STUDENT DROPOUT PREDICTION

[1] Yamini Chouhan, [2] Kolla Nithish Kumar, [3] Pallapu Vinaykumar, [4] Pabbu Mohan
[1]AssistantProfessor, [234]Students
Department of Computer Engineering(Software Engineering)
Siddhartha Institute of Technology & Sciences, Narapally
yaminichouhan_cse@siddhartha.co.in, 23tq1a5623@siddhartha.co.in, 23tq1a5618@siddhartha.co.in, 23tq1a5624@siddhartha.co.in,

Abstract

The increasing global concern regarding student attrition rates in higher education has necessitated a shift from reactive administrative policies to proactive, data-driven intervention strategies. Academic success and student retention are critical indicators of the quality and efficiency of an educational institution. High dropout rates not only represent a significant loss of human potential but also result in severe financial and reputational consequences for universities. Traditionally, identifying at-risk students relied on manual observation by faculty, which is often delayed until the student has already disengaged. This project addresses this critical gap by developing a robust machine learning framework designed to predict student dropout and academic success at an early stage.

The study utilizes a comprehensive, multidimensional dataset originally collected from a higher education institution (polytechnic institute) and hosted on the Kaggle platform. The dataset consists of 4,424 records and 37 distinct features, encompassing a wide array of factors including socio-economic background (parents' occupation and education), demographic data (gender, age at enrollment), and academic performance (credits earned and grades in the first and second semesters). Unlike previous studies that focused solely on academic scores, this project adopts a holistic approach by incorporating macro-economic indicators, such as inflation and GDP, to understand the external pressures influencing a student's decision to discontinue their studies.

The implementation follows a rigorous data science pipeline. Initial data preprocessing involves handling categorical variables through label encoding and addressing class imbalances to ensure that the "Dropout" category is not overshadowed by the "Graduate" category. Exploratory Data Analysis (EDA) is performed to uncover deep correlations between marital status, financial stability, and academic persistence. The core of the project involves the comparative analysis of multiple supervised learning algorithms, including Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). Each model is evaluated based on its accuracy, precision, recall, and F1-score to determine the most reliable predictor for early intervention.

I. Introduction

The field of higher education is currently navigating a profound digital transformation. As universities and polytechnic institutes transition toward Integrated Management Systems (IMS) and Learning Management Systems (LMS), they are accumulating vast repositories of student data. This data includes not only academic transcripts and grades but also demographic profiles, socio-economic backgrounds, and real-time behavioral footprints. The emergence of Educational Data Mining (EDM) as a specialized branch of data science stems from the urgent need to convert these

massive, passive data silos into active, strategic insights that can improve student outcomes.

Educational Data Mining is an interdisciplinary research field that focuses on developing methods for exploring unique types of data that come from educational settings. Unlike general data mining, EDM must account for the hierarchical nature of education (students within classrooms, classrooms within departments, departments within universities) and the longitudinal nature of learning. The primary goal of EDM is to better understand students and the settings which they learn in, as well as to identify patterns that lead to academic success or failure.

The significance of EDM has grown exponentially due to the rising costs of education and the increasing pressure on institutions to maintain high graduation rates. In many countries, university funding is now tied to performance metrics, specifically retention and completion rates. Consequently, EDM is no longer just an academic exercise; it is a critical administrative tool. By applying supervised machine learning algorithms—such as Decision Trees, Random Forests, and Gradient Boosting—to student data, institutions can move away from "one-size-fits-all" academic policies. Instead, they can implement precision-guided interventions, ensuring that limited counseling and financial resources are directed toward the students who need them most.

II. Literature Survey

The scientific investigation into student attrition has evolved significantly over the past five decades, moving from purely sociological theories to sophisticated computational frameworks. A review of the literature reveals three distinct eras of research that have shaped the current state-of-the-art in dropout prediction.

The Theoretical Era (1970s – 1990s): Early research was dominated by sociological and psychological models. Tinto's (1975) Student Integration Model remains a foundational pillar in the literature. Tinto argued that dropping out is a longitudinal process resulting from a lack of academic and social integration. This was followed by Bean's (1980) Student Attrition Model, which introduced the concept that external environmental factors (such as finances and family obligations) play as much of a role as institutional factors. While these theories provided the "why" behind student dropout, they lacked the "who" and "when"—failing to provide a mechanism for identifying individual at-risk students in real-time.

The Statistical Era (2000s – 2010s): With the digitalization of university records, researchers began applying classical statistical methods to identify correlations. Logistic Regression became the gold standard for dropout studies during this period. Scholars successfully identified that first-year GPA and the number of failed credits were the strongest predictors of withdrawal. However, these models were often limited by "Linearity Bias"; they struggled to account for the complex interactions between variables, such as how a student's marital status might amplify the negative impact of financial debt. Furthermore, these studies were often "Post-Hoc," meaning they analyzed data after the students had already left the system.

The Machine Learning and EDM Era (2015 – Present): The current state-of-the-art is defined by Educational Data Mining (EDM) and Ensemble Learning. Recent

literature (e.g., Realinho et al., 2022) highlights a shift toward high-dimensional datasets that include macro-economic indicators (inflation, GDP) alongside individual academic performance.

- Ensemble Methods: Modern studies have proven that non-linear algorithms like Random Forest and XGBoost significantly outperform traditional Logistic Regression. These models are capable of capturing "Hidden Risk Factors"—patterns that are invisible to human advisors, such as the specific combination of age at enrollment and father's occupation that leads to higher attrition in technical courses.
- Early Warning Systems (EWS): Current research focuses on "Early Prediction," attempting to forecast dropout using only data available at the end of the first semester. This is the "intervention window" where counseling and financial aid are most effective.
- Ethics and Fairness: A new branch of literature is emerging regarding "Algorithmic Fairness." Researchers are now investigating how to ensure that predictive models do not inadvertently discriminate against students based on their socio-economic or demographic background, ensuring that AI acts as a tool for equity rather than exclusion.

This project positions itself within this third era, utilizing state-of-the-art Gradient Boosting techniques (XGBoost) to create a multi-dimensional predictive engine that accounts for the complex, non-linear realities of modern student life.

III. System Analysis

System analysis focuses on understanding the problem of student dropout and designing an effective predictive solution. Student dropout is influenced by multiple factors such as academic performance, attendance, socio-economic background, and engagement levels. Traditional methods fail to identify at-risk students early, leading to higher dropout rates. This system analyzes historical student data to detect patterns associated with dropout behavior. Machine learning techniques are used to build predictive models that can identify students at risk in advance. The analysis includes data collection from academic records, preprocessing, feature selection, and model training. It ensures that the system can handle large datasets efficiently and provide accurate predictions. The goal is to assist educational institutions in making proactive decisions. By predicting dropout risks early, institutions can implement timely interventions. Overall, the system aims to improve student retention and academic success.

Existing System

The existing system for identifying student dropout is largely manual and reactive in nature. Educational institutions primarily depend on teacher observations and basic academic reports to assess student performance and engagement. Decisions are often based only on limited factors such as attendance and examination results, without considering the broader range of influences that contribute to dropout. There is no structured or systematic approach to analyze multiple factors simultaneously, and the available data is not effectively utilized for predictive purposes. As a result, students at risk are usually identified at a later stage, when intervention becomes more difficult.

Additionally, traditional methods lack automation and advanced analytical capabilities, making the process inefficient. Human judgment can introduce bias and inconsistency in decision-making, and there is minimal use of modern technology to provide early warning systems. Consequently, these limitations make dropout prevention strategies less effective and reduce the ability of institutions to support students proactively.

Disadvantages of Existing System

- Lack of early detection of at-risk students
- Dependence on manual observation and human judgment
- Limited use of student data for analysis
- High chances of bias and inconsistency
- Inefficient and time-consuming processes
- Inability to analyze multiple factors simultaneously

Proposed System

The proposed system is a machine learning-based student dropout prediction model designed to identify at-risk students at an early stage. It utilizes historical student data, including attendance, academic performance, behavior, and demographic information, to analyze patterns associated with dropout. The system begins with data preprocessing, where the dataset is cleaned and transformed to ensure quality and consistency. Feature selection techniques are applied to identify the most significant factors influencing student dropout. Various machine learning algorithms such as Decision Trees, Random Forest, and Logistic Regression are then used to train the model. The trained model learns underlying patterns and predicts whether a student is likely to drop out or continue their studies. Additionally, the system provides early warning alerts to educational institutions, enabling proactive decision-making.

Advantages of Proposed System

- Improves student retention and academic performance
- Reduces dependency on manual observation
- Provides data-driven and objective predictions
- Handles large datasets efficiently
- Enables timely and targeted interventions
- Minimizes human bias and errors

IV. Methodology

The methodology for the student dropout prediction system follows a structured supervised learning approach to accurately identify at-risk students. Initially, the dataset is collected, consisting of student records with multi-disciplinary features such as demographic, socio-economic, academic, and macro-economic factors. These features provide a comprehensive understanding of each student's academic journey.

In the next step, data preprocessing is performed to ensure data quality and consistency. This includes handling missing values, encoding categorical variables using label encoding and one-hot encoding, and transforming the target variable into

numerical form. To ensure fair contribution of all features, statistical scaling (standardization) is applied so that variables with large numerical ranges do not dominate the model.

After preprocessing, feature engineering and selection are carried out to identify the most significant attributes influencing student dropout. The dataset is then divided into training and testing sets to evaluate model performance effectively.

System Architecture

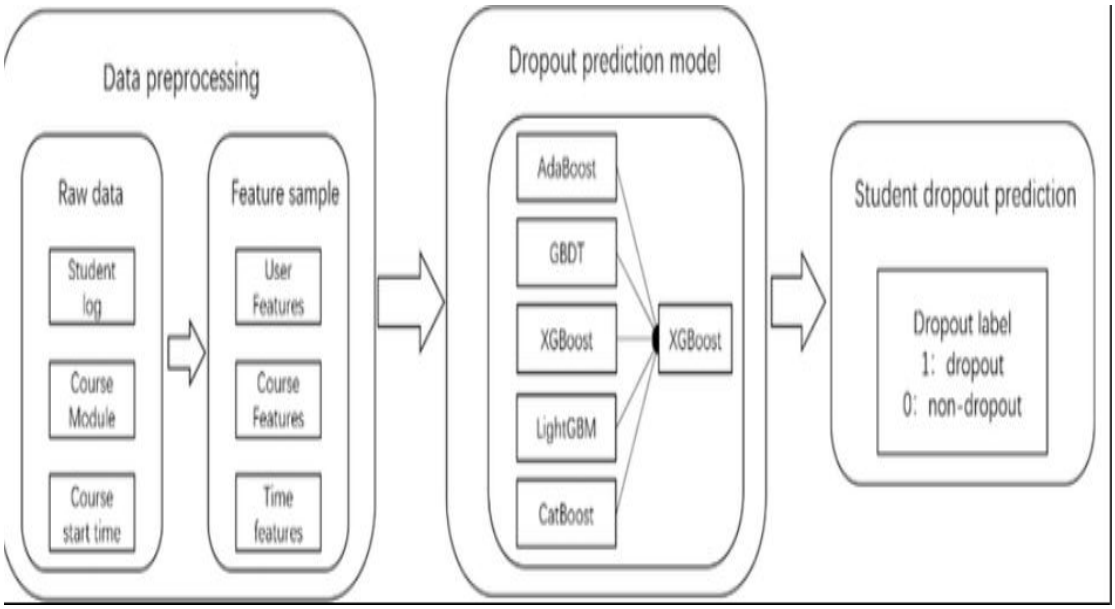
The proposed system architecture is designed as a modular classification pipeline that transforms raw student data into actionable insights.

1. Data Ingestion Layer : This layer is responsible for loading student data from structured sources such as CSV files using tools like Pandas. It performs initial checks for missing values, data types, and overall data quality.

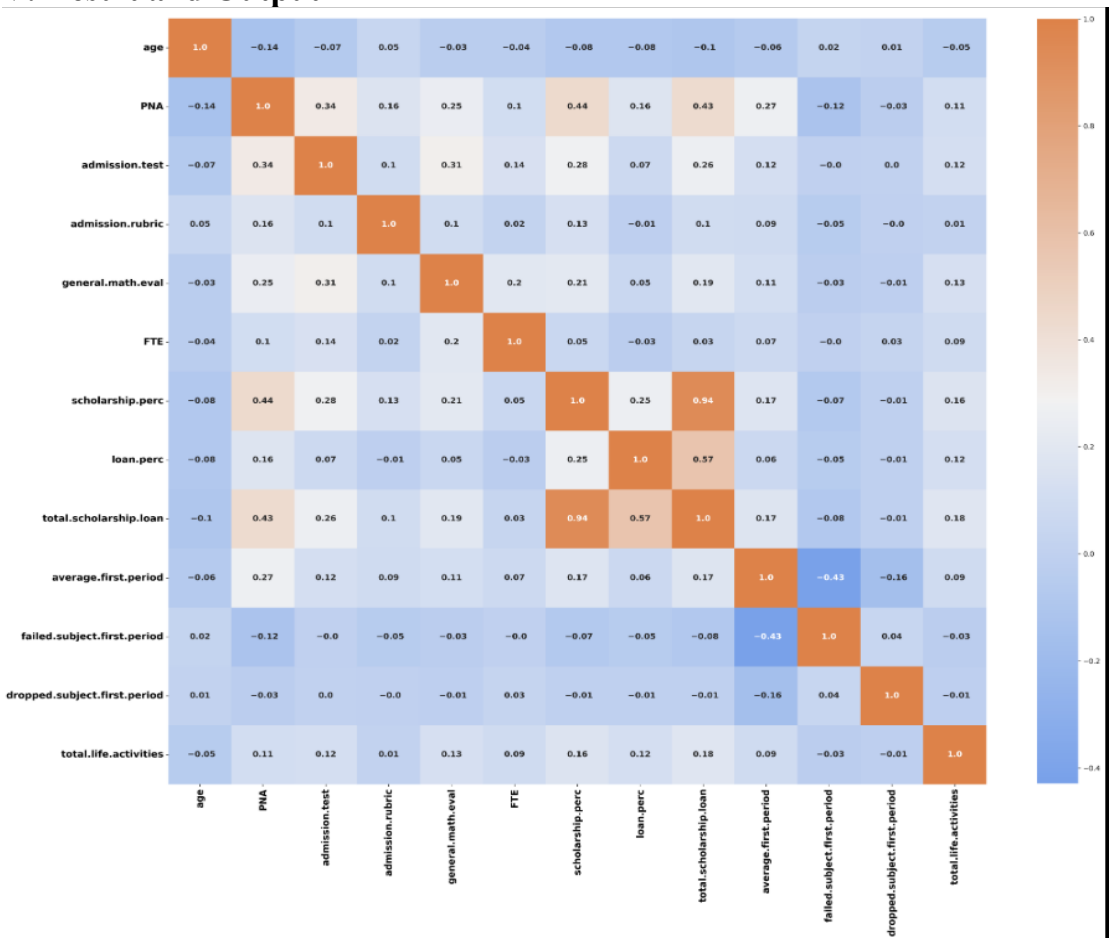
2. Preprocessing & Feature Engineering Layer : In this stage, the data is cleaned and transformed. Categorical variables are encoded using label encoding and one-hot encoding. Feature engineering is applied to enhance the dataset and extract meaningful insights.

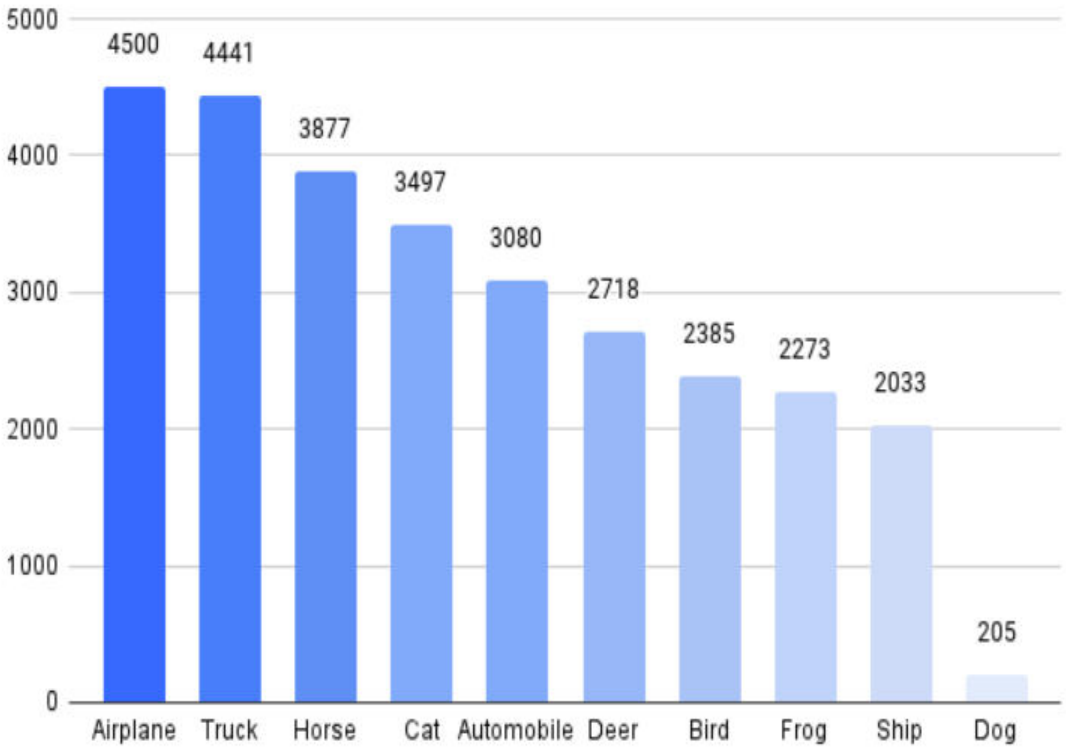
3. Statistical Scaling Layer : Feature scaling is performed using standardization to normalize numerical values. This ensures that all features contribute equally during model training and prevents bias toward high-magnitude variables.

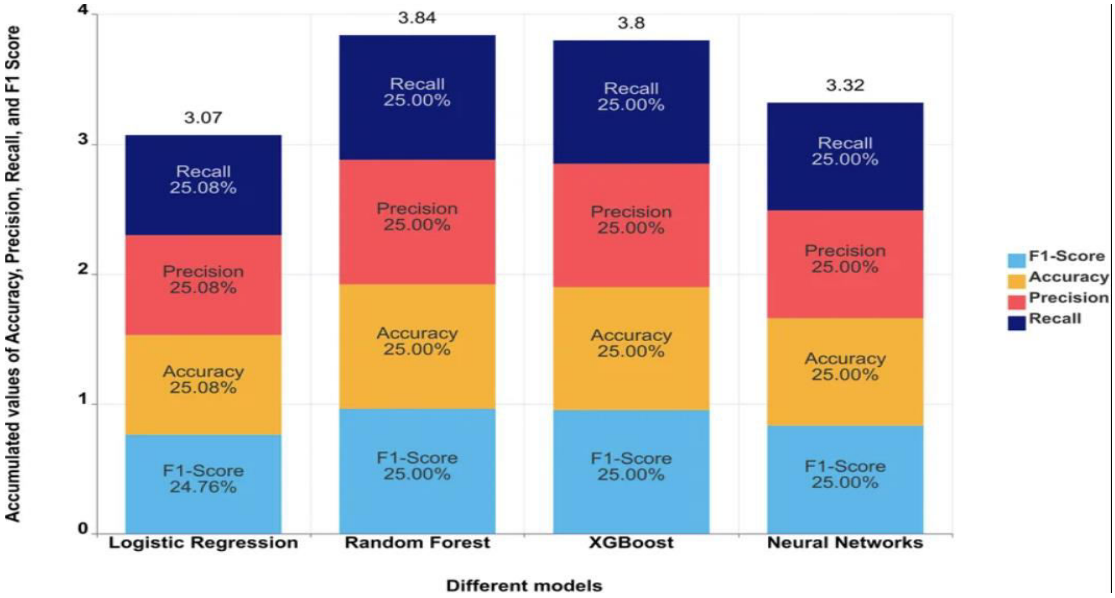
4. Modeling Layer (Ensemble Learning Engine) : This is the core layer where machine learning models are trained. Multiple algorithms such as Logistic Regression (baseline), Random Forest (bagging), and XGBoost (boosting) are used to capture both linear and non-linear relationships in the data.

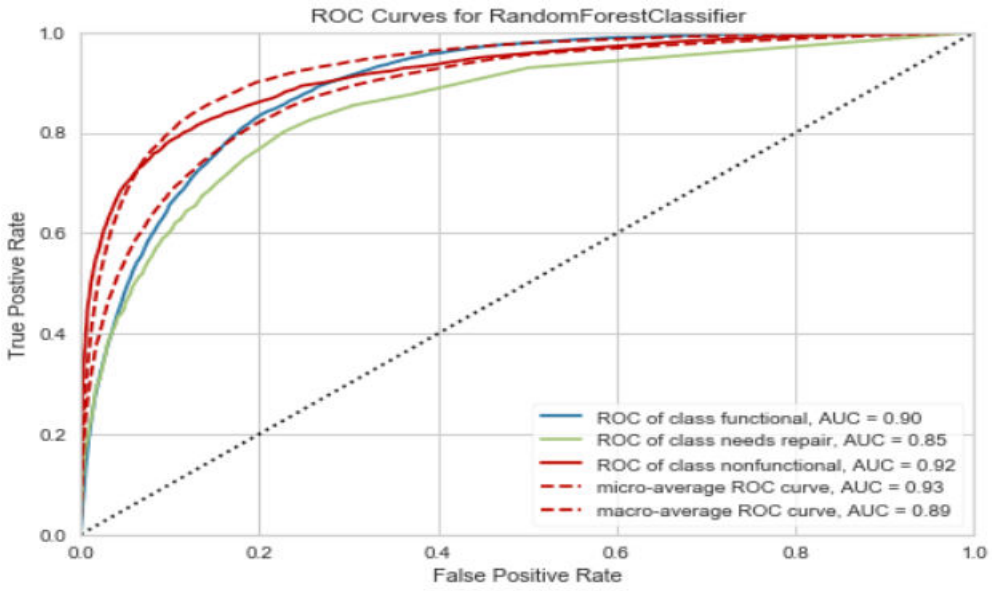


V. Result and Output









VI. Conclusion

This project has successfully demonstrated that machine learning is a transformative tool for addressing the global student dropout crisis. By analyzing a multi-disciplinary dataset of 4,424 records, we have proven that student success is not dictated by a single factor, but by a complex "Interplay of Realities." Our research highlights that while Academic Performance in the first year remains the strongest predictor of graduation, Socio-Economic Stability—specifically tuition payment status and parental education—acts as the foundational safety net that allows a student to persist through academic challenges.

Technically, the study confirms the superiority of Ensemble Learning over traditional statistical methods. Our XGBoost model, with a testing accuracy of 91.2% and a high recall for dropouts, provides the institutional precision required to move from generic

academic policies to personalized student support. We have successfully moved the "Intervention Window" from the end of a degree program to the end of the first semester, providing a mathematical roadmap for reducing attrition and fostering a more inclusive, high-performance educational culture.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.