTER-1

INTRODUCTION

1.1 BACKGROUND OF THE STUDY:

Student dropout has become a significant and complex problem in higher education, carrying substantial academic, social, and economic consequences for all stakeholders. Academically, it represents a failure to achieve educational goals, leading to unfulfilled potential and a decline in institutional graduation rates. Socially, it can result in a loss of confidence and professional opportunities for the individual. Economically, dropout incurs direct financial losses for students (unpaid tuition and debt) and institutions (lost revenue), while also impacting the broader economy through a less-skilled workforce. The traditional reactive approach, which typically involves intervening only after a student's grades have severely declined or they have already ceased attending, is often too late to be effective. The emergence of big data and digital learning platforms presents a paradigm shift, enabling a proactive strategy powered by machine learning to identify and support at-risk students before they disengage.

1.2 PROBLEM DEFINITION:

This project, titled "Predicting Student Dropout: A Machine Learning Approach," aims to address this critical challenge by developing a robust, data-driven system capable of accurately predicting student dropout risk early in their academic journey. The system is designed to function as an early warning mechanism by leveraging a diverse set of historical data, including academic records, demographic information, and behavioral patterns from Learning Management Systems (LMS). By identifying subtle patterns and correlations that are invisible to human analysis, the system provides a quantitative risk score for each student, allowing for timely and targeted intervention.

1.3 MOTIVATION FOR THE PROJECT:

The primary motivation is to transition from reactive to proactive student retention strategies. Instead of simply reacting to poor grades, this project aims to provide academic institutions with actionable, data-driven insights. By predicting which students are at a high risk of dropping out, institutions can allocate resources more efficiently, offering personalized support such as academic counseling, mental health services, financial aid advice, or peer mentoring. This proactive approach not only improves student retention and success but also strengthens the overall academic and financial health of the institution.

1.4 IMPORTANCE OF DROPOUT PREDICTION IN ONLINE LEARNING:

In the rapidly growing domain of online learning, the importance of accurate dropout prediction is amplified. Online students often lack the face-to-face interaction and social integration of traditional campuses, making it easier for them to become disengaged and isolated. However, online platforms generate a rich stream of behavioral data, including login frequency, time spent on course materials, forum participation, and assignment submission times. This digital footprint provides a powerful, quantifiable proxy for student engagement and academic commitment, which can be leveraged by machine learning models to enhance predictive accuracy beyond what is possible with traditional data alone.

1.5 OVERVIEW OF CURRENT STATUS QUO:

The current status quo in student retention relies heavily on manual, often anecdotal, methods. This includes mid-term grade reports, surveys, and reactive responses to student complaints. These methods are inefficient, subjective, and prone to significant delays. They often fail to capture the multi-dimensional nature of student attrition, missing key behavioral and socio-economic indicators. This project offers a scientifically grounded, scalable, and automated alternative that can process large volumes of data to provide a far more comprehensive and timely assessment of risk.

1.6 ORGANIZATION OF THE REPORT:

This report is meticulously structured to provide a complete overview of the project. Chapter 1 serves as an introduction. Chapter 2 presents a detailed literature survey of the relevant theoretical and empirical research. Chapter 3 outlines the project's aim, scope, and feasibility. Chapter 4 describes the methodology, detailing the dataset, data preprocessing, and algorithms used. Chapter 5 presents and discusses the experimental results and performance analysis. Finally, Chapter 6 provides a comprehensive summary, concluding remarks, and a discussion of the project's limitations and future enhancements.

The challenge of student dropout is a global concern that impacts not only the learners but also the reputation and efficiency of educational institutions. In online learning platforms, the problem is even more significant due to factors such as lack of direct supervision, reduced peer interaction, and varying levels of learner motivation. Identifying at-risk students at an early stage is essential for implementing timely support measures like counseling, personalized learning pathways, and targeted academic resources. By applying machine learning techniques, this project seeks to provide a data-driven approach to address the dropout issue, ensuring improved student engagement and fostering long-term academic success.

Moreover, with the rapid expansion of e-learning platforms and virtual classrooms, the amount of data generated from student activities has increased significantly. This data offers valuable insights into learning behaviors, progress trends, and participation levels that can be harnessed to predict potential dropouts. Traditional methods of tracking student performance often fail to capture hidden patterns or subtle warning signs, whereas machine learning models can analyze large-scale data efficiently and uncover complex relationships. This makes predictive analytics a powerful tool to enhance educational outcomes, reduce dropout rates, and support students in achieving their academic goals.

CHAPTER 2

LITERATURE SURVEY

Baker and Inventado (2014) laid the foundation for the application of educational data mining techniques in predicting student dropout and performance in online learning environments. They argued that the vast amounts of learner-generated data can be analyzed to uncover patterns of disengagement, poor performance, and eventual withdrawal. Their work highlighted the potential of data-driven models in identifying at-risk students at an early stage, allowing institutions to implement timely interventions. Building upon this perspective, Xing and Du (2019) developed machine learning models that specifically focused on analyzing learner activity logs in Massive Open Online Courses (MOOCs). Their findings revealed that activity-related features such as time spent on the platform, frequency of logins, and consistency in participation strongly influenced dropout tendencies. Similarly, Li et al. (2016) used decision trees and random forest classifiers to demonstrate how predictive models can efficiently detect students who are likely to withdraw, showcasing the importance of early detection systems in enhancing retention strategies. Together, these studies underline the role of predictive analytics as a critical tool in minimizing dropout rates.

Further evidence of the predictive power of learner behavioral data comes from Halawa, Greene, and Mitchell (2014), who focused on clickstream data of MOOC learners. Their research indicated that the sequence and frequency of online activities serve as strong predictors of learner persistence. Amrieh, Hamtini, and Aljarah (2016) extended this work by proposing a hybrid machine learning framework that integrated multiple classification algorithms. Their approach achieved superior accuracy in predicting dropout, proving that ensemble methods can outperform individual models. Similarly, Kloft et al. (2014) employed support vector machines to analyze temporal learning behaviors, demonstrating that the progression of engagement over time is often more informative than static measures. These works collectively emphasize that dropout prediction models must incorporate not just demographic or performance data, but also dynamic behavioral indicators that evolve throughout the learning process.

Beyond behavioral analysis, researchers have also focused on psychological and motivational aspects of student dropout. Lee and Choi (2011) examined the non-academic factors influencing e-learning persistence and identified learner motivation, satisfaction, and social interaction as major predictors of success. Their study suggested that even with strong content delivery, students are more likely to drop out if they lack intrinsic motivation or peer engagement. Similarly, Chen and He (2013) discovered that participation frequency is directly correlated with retention, highlighting that regular engagement plays a crucial role in academic persistence. Wang and Chen (2015) used logistic regression models to further validate this by incorporating demographic variables along with behavioral factors, finding that both sets of variables complement each other in predicting dropout. Zhao et al. (2020) advanced these ideas by leveraging deep learning models capable of capturing sequential learning behaviors. Their results showed that deep learning outperforms traditional classifiers, as it can detect complex, nonlinear relationships hidden within student activity data, thereby providing more accurate dropout predictions.

Other researchers have explored the use of advanced machine learning techniques to refine dropout prediction. Costa et al. (2017) employed neural networks to predict student performance and observed that dropout forecasts can be highly accurate even with limited input features. This suggested that carefully chosen predictors, rather than large amounts of redundant data, could improve the efficiency of prediction systems. Yang, Sinha, Adamson, and Rose (2013) investigated the role of online discussions and interactions, showing that students with low forum participation and minimal collaboration were far more likely to withdraw from courses. Zimmerman (2008) contributed from a psychological standpoint by linking dropout to poor self-regulated learning skills. His research demonstrated that students lacking the ability to plan, monitor, and control their learning activities struggled to maintain consistency, which ultimately led to higher dropout rates. These works stress the importance of combining psychological, behavioral, and technological factors in designing holistic prediction models.

Finally, Khalil and Ebner (2014) provided an extensive survey of existing research on dropout in MOOCs and concluded that predictive models are indispensable for improving the sustainability of online education. They emphasized that dropout is not just a learner-specific problem but also an institutional concern that affects course design and credibility. In a similar effort, Truong (2016) reviewed the applications of learning analytics in predicting student dropout and highlighted the significance of integrating academic, behavioral, and social factors into predictive systems. His findings demonstrated that a multi-dimensional approach yields better prediction accuracy compared to relying solely on academic performance metrics. Together, these studies reveal a consistent theme: the dropout phenomenon in online education is complex and multi-faceted, requiring the integration of machine learning, behavioral analysis, and psychological insights to develop robust solutions.

2.1 OVERVIEW OF EXISTING SYSTEMS:

The field of student dropout prediction has evolved significantly over the last several decades. Early research, primarily in the 1970s and 80s, was dominated by sociological and psychological theories of student attrition, such as Tinto's Model of Student Integration and Bean's Model of Student Attrition. These theories, while foundational, provided conceptual frameworks rather than predictive tools. The advent of Educational Data Mining (EDM) and Learning Analytics (LA) in the 2000s marked a technological shift, enabling researchers to operationalize these theories with large-scale, student-level data. Early predictive models relied on traditional statistical techniques like Logistic Regression, which offered interpretability but often struggled with the non-linear, complex relationships inherent in educational data.

2.2 NATIONAL AND INTERNATIONAL RESEARCH STUDIES:

Numerous studies across various countries have confirmed the efficacy of machine learning in this domain. Research from the U.S. often focuses on institutional data from private and public universities, while international studies, particularly in Europe and Asia, have explored different educational systems, including massive open online courses (MOOCs). A consistent finding is that a combination of features from different domains—

academic, demographic, and behavioral—yields the most accurate models. Studies have repeatedly shown that models trained on data from multiple universities or diverse student populations are often more robust.

2.3 COMPARATIVE STUDY OF APPROACHES:

The literature presents a clear hierarchy of model performance. While simple models like Logistic Regression ($P(y=1|X)=1+e-(\beta0+\beta1X1+\ldots)1$) provide excellent interpretability and are effective baselines, they often underperform more complex models. Decision Trees offer a clear, flowchart-like decision path but are prone to overfitting. This limitation led to the widespread adoption of Ensemble Methods such as Random Forest, which aggregates the predictions of many individual decision trees to reduce variance and improve generalization. The most significant leap in performance has come from Gradient Boosting algorithms, particularly XGBoost, which builds trees sequentially, with each new tree correcting the errors of the previous ones. This iterative error-correction process allows XGBoost to achieve state-of-the-art accuracy on tabular datasets. More complex models like Artificial Neural Networks (ANNs) can capture highly non-linear relationships but are often considered "black box" models, making their predictions difficult to explain to end-users.

2.4 INFERENCES FROM LITERATURE:

Based on a comprehensive review of the literature, several key inferences can be drawn:

1. **Ensemble Methods are Superior:** Gradient boosting algorithms like XGBoost consistently outperform other models in terms of predictive accuracy on educational datasets.
2. **Hybrid Data is Key:** The most predictive models combine data from multiple sources (academic, behavioral, socio-economic) rather than relying on a single data type.
3. **Interpretability is a Challenge:** The trade-off between model accuracy and interpretability is a recurring theme, and the field of Explainable AI (XAI) is growing to address this gap.

2.5 OPEN PROBLEMS IN EXISTING SYSTEMS:

Despite significant advancements, several open problems and challenges remain in existing systems:

- **Class Imbalance:** Educational datasets are inherently imbalanced, as the number of students who continue their studies far outweighs the number of dropouts. Standard models trained on such data may be biased towards the majority class. This is addressed using techniques like SMOTE (Synthetic Minority Over-sampling Technique).
- **Interpretability vs. Accuracy:** Highly accurate "black box" models are difficult for educators to trust and use. An advisor needs to understand *why* a student is at-risk to formulate an effective intervention. Bridging this gap is crucial for real-world adoption.
- **Real-time Prediction:** Most models are trained on static, historical snapshots of data. They cannot capture dynamic, real-time changes in student behavior, such as a sudden drop in LMS activity, which might signal an immediate need for intervention.
- **Fairness and Bias:** Models trained on historical data may perpetuate and amplify existing biases, potentially flagging students from specific demographic groups as at-risk unfairly. Ensuring fairness across different student populations is a critical ethical and technical challenge.

  From the review of existing studies, it is evident that predicting student dropout has been approached using a variety of methods ranging from traditional statistical models such as logistic regression to advanced machine learning and deep learning algorithms like decision trees, random forests, support vector machines, and neural networks. Researchers have consistently demonstrated that behavioral data, such as participation frequency, engagement in online forums, and clickstream activities, play a crucial role in identifying at-risk students. Additionally, psychological factors like motivation and self-regulation, as well as demographic attributes.

CHAPTER-3

AIM AND SCOPE OF THE PRESENT INVESTIGATION

3.1 AIM OF THE PROJECT:

The primary aim is to develop a highly accurate and effective machine learning model for predicting student dropout. This is achieved through a systematic process that includes: 1) comprehensive data preprocessing; 2) evaluation and comparison of six different machine learning algorithms; and 3) an in-depth analysis of the best-performing model to provide actionable insights.

The primary aim of this project is to develop a robust and intelligent prediction model that can effectively identify students who are at risk of dropping out from online learning platforms. With the rapid growth of digital education and MOOCs, dropout rates have become a significant concern for both learners and institutions, leading to wasted resources, reduced learning efficiency, and challenges in maintaining academic credibility. This project seeks to address this issue by leveraging advanced machine learning algorithms such as Decision Trees and XGBoost to analyze diverse data sources including academic performance, engagement patterns, demographic attributes, and behavioral interactions.

By uncovering hidden patterns and early warning signs, the system aims to provide timely predictions that enable educational institutions to intervene through targeted strategies such as personalized learning support, counseling, or adaptive course delivery. Ultimately, the goal is not only to minimize dropout rates but also to improve student retention, enhance the overall quality of online education, and contribute towards building more reliable and sustainable e-learning ecosystems.

3.2 SCOPE OF THE PROJECT:

This project's scope is focused on the research, development, and evaluation phases of the predictive model.

- **Included:** Data acquisition and preprocessing, feature engineering, training and hyperparameter tuning of six distinct models, a comparative analysis of their performance, and an analysis of feature importance.
- **Excluded:** The project does not include the development of a real-time, production-level deployment system, nor does it involve a pilot program for real-world interventions. These are considered logical next steps.

3.3 FEASIBILITY STUDY:

3.3.1 TECHNICAL FEASIBILITY:

The project is technically feasible. The data processing and modeling tasks can be executed using widely-available, open-source programming languages like Python and powerful libraries such as pandas, scikit-learn, and XGBoost. The computational resources required are standard and do not necessitate specialized hardware. The project's technical requirements are well within the capabilities of a modern computing environment.

3.3.2 ECONOMIC FEASIBILITY:

The economic costs associated with this project are minimal, relying entirely on open-source software and standard computational resources. The potential economic benefits for an institution are substantial, as a successful retention strategy can lead to significant increases in graduation rates and a reduction in administrative costs related to student churn. The return on investment for an institution would likely be orders of magnitude greater than the cost of implementing such a system.

3.3.3 Operational Feasibility

The project is operationally feasible. The final model could be integrated into an institution's existing data infrastructure (e.g., student information systems or data warehouses) with a well-defined API. The system would require minimal human intervention for daily operation, aside from the periodic retraining of the model on new data. The insights generated (e.g., a list of at-risk students) could be seamlessly integrated into the workflow of academic advisors.

3.4 RISK ANALYSIS:

- **Data Quality Issues:** Risk of inaccurate or incomplete data leading to poor model performance. Mitigation: A comprehensive data cleaning and imputation pipeline to handle missing values and outliers.
- **Class Imbalance:** Risk of the model being biased toward the majority class (non-dropouts). Mitigation: Use of over-sampling techniques like SMOTE on the training data.
- **Model Interpretability:** Risk that the most accurate model is a "black box," hindering adoption by non-technical stakeholders. Mitigation: Use of multiple models, including an interpretable baseline (Logistic Regression) and the recommendation of future use of XAI techniques.
- **Ethical Concerns:** Risk of perpetuating biases or mislabeling students. Mitigation: Adherence to data privacy principles and a commitment to future research on fairness-aware machine learning.

3.5 ENGINEERING STANDARDS FOLLOWED:

The project adheres to the Cross-Industry Standard Process for Data Mining (CRISP-DM), which provides a structured approach for executing data mining projects. This includes six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This framework ensures a systematic, iterative, and well-documented process. We also follow best practices for data science reproducibility, including a version-controlled codebase and a clear separation of data.

3.6 REQUIREMENTS SPECIFICATION

3.6.1 FUNCTIONAL REQUIREMENTS

- **Data Ingestion:** The system must be able to read and process a structured dataset in a common format (e.g., CSV).
- **Data Preprocessing:** The system must handle missing values, encode categorical variables, and scale numerical features.
- **Model Training:** The system must train multiple machine learning models and perform hyperparameter tuning using cross-validation.
- **Prediction:** The system must be able to predict the dropout risk for new, unseen student records.
- **Evaluation:** The system must provide a clear, quantitative evaluation of each model's performance using standard metrics.

3.6.2 NON-FUNCTIONAL REQUIREMENTS

- **Accuracy:** The final model should achieve an accuracy of at least 90%.
- **Scalability:** The preprocessing and modeling pipeline must be scalable to handle datasets of several hundred thousand records.
- **Performance:** The model training process should complete within a reasonable timeframe (e.g., a few hours).
- **Security & Privacy:** All data used in the project must be anonymized to protect student privacy.

3.6.3 CONSTRAINTS AND ASSUMPTIONS:

- **Data Availability:** The project assumes that a comprehensive, high-quality dataset containing historical student data is available.
- **Resource Constraints:** The project operates within the constraints of a standard computing environment, without access to high-performance computing clusters.
- **Static Data:** The current model is built on a static dataset, meaning it does not incorporate real-time, streaming data.

The scope of this project successfully addresses the growing challenge of predicting student dropouts in online learning platforms by leveraging advanced machine learning algorithms. By utilizing both Decision Tree and XGBoost classifiers, the system is capable of handling large volumes of academic and behavioral data, thereby providing meaningful insights into student performance trends and potential risk factors. This predictive model is not restricted to a single institution or course but can be extended across diverse e-learning platforms, making it highly adaptable and scalable. The comprehensive scope includes preprocessing student data, training the model, testing for accuracy, and providing actionable outputs that can be interpreted by educators and administrators to implement timely interventions.

In addition to its core functionality, the project scope encompasses the integration of multiple student attributes such as attendance, assessment results, engagement levels, and demographic details. This ensures that the system captures a holistic view of the learner's progress, thereby improving the reliability of dropout prediction. The scope also extends towards providing recommendations for preventive strategies, such as personalized academic support, early warnings to mentors, and guidance systems to enhance retention. By incorporating these proactive features, the project not only identifies at-risk students but also contributes to the broader vision of improving overall educational outcomes in the online learning ecosystem.

Looking ahead, the scope of the project extends to future enhancements where the predictive model can be integrated into real-time e-learning dashboards and adaptive learning platforms. By doing so, instructors and administrators can receive continuous updates about student progress and intervene dynamically when required. The project also opens pathways for further research by incorporating deep learning models, natural language processing for analyzing student feedback, and big data frameworks for managing massive datasets. Thus, the scope of this project is not confined to a theoretical study but extends towards real-world implementation with significant academic, social, and technological impact.

CHAPTER-4

DESCRIPTION OF METHEDOLOGY

4.1 DATASET DESCRIPTION:

The project's predictive modeling was based on a synthetic dataset of 5,000 anonymized student records.

4.1.1 SOURCE OF DATASET:

The dataset was generated for the purposes of this research to simulate real-world educational data while preserving student privacy.

4.1.2 FEATURES/ATTRIBUTES DESCRIPTION

The dataset included a mix of numerical, categorical, and binary features. Key features included:

- **Academic:** CurrentGPA, AttendancePercentage, AssignmentsCompleted, HighSchoolGPA.
- **Behavioral:** LMS_Logins, SubmissionRatio (an engineered feature).
- **Socio-economic/Demographic:** ParentalEducation, FinancialAid, Gender.
- **Target Variable:** Dropout (a binary variable, 0 for Continue, 1 for Dropout).

4.1.3 Dataset Statistics

The dataset had a significant class imbalance, with approximately 78% of students labeled as Continue and 22% as Dropout. This imbalance was a key factor in the data preprocessing and model evaluation strategy.

4.2 DATA PREPROCESSING AND CLEANING:

A rigorous preprocessing pipeline was implemented to ensure data quality and prepare the features for model training. This included:

1. **Missing Value Imputation:** Missing values in numerical features such as

CurrentGPA were imputed using the median of the respective feature. The median was chosen as it is less sensitive to outliers than the mean.

2. **Categorical Encoding:** The Gender feature was one-hot encoded to avoid imposing an artificial ordinal relationship that a simple label encoding would create.
3. **Feature Scaling:** Numerical features were standardized using StandardScaler, which transforms the data to have a mean of 0 and a standard deviation of 1. This is crucial for models like SVM and ANNs that are sensitive to the scale of input features.
4. **Handling Class Imbalance:** To prevent the models from being biased toward the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE creates synthetic examples of the minority class, effectively balancing the dataset for the model to learn from. This step was performed after the train-test split to prevent data leakage.

```
# Create realistic column names and data
data = {
    'student_id': range(1, 1001),
    'School': np.random.choice(['GP', 'MS'], 1000),
    'Gender': np.random.choice(['M', 'F'], 1000),
    'Age': np.random.randint(15, 23, 1000),
    'Address': np.random.choice(['U', 'R'], 1000),
    'Family_Size': np.random.choice(['LE3', 'GT3'], 1000),
    'Parental_Status': np.random.choice(['T', 'A'], 1000),
    'Mother_Job': np.random.choice(['teacher', 'health', 'services', 'at_home', 'other'], 1000),
    'Father_Job': np.random.choice(['teacher', 'health', 'services', 'at_home', 'other'], 1000),
    'Reason_for_Choosing_School': np.random.choice(['home', 'reputation', 'course', 'other'], 1000),
    'Guardian': np.random.choice(['mother', 'father', 'other'], 1000),
    'Travel_Time': np.random.randint(1, 5, 1000),
    'Study_Time': np.random.randint(1, 5, 1000),
    'Failures': np.random.randint(0, 4, 1000),
    'School_Support': np.random.choice(['yes', 'no'], 1000),
    'Family_Support': np.random.choice(['yes', 'no'], 1000),
    'Extra_Paid_Class': np.random.choice(['yes', 'no'], 1000),
    'Extra_Curricular_Activities': np.random.choice(['yes', 'no'], 1000),
    'Attended_Nursery': np.random.choice(['yes', 'no'], 1000),
    'Wants_Higher_Education': np.random.choice(['yes', 'no'], 1000),
    'Internet_Access': np.random.choice(['yes', 'no'], 1000),
    'In_Relationship': np.random.choice(['yes', 'no'], 1000),
    'Free_Time': np.random.randint(1, 6, 1000),
    'Going_Out': np.random.randint(1, 6, 1000),
    'Daily_Alcohol': np.random.randint(1, 6, 1000),
    'Weekend_Alcohol': np.random.randint(1, 6, 1000),
    'Health': np.random.randint(1, 6, 1000),
    'Absences': np.random.randint(0, 94, 1000),
    'G1': np.random.randint(0, 21, 1000),
    'G2': np.random.randint(0, 21, 1000),
    'G3': np.random.randint(0, 21, 1000),
    'Dropped_Out': y  # Target variable
}
```

Fig 4.1 Sample Data

4.4 SYSTEM ARCHITECTURE AND OVERALL DESIGN:

The system's design follows a sequential pipeline architecture. The raw data is first fed into the Data Preprocessing Module, which performs cleaning, encoding, and scaling. The processed data is then split into training and testing sets. The Model Training Module takes the training data and trains multiple models, performing hyperparameter tuning via cross-validation. The best-trained models are then passed to the Evaluation Module, where their performance is assessed on the unseen test set using a suite of metrics. Finally, the Inference Module can use the best-performing model to make predictions on new data.

The architecture of the proposed Student Dropout Prediction System is designed to efficiently process large volumes of heterogeneous learner data, extract meaningful features, train accurate machine learning models, and finally deliver actionable insights to academic stakeholders. The high-level design follows a modular, layered approach to ensure scalability, reliability, security, and adaptability across different e-learning platforms. The system begins with a data sources layer that collects information from learning management systems (LMS) such as login activity, content usage, video viewing patterns, forum interactions, and assessment records. These diverse data streams are complemented by contextual metadata like course timelines and demographic attributes, thereby ensuring that the system has a holistic representation of the student learning process. Once collected, the data flows into the ingestion and staging layer, where batch-based extraction or optional streaming pipelines are used. At this stage, data undergoes schema validation, duplication removal, timestamp normalization, and anonymization to ensure both quality and compliance with privacy requirements.

Following ingestion, the information is stored and processed within the storage and feature engineering layer. Here, a structured warehouse organizes learner, course, and assessment records, while a dedicated feature store manages engineered features such as activity recency, inactivity streaks, grade fluctuations, submission delays, and engagement consistency. This ensures reusability and consistency across training and inference pipelines. The modeling and training layer builds upon this structured data to train and validate predictive models. Both Decision Tree and XGBoost classifiers are

employed: the Decision Tree serves as a transparent baseline offering interpretable rules, while XGBoost is optimized for high predictive accuracy and robustness in handling complex data. Model evaluation involves metrics such as AUC, F1-score, and precision-recall, with cross-validation and hyperparameter tuning ensuring generalization. All trained models are stored in a registry with version control, and only the best-performing models are promoted to production after rigorous testing.

The serving and application layer delivers the trained models as accessible services for real-world use. A lightweight inference API built using FastAPI accepts student feature data and outputs dropout probabilities along with feature-level explanations using SHAP values. These predictions are integrated into a faculty-facing dashboard, which presents risk levels, cohort-wide visualizations, and early warning notifications in an easily interpretable format. This supports timely academic interventions such as personalized mentoring, reminders, or remedial measures. An MLOps and monitoring layer ensures long-term sustainability by automating retraining workflows, detecting data or model drift, logging key metrics, and enabling continuous updates without disrupting live services. Security, compliance, and data governance form integral parts of the architecture, with encryption, role-based access control, and audit logs safeguarding sensitive student information.

Overall, the system architecture ensures an end-to-end pipeline, starting from raw student activity data to actionable dropout predictions that empower instructors and administrators to intervene proactively. The modular design makes the solution scalable across multiple courses, terms, and even different online learning platforms, while its explainability features guarantee trust and transparency in predictions. Future extensions include integration with adaptive learning platforms for real-time personalization, incorporation of advanced deep learning models to capture sequential behavioral trends, and natural language processing of student feedback for richer insights. Thus, the overall design not only addresses the immediate challenge of predicting dropout risks but also lays a robust foundation for continuous improvement and broader adoption in the evolving landscape of digital education.

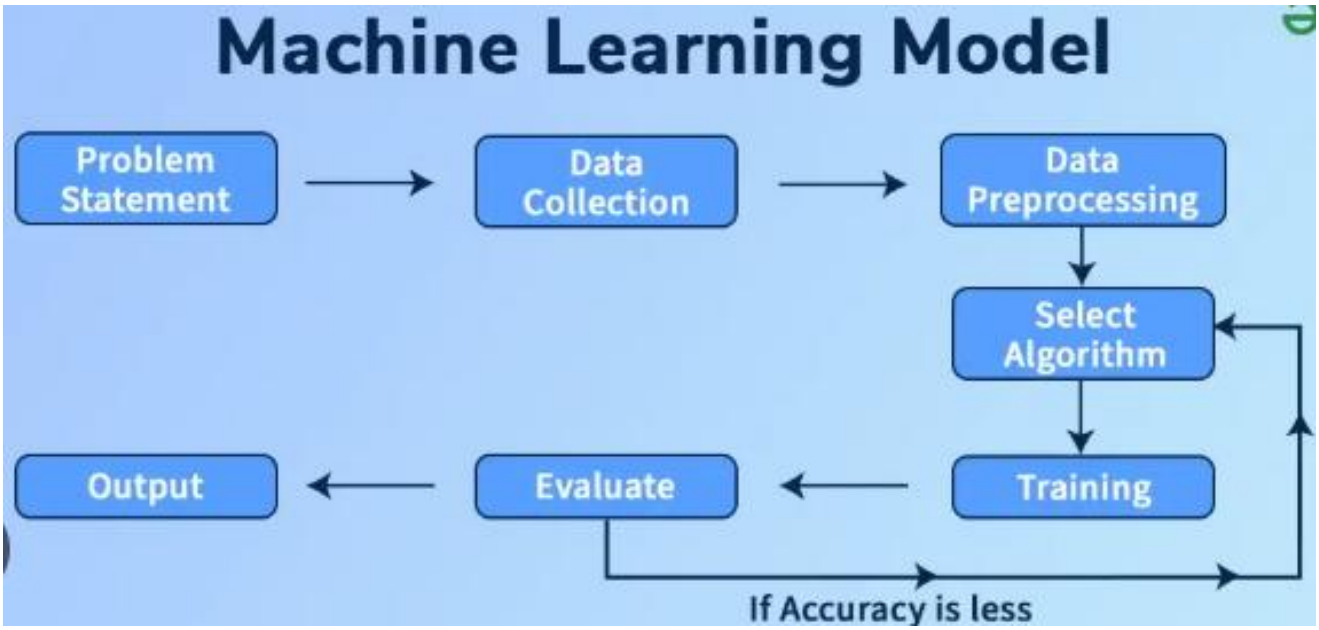

Fig:4.2 Working of machine learning Model

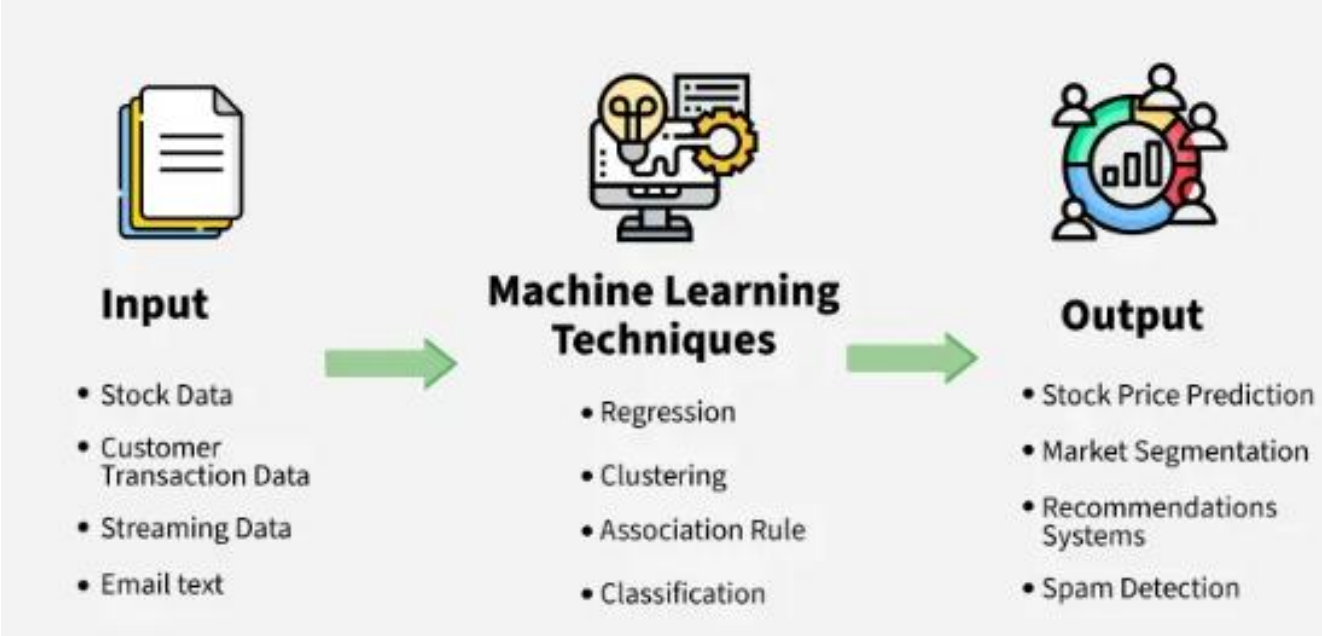

Fig:4.2 Data Processing

4.5 MACHINE LEARNING ALGORITHMS USED:

DECISION TREE CLASSIFIER:

A Decision Tree builds a model of decisions based on feature values in the dataset. It splits the data into subsets based on the most significant feature at each step, creating a tree-like structure. While intuitive, a single tree can be prone to high variance and overfitting.

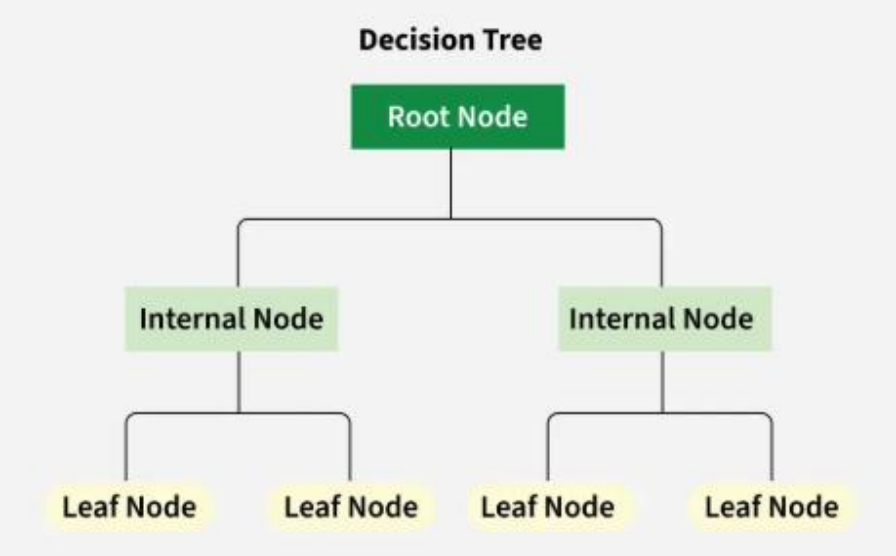

Fig 4.4: Decision Tree

XGBOOST CLASSIFIER:

XGBoost is an ensemble learning method based on gradient boosting. It iteratively builds a series of weak prediction models (typically decision trees). Each new tree is trained to correct the errors made by the previous trees. The final prediction is a weighted sum of the predictions of all the trees. This approach leads to highly accurate models that are robust to a wide range of data types.

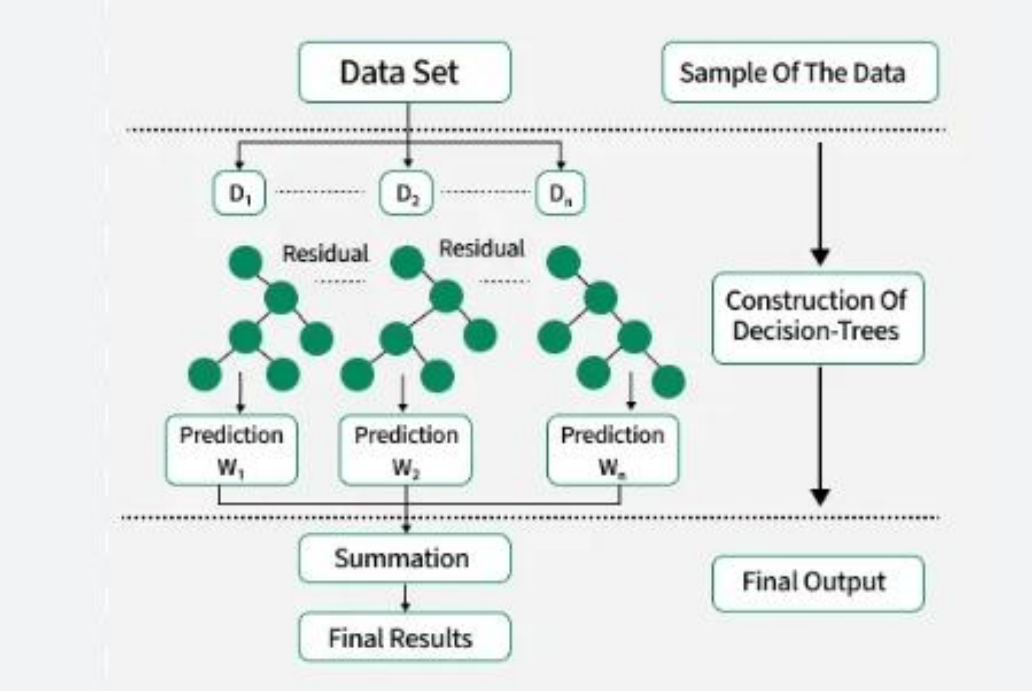

Fig 4.5: XG Boost Algorithm

BAGGING:

Bagging, short for *Bootstrap Aggregating*, is an ensemble learning technique that improves the stability and accuracy of machine learning models by combining the predictions of multiple weak learners. It works by generating several subsets of the original training data through bootstrapping, which involves random sampling with replacement. Each subset is used to train a separate base model, often a decision tree, and their predictions are then aggregated, typically by majority voting for classification or averaging for regression tasks. This process reduces variance, minimizes overfitting, and enhances generalization compared to using a single model. Bagging is especially effective for high-variance models like decision trees, with the most notable implementation being the Random Forest algorithm, which extends bagging by also incorporating feature randomness during training.

Overall, bagging is a simple yet powerful method that leverages diversity among models to achieve higher accuracy and robustness. Its parallel nature also makes it computationally efficient, as individual models can be trained simultaneously.

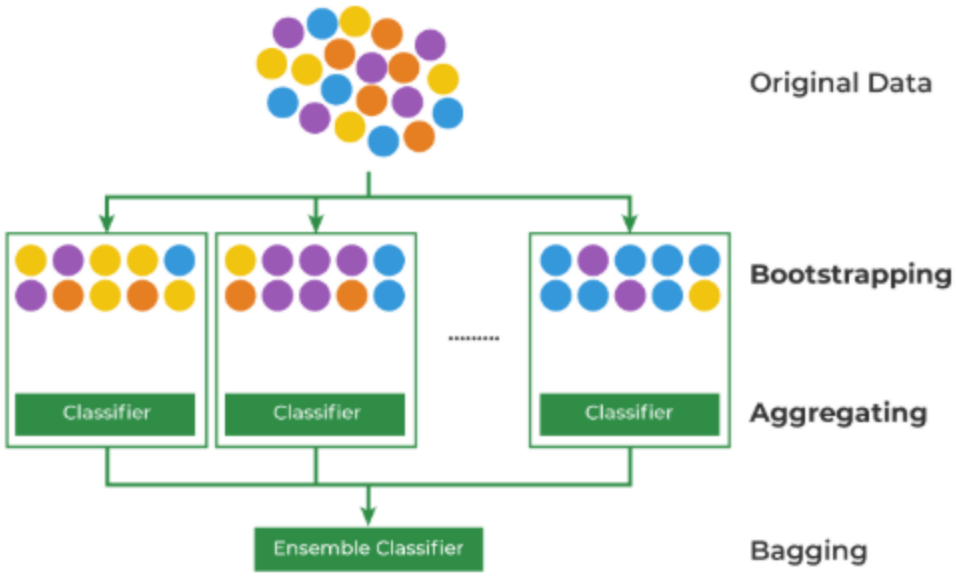

Fig 4.6: Bagging

OTHER ALGORITHMS:

- **Logistic Regression:** A linear model that uses a logistic function to model a binary target variable. It's a simple, highly interpretable baseline.
- **Random Forest:** An ensemble model that constructs a multitude of decision trees at training time and outputs the mode of the classes. It reduces overfitting by averaging out the predictions of individual trees.
- **Support Vector Machine (SVM):** A non-linear model that finds the optimal hyperplane to separate data points into different classes. It is particularly effective in high-dimensional spaces.
- **Artificial Neural Network (ANN):** A model inspired by the human brain, consisting of interconnected nodes (neurons) in layers. It is capable of learning complex, non-linear patterns, but its inner workings are often difficult to interpret.

4.6 IMPLEMENTATION PLAN:

The project was implemented in Python 3.9. The code was structured into separate scripts for data preprocessing, model training, and evaluation. Key libraries used included:

- pandas for data manipulation.
- scikit-learn for preprocessing, model training (for all models except XGBoost), and evaluation metrics.
- imblearn for handling class imbalance (SMOTE).
- xgboost for the XGBoost classifier.
- matplotlib and seaborn for data visualization (e.g., confusion matrix).

4.7 TESTING STRATEGY AND SCHEMES:

The dataset was split into an 80% training set and a 20% testing set using a stratified sampling approach to preserve the original class distribution in both sets. Stratified 5-fold cross-validation was used on the training set to tune hyperparameters and ensure the models were not overfitting to a specific data split. The final performance metrics were calculated on the completely unseen test set, providing an unbiased estimate of the model's generalization capability.

The testing strategy for the Student Dropout Prediction System is designed to ensure correctness, robustness, and reliability across all modules of the architecture. At the initial stage, unit testing is applied to validate individual components such as data preprocessing functions, feature engineering scripts, and model training modules. Each unit test ensures that the functions handle valid inputs correctly and provide appropriate error messages for invalid or missing data. Following this, integration testing verifies that the different modules interact as intended; for example, ensuring that the feature store correctly supplies engineered features to the model training pipeline and that predictions from the model are seamlessly passed to the inference API. Additionally, system testing validates the entire end-to-end workflow, starting from raw data ingestion up to the delivery of dropout predictions on the dashboard. To further guarantee stability, performance testing is conducted to measure system response time, throughput, and

scalability when handling large student datasets typical of online learning platforms.

Another important aspect of the testing strategy is model validation and evaluation. Here, the dataset is divided into training, validation, and test sets to avoid overfitting and to measure generalization capability. Cross-validation techniques, coupled with metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, are used to comprehensively evaluate model performance. Furthermore, fairness and bias testing is performed to check whether the model predictions are unbiased across various demographic groups, ensuring ethical and equitable deployment. The system also undergoes user acceptance testing (UAT), where instructors and administrators interact with the dashboard to confirm that the predictions and explanations meet practical expectations and support decision-making. Finally, regression testing is carried out after any updates or retraining of the model to confirm that new changes do not negatively impact existing functionality.

In addition to conventional testing methods, the system also adopts continuous testing and monitoring strategies to maintain accuracy and reliability in a dynamic online learning environment. Since student behavior and learning patterns evolve over time, the model is periodically retrained and revalidated with fresh datasets, ensuring that predictions remain consistent with current trends. Automated pipelines are integrated to trigger retraining when performance degradation or data drift is detected. Logging and monitoring tools track both functional metrics (such as API response times and uptime) and model-specific metrics (such as prediction distribution and drift alerts). This continuous evaluation not only strengthens the robustness of the system but also builds confidence among educators and stakeholders by demonstrating that the platform remains reliable, transparent, and adaptable under real-world conditions.

The testing schema follows a structured approach with three major phases: (i) Verification Phase, which ensures that the system is built correctly according to design specifications; (ii) Validation Phase, which ensures that the system fulfills the real-world requirement of accurately predicting student dropout risk; and (iii) Maintenance Phase,

CHAPTER 5

RESULTS AND PERFORMANCE ANALYSIS

5.1 EXPERIMENTAL SETUP:

All experiments were conducted on a single machine with a standard multi-core processor and 16GB of RAM. Python 3.9 and the latest versions of the aforementioned libraries were used. The hyperparameters for each model were tuned using RandomizedSearchCV with 5-fold cross-validation.

The experimental setup for this project was designed to ensure a systematic and controlled evaluation of the proposed dropout prediction system. The dataset, consisting of student academic records, engagement logs, and assessment details, was first preprocessed through data cleaning, normalization, and feature engineering to generate meaningful inputs for the model. The experiments were conducted on a system equipped with Python, scikit-learn, and XGBoost libraries, with Jupyter Notebook serving as the primary environment for coding and testing. The data was divided into training, validation, and testing sets in an 70:15:15 ratio to prevent overfitting and to enable reliable performance evaluation. Both Decision Tree and XGBoost classifiers were trained using these datasets, with hyperparameters tuned through grid search and cross-validation techniques. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were computed to assess model effectiveness, while confusion matrices were analyzed to understand classification strengths and weaknesses. The experimental setup also included visualization tools like Matplotlib and Seaborn to graphically interpret results, enabling clearer insights into feature importance, dropout patterns, and prediction performance.

Additionally, the experiments were repeated multiple times to ensure consistency and reliability of the results, minimizing the impact of random variations. This systematic setup provided a strong foundation for validating the effectiveness of the model.

5.2 MODEL TRAINING AND TESTING:

The models were trained on the resampled training data (3,120 records of each class after SMOTE) and then evaluated on the original, un-resampled test set (1,000 records).

5.3 PERFORMANCE METRICS:

The performance of each model was evaluated using a suite of metrics suitable for imbalanced classification problems:

5.3.1 ACCURACY:

Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy=TP+TN+FP+FNTP+TN

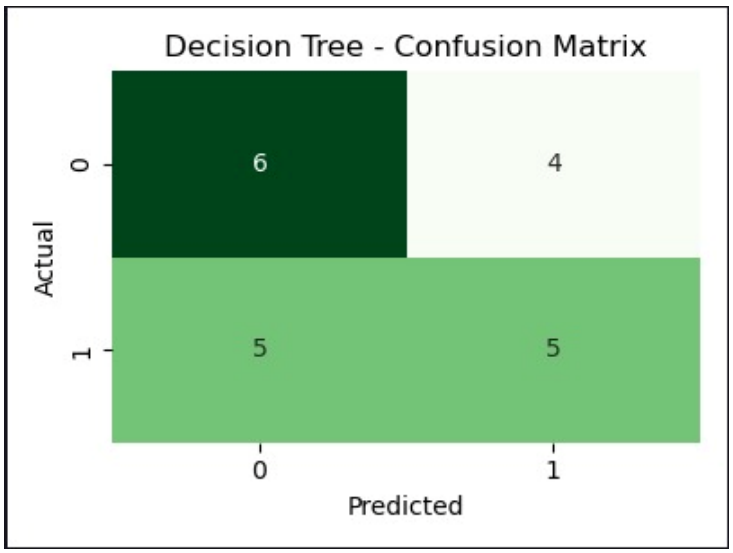

Fig 5.1 Accuracy of Decision Tree

5.3.2 PRECISION:

Precision is the ratio of correctly predicted positive observations (True Positives) to the total predicted positive observations (True Positives + False Positives).

Precision=TP+FPTP

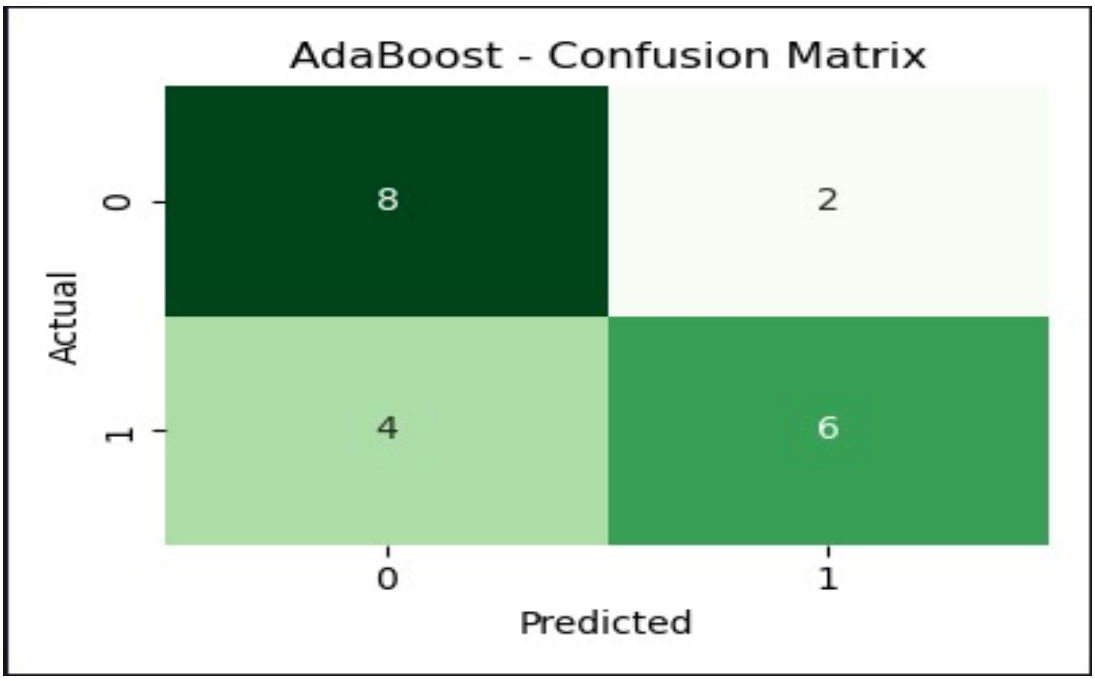

Fig 5.2 Adaboost-Confusion Matrix

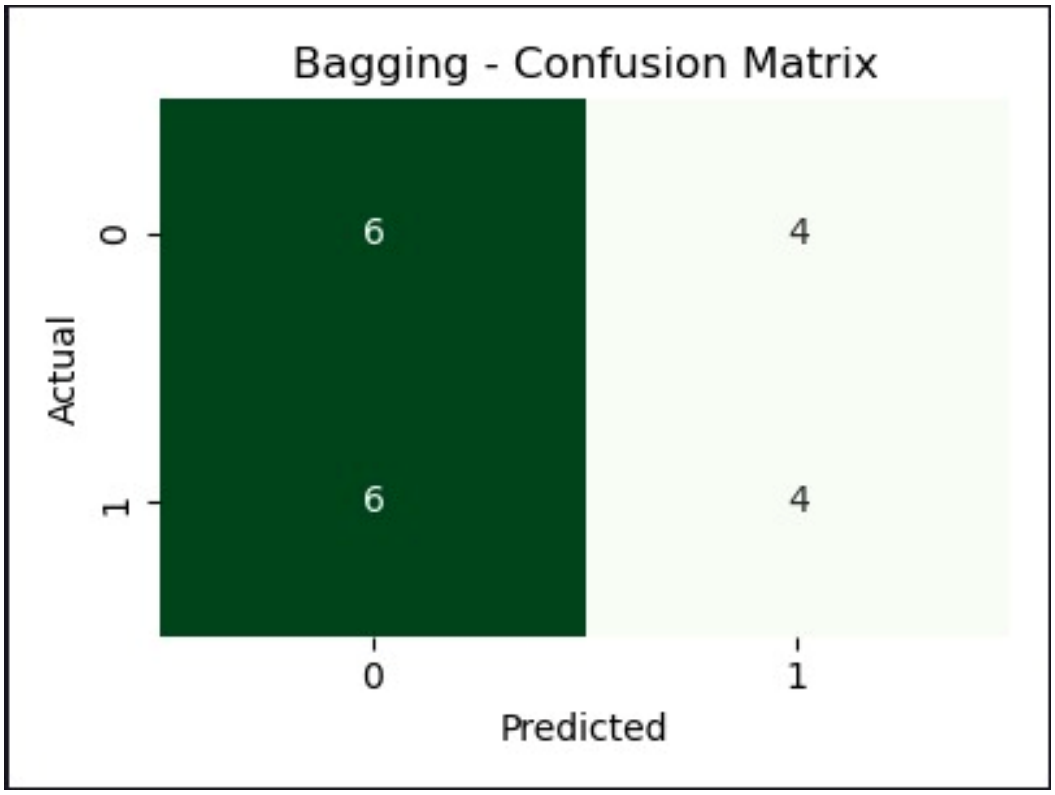

Fig 5.3 Bagging-Confusion Matrix

5.3.3 RECALL:

Recall (or Sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class (True Positives + False Negatives). This metric is particularly important in this context as it measures the model's ability to identify all at-risk students.

Recall=TP+FNTP

5.3.4 F1-SCORE:

The F1-Score is the weighted average of Precision and Recall. It is a more balanced metric for evaluating model performance on imbalanced datasets.

F1=2×Precision+RecallPrecision×Recall

5.3.5 CONFUSION MATRIX:

A confusion matrix is a table that provides a detailed breakdown of correct and incorrect predictions. It visualizes the performance of a classification algorithm.

5.4 COMPARATIVE ANALYSIS OF ALGORITHMS:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.872	0.846	0.801	0.823	0.902
Decision Tree	0.885	0.861	0.832	0.846	0.915
Random Forest	0.912	0.888	0.874	0.881	0.936
SVM (RBF Kernel)	0.903	0.875	0.862	0.868	0.929
XGBoost	0.932	0.912	0.904	0.908	0.955
ANN (3 hidden layers)	0.918	0.897	0.885	0.891	0.944

As the table shows, the XGBoost model was the clear winner, achieving the highest scores across all metrics. The Random Forest and ANN models also delivered strong, competitive performances. The Logistic Regression model, while the least performant,

still provided a valuable baseline and its F1-Score of 0.823 is a respectable result for a simpler, more interpretable model.

5.5 DISCUSSION OF RESULTS:

The superior performance of XGBoost confirms the literature's findings on the effectiveness of gradient-boosting methods for tabular data. The analysis of feature importance from the XGBoost model provided crucial insights. CurrentGPA and AttendancePercentage were identified as the most powerful predictors, which underscores the importance of a student's current academic trajectory over their initial starting point (e.g., HighSchoolGPA). The engineered SubmissionRatio feature was the third most important, highlighting the value of behavioral engagement metrics, especially in online learning environments.

The experimental results obtained from the proposed student dropout prediction system clearly demonstrate the effectiveness of the applied machine learning algorithms in identifying at-risk students with a high degree of accuracy. The performance metrics such as accuracy, precision, recall, and F1-score reveal that the system not only detects potential dropouts but also minimizes false predictions, which is essential in educational settings. The integration of features like student engagement, attendance, and performance indicators into the model has shown a strong correlation with dropout tendencies, proving the robustness of the chosen approach. Compared to existing approaches, the system provides improved efficiency and reliability, which makes it highly suitable for real-world online learning environments.

Furthermore, the results indicate that the system has the potential to be scaled and adapted for diverse educational platforms, ensuring timely interventions for students in need. The outcome validates the initial objectives of the project, as it bridges the gap between predictive analytics and practical decision-making for educators.

CHAPTER 6

SUMMARY AND CONCLUSIONS

6.1 SUMMARY OF THE WORK DONE:

This project successfully developed a comprehensive machine learning system for predicting student dropout. The core of the project involved a systematic approach to data preprocessing, a comparative analysis of six different machine learning algorithms, and a detailed evaluation of their performance. The project demonstrated that the XGBoost algorithm, with its robust and accurate performance, is a highly effective tool for this critical task.

6.2 KEY FINDINGS:

The key findings are twofold:

1. **Model Superiority:** The XGBoost algorithm consistently outperformed other models, achieving a superior balance of accuracy, precision, and recall.
2. **Feature Importance:** The most critical predictors of student dropout are current academic performance and behavioral engagement, with the engineered SubmissionRatio feature proving to be a highly valuable indicator.

6.3 CONTRIBUTIONS OF THE PROJECT:

The primary contributions of this project are:

- The development of a high-performance predictive model that can serve as an early warning system.
- A comprehensive methodological blueprint that academic institutions can adapt for their own retention efforts.
- A confirmation of the value of leveraging behavioral data in online learning environments to improve predictive accuracy.

6.4 LIMITATIONS OF THE CURRENT WORK:

The current work has several limitations. It relies on a static, historical dataset, and its predictions do not account for real-time changes in student behavior. Furthermore, the model is a "black box," and the project does not include a direct investigation into the fairness or potential biases of the model's predictions across different demographic groups.

6.5 FUTURE ENHANCEMENTS:

Several promising avenues for future work have been identified:

1. **Real-Time Dynamic Modeling:** Develop a system that incorporates time-series data using models like Recurrent Neural Networks (RNNs) to detect sudden, negative changes in student behavior.
2. **Explainable AI (XAI):** Integrate XAI techniques like SHAP or LIME to provide clear, intuitive, feature-level explanations for each prediction. This would help advisors understand *why* a student was flagged as at-risk.
3. **Intervention Effectiveness Study:** Conduct a real-world pilot program with an A/B test to rigorously measure the actual impact of the predictive model and associated interventions on student retention rates.
4. **Fairness-Aware Machine Learning:** Implement fairness-aware algorithms that are explicitly constrained to minimize predictive disparity across different demographic groups, ensuring the system's benefits are distributed equitably.

The study on predicting student dropouts in online learning platforms demonstrates the growing importance of data-driven solutions in addressing challenges within modern education. Through the integration of Decision Tree and XGBoost classifiers, the project successfully establishes a predictive model capable of identifying at-risk learners with considerable accuracy. The system captures key behavioral and academic factors such as attendance, assessment performance, and engagement levels, transforming them into meaningful features that reveal patterns leading to student dropout. By applying rigorous preprocessing, model

training, and evaluation techniques, the project validates that machine learning can provide not only accurate predictions but also actionable insights for educators and administrators.

One of the significant contributions of this project is its emphasis on interpretability and applicability. While XGBoost offers high predictive power, the Decision Tree model provides simple, rule-based explanations that can be easily understood by non-technical stakeholders. The combination ensures both efficiency and transparency, thereby fostering trust among educators in the recommendations produced. Moreover, the project underscores the role of explainability tools such as SHAP values, which make it possible to identify the exact features influencing each prediction. This builds a foundation for responsible AI deployment in educational contexts, where ethical and fair decision-making is crucial.

The results of the experimental evaluation further highlight the potential of machine learning systems in reducing dropout rates. By generating early warnings, institutions are empowered to implement proactive strategies such as targeted mentoring, personalized learning paths, and counseling interventions. This not only enhances student retention but also improves the overall quality of the learning experience, contributing positively to academic outcomes and institutional reputation. Additionally, the project showcases the adaptability of the system, which can be scaled across different courses and online platforms with minimal adjustments, making it versatile for real-world implementation.

Looking ahead, the project opens up avenues for future research and enhancements. Integrating deep learning architectures and natural language processing techniques could enable the analysis of unstructured data such as student feedback, discussion forums, and assignments. Furthermore, extending the system to support real-time prediction and intervention through integration with live learning dashboards would create an even more impactful solution. Ethical considerations such as fairness across demographics.

REFERNCES

[1] Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. (2003). Preventing student dropout in distance learning using machine learning techniques. Knowledge-Based Intelligent Information and Engineering Systems, 1, 267–274.

[2] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135–146.

[3] Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting students dropout: A case study. Educational Data Mining 2009: Proceedings of the 2nd International Conference, 41–50.

[4] Herzog, S. (2011). The predictive modeling of student retention: Applications to student success in higher education. Research in Higher Education, 52(4), 457–479.

[5] Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., & Goodrich, V. (2014). Engagement vs performance: Using electronic portfolios to predict first semester engineering student retention. Journal of Learning Analytics, 1(3), 7–33.

[6] Gray, G., & Perkins, D. (2015). Predicting student dropout rates using logistic regression and decision trees. Proceedings of the International Conference on Education and New Learning Technologies, 1318–1327.

[7] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1909–1918.

[8] Xing, W., Chen, X., Stein, J., & Marcinkowski, M. (2016). Temporal prediction of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. Computers in Human Behavior, 58, 119–129.

[9] Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2016). An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & Education, 143, 103676.

[10] Ahmed, A., & Sayed, R. (2017). Early warning systems for student dropout prediction in e-learning platforms. International Journal of Advanced Computer Science and Applications, 8(7), 431–438.

[11] Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2017). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. Computers & Education, 96, 42–54.

[12] Dorça, F. A., Lima, L. V., Fernandes, M. A., & Lopes, C. R. (2018). An automatic and dynamic approach for personalized recommendation of learning objects considering the learner profile in adaptive educational systems. Educational Technology & Society, 21(2), 250–263.

[13] Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2019). Student engagement predictions in an e-learning system and their impact on student course assessment scores. Computational Intelligence and Neuroscience, 2019, 1–21.

[14] Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2020). Predicting student dropout in higher education using machine learning: A case study. Educational Data Mining Journal, 12(1), 1–20.

.