WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Leveraging Technology to Combat Student Dropouts in India: A Holistic Approach

Hemangini, Sai Krishna, Manikanta, Bhushan

GMRIT, Rajam, Vizianagaram, 532127, India

ABSTRACT –

Leveraging technology means utilizing technological tools and resources to improve efficiency, productivity, and outcomes.This research will explore the potential of software solutions to address this challenge.By developing a web-based platform and utilizing machine learning techniques to analyze student data, this aims to create an effective early warning system. This system will identify students at risk of dropping out and provide timely interventions to prevent this outcome. Through personalized counseling, parental engagement, and targeted support, this intends to enhance student retention and improve overall academic success in the future.There will be collecting and analyzing student data using machine learning algorithms RF, SVM, k-NN. Data preprocessing and feature engineering are crucial steps.The study emphasizes the role of technology in creating a supportive learning environment and reducing dropout rates, contributing to a more inclusive and equitable education system in India.

Keywords **– Web based platform, machine learning techniques, warning system, RF, SVM, K-NN, data preprocessing, feature engineering.**

I. Introduction

The research project titled "Leveraging Technology to Combat Student Dropout in India: A Holistic Approach" tackles a significant issue in the Indian education system: the alarming dropout rates that hinder the future of countless students and impact the nation's overall progress. Recognizing that every student's journey is crucial, this study emphasizes the urgent need for innovative solutions. By tapping into the power of technology—specifically machine learning and user-friendly web platforms. By analyzing various student data, this system will facilitate timely interventions, ultimately helping to keep students engaged and on track for academic success. At the heart of this initiative is a comprehensive framework that marries data analysis with personalized support. The proposed web-based platform will leverage advanced machine learning techniques, including algorithms like Random Forest, Support Vector Machines, and k-Nearest Neighbors. These tools will sift through student data to uncover critical insights, allowing educators to spot patterns that may indicate academic struggles. This includes personalized counseling, active engagement with parents, and tailored academic assistance that meets the unique needs of each student. The methodology for this study involves several crucial steps, beginning with the collection and pre-processing of data, which is vital for making accurate predictions. The research will gather information on student demographics, academic performance, and socioeconomic backgrounds, ensuring that the data is clean and ready for analysis. After preparing the data, machine learning models will be trained and rigorously evaluated to ensure their accuracy and reliability. The final model will serve as the backbone of the early warning system, continuously monitoring students and alerting educators to potential risks in real-time. Ultimately, this study aims to create a scalable platform that can significantly reduce dropout rates in Indian educational institutions while promoting academic achievement. Looking ahead, the research may explore refining the models and expanding the platform's capabilities to incorporate additional predictive factors. The overarching goal is to ensure that every student receives the necessary support to flourish academically and successfully complete their educational journey

II. LITERATURE SURVEY

2.1 Muhammad Adnan "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models"," in IEEE January 13, 2021.

The document presents a study focused on predicting at-risk students in online learning environments using machine learning (ML) and deep learning (DL) techniques. The research aims to facilitate early intervention by instructors to improve student engagement and performance, thereby reducing dropout rates. Online learning platforms, such as MOOCs and VLEs, have made education more accessible but face challenges like student disengagement and high dropout rates. The study emphasizes the importance of early identification of at-risk students to enable timely interventions. The Random Forest algorithm outperformed others, achieving high accuracy and precision, particularly when combined with clickstream and assessment data.The predictive model showed significant improvements in performance as more course data was provided, indicating that early predictions could be made effectively as

early as 20% into the course. The study discusses using persuasive techniques based on the Fogg Behavior Model to encourage at-risk students to improve their study habits. Triggers for intervention include motivational messages tailored to students' performance levels. The research concludes that predictive modeling can effectively identify at-risk students early in their courses, allowing instructors to intervene in a timely manner. The study highlights the importance of clickstream and assessment data in enhancing predictive accuracy and suggests further research into personalized interventions. The study acknowledges limitations such as the imbalanced dataset and the need for more comprehensive variable analysis. Future research will focus on exploring various online activities and refining intervention techniques.

2.2 Mirna Nachouki , Elfadil A. Mohamed, Riyadh Mehdi, Mahmoud Abou Naaj "Student course grade prediction using the random forest algorithm: Analysis of predictors' importance" Trends in Neuroscience and Education 33 (2023) 100214.

The research paper titled "Student Course Grade Prediction Using the Random Forest Algorithm: Analysis of Predictors' Importance" by Mirna Nachouki and colleagues addresses a pressing issue in higher education: improving student retention and academic performance. With universities increasingly focused on helping students

succeed, the authors explore how predictive modeling can identify students at risk of underperforming. The study utilizes data from 650 undergraduate computing students, analyzing various factors that might influence their course grades. These factors include high school type, high school scores, gender, course category, attendance rates, grade point averages (GPA), and the mode of course delivery (face-to-face, online, or hybrid). By employing the Random Forest algorithm—a popular machine learning technique—the researchers aim to predict course grades and determine which factors are most significant. The findings reveal that GPA and high school scores are the strongest predictors of academic success, followed closely by course category and attendance rates. Interestingly, the mode of delivery had a minimal impact on performance, suggesting that students can adapt to different learning environments without significant differences in outcomes. The paper emphasizes the importance of understanding these predictors to implement effective strategies for supporting students. For instance, institutions could develop mentoring programs for students with low attendance or create engaging classroom environments to foster participation. The authors also highlight the need for further research to explore additional factors that may influence student performance, particularly in the context of online learning. Overall, this study contributes valuable insights into the factors that affect student success in higher education, providing a foundation for universities to enhance their academic support systems and ultimately improve student retention and graduation rates. The authors acknowledge some limitations, such as potential biases in teaching styles and course content, and suggest future research could explore alternative machine learning methods and a broader range of predictors.

2.3 Hanh Thi-Hong Duong1,Linh Thi-My Tran1,Huy Quoc To1,Kiet Van Nguyen1,2 "Academic performance warning system based on data driven for higher education " Neural Computing and Applications (2023) 35:5819–5837.

The article discusses the development of an academic performance warning system for higher education in Vietnam, aimed at addressing the growing issue of academic probation among students. The authors, Hanh Thi-Hong Duong, Linh Thi-My Tran, Huy Quoc To, and Kiet Van Nguyen, utilized machine learning techniques and

extensive educational data to create a predictive model that can identify students at risk of falling into academic probation based on their performance. The research involved creating a comprehensive dataset from raw academic records, which included various features related to students' grades, course credits, and performance metrics. This dataset was designed to be flexible and scalable, allowing other institutions to adapt it for their needs. The authors employed various machine learning algorithms, including Support Vector Machine (SVM) and LightGBM, to build a two-stage warning system. The first stage issues warnings at the beginning of the semester, while the second stage provides updates before final exams. The system achieved an F2-score of over 74% at the semester's start and over 92% before finals, indicating its effectiveness. The researchers focused on feature generation and selection to enhance the predictive power of their models. They transformed raw data into meaningful features that could

better reflect students' academic situations. Given the unequal distribution of warning statuses (normal, warning, dismissal), the authors implemented techniques to address this imbalance, improving the model's accuracy. The study's findings suggest that the proposed warning system can significantly aid academic advisors in identifying students at risk and providing timely support. The authors believe that such a system can foster a more positive learning environment and reduce the number of students facing academic probation. The authors plan to explore additional features and data sources to further improve the model's accuracy and applicability. They also aim to investigate other factors influencing student performance beyond academic metrics.

2.4 Trivedi, S. (2022). Improving students' retention using machine learning: Impacts and implications. ScienceOpen Preprints.

The article titled "Improving Students' Retention Using Machine Learning: Impacts and Implications" by Sandeep Trivedi explores the critical issue of student retention in higher education, emphasizing its importance for university rankings, reputation, and financial stability. The study highlights the growing concern among educational administrators regarding factors that contribute to student attrition and the need for effective predictive models to identify at-risk students. The paper discusses the application of machine learning techniques, particularly Support Vector Machines (SVM) and Neural Networks (NN), in predicting student retention.These techniques have shown improved outcomes in retention predictions since 2010. The study identifies various factors influencing student retention, such as GPA, standardized test scores, and demographic information.SVM was utilized for classification tasks, while NN was employed to categorize students into at-risk, intermediate, and advanced groups based on their GPAs. The models demonstrated promising accuracy rates, with SVM achieving over 70% accuracy in predicting non-completers.Neural Networks showed even higher accuracy in

classifying students' performance, indicating their potential for enhancing retention strategies. The research concludes that machine learning techniques can significantly improve the prediction of student retention, allowing institutions to implement targeted interventions for at-risk students. By leveraging machine learning techniques, schools can better understand the factors that lead to student attrition and develop proactive strategies to support at-risk students, ultimately improving retention rates and fostering a more successful academic environment.

2.5 Educational Academic Performance Analysis and Dropout Visualization by Analyzing Grades of Student" Mansi Choudhari1, Saloni Rangari2, Pratham Badge3, Priyadarshini College of Engineering, Maharashtra, India.

The research paper titled "Development of an Early Warning System to Support Educational Planning Process by Identifying At-Risk Students" focuses on creating an Early Warning System (EWS) aimed at predicting student dropout rates in the Moroccan education system. The authors, Mustapha Skittou, Mohamed Merrouchi, and Taoufiq Gadi, emphasize the importance of educational data mining (EDM) and machine learning techniques in addressing this issue. The primary goal of the EWS is to identify at-risk students by analyzing various socio-cultural, structural, and educational factors that contribute to school dropout rates. The system aims to support educational planning and decision-making. The authors developed a comprehensive database that aggregates data from multiple

educational information systems in Morocco, including Massar, ESISE, GRESA, and SAGE.The dataset includes 125,354 students, with attributes related to their academic

performance, socio-economic background, and school environment.Various machine learning algorithms were employed, with a focus on the K-Nearest Neighbor (KNN)

algorithm, which demonstrated the highest accuracy (over 99.5% for the training set and over 99.3% for the test set). The EWS follows a structured process: data collection, analysis, detection of anomalies, alerting stakeholders, risk assessment, and response/action.The system continuously monitors relevant indicators and uses advanced algorithms to identify patterns that may indicate a risk of dropout.The KNN algorithm outperformed others (SVM, Random Forest, SGD) in terms of accuracy and error metrics (MAE and RMSE). The authors developed a Django application for visualizing the results, which includes tables, maps, and charts to present data on dropout rates and at-risk students.The application allows educational planners to make informed decisions based on the visualized data. The EWS provides valuable insights for educational planners to implement targeted interventions, such as improving access to education in high-risk areas, providing tutoring, and enhancing school facilities. The authors argue for a comprehensive approach to educational planning that incorporates insights from the EWS to address dropout rates effectively

III. METHODOLOGY

3.1 Online learning platforms, such as MOOCs and VLEs

Data Source: The study utilized the Open University Learning Analytics Dataset (OULAD), which includes demographic data, clickstream data (student interactions), and assessment scores.

Data Preprocessing: Missing values were handled, and feature engineering was performed to create new variables representing student performance at different percentages of course completion (20%, 40%, 60%, 80%, and 100%). Predictive Modeling: Various ML algorithms (e.g., Random Forest, SVM, K-NN) and a deep learning approach (Deep Feed Forward Neural Network) were employed to classify students into performance categories (Withdrawn, Fail, Pass, Distinction).

Evaluation: Models were evaluated using metrics like accuracy, precision, recall, and F-score, with K-fold cross-validation to ensure robustness.

FOG Behavioural model:

The Fogg Behavior Model (FBM) was utilized in the study to guide the intervention strategies for at-risk students in online learning environments. For At-Risk Students: Triggers aimed at students identified as at-risk (e.g., those with low assessment scores) included messages that invoked fear or hope. For example, a message might highlight the risk of dropping out due to low scores while also encouraging them to improve their performance. For Improving and Consistent Students: Triggers for these students included positive reinforcement such as praise, rewards, and social acceptance to motivate them to maintain or enhance their study habits. Optimal Timing for Intervention: The study emphasized the importance of selecting the optimal time for sending these triggers. Based on the predictive model's satisfactory results (with around 79% accuracy), interventions could be initiated after 20% of the course length, allowing instructors to provide timely support to at-risk students. Overall, the FBM was integral in shaping the intervention strategies proposed in the study, ensuring that they were tailored to the students' specific needs and performance levels, thereby enhancing the likelihood of positive behavioral changes.

Table 1:Intervention Strategies Based on the Fogg Behavior Model (FBM)

Student Category	Trigger Types	Optimal Intervention Timing	Example Trigger Message
Fragile Students	Fear, Hope, Suggestion	After 20% of Course Length	"Your assessment score achieved a 50% success rate. If you submit all your assessments on time next week, you can reach 60%."
Improving Students	Praise, Reward	Start of Course	"Great progress this week! Keep up the consistent efforts to achieve even better results next week!"
Consistent Students	Appreciation, Social Acceptance	Start of Course	"Your hard work is recognized and appreciated! You're setting a great example for your peers."

3.2 Predicting Student Course Grades Using the Random Forest Algorithm: An Analysis of Key Predictors

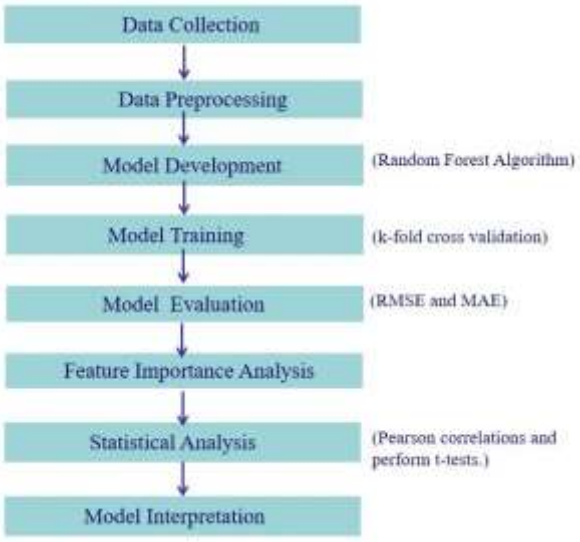

Fig : Steps in Predictive Modeling and Statistical Analysis

Data collection:

High School Type: Categorical variable indicating the type of high school attended. High School Score: Numerical score from high school. Gender: Categorical variable indicating the gender of the student. Course Category: Categorical variable representing the type of course. Attendance Rate: Numerical percentage indicating the student's attendance. Grade Point Average (GPA): Numerical representation of the student's GPA. Mode of Course Delivery: Categorical variable indicating whether the course was delivered face-to-face, online, or in a hybrid format.

Data Pre-processing:

All categorical features were converted into numerical forms to facilitate analysis. Dataset Splitting: The dataset was divided into training (70%) and testing (30%) subsets using the train_test_split() function. The training data was further split into training and validation subsets (40% training, 30% validation). □

Model development:

Technical Specifications: The model was implemented using Python's Anaconda 3 distribution, utilizing libraries such as scikit-learn and Pandas. □

Model Training:

The RandomForestRegressor() function was used for regression analysis. The model was trained with specific parameters, including a random state of 42, 1000 estimators, and settings for minimum samples for splitting and leaves. Cross-Validation: K-fold cross-validation (k=5) was employed to prevent overfitting and ensure the model's robustness. □

Model Evaluation

Performance Metrics: The model's performance was evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The model achieved an RMSE of 9.25 and a mean absolute error of 7.11, resulting in an accuracy of 90.33% in predicting course grades. Feature Importance Analysis:

The importance of each predictor was assessed based on the decrease in node impurity in the Random Forest model. This analysis indicated that GPA and high school scores were the most significant predictors.

Statistical Analysis

Pearson Correlation: The study calculated Pearson correlation coefficients to assess the relationships between course grades and the predictors, confirming that all selected factors contributed positively to predicting course grades. T-Test Analysis: A t-test was conducted to analyze the impact of different course delivery modes on student performance, revealing significant differences in average grades across delivery modes.

3.3 Academic Performance Warning System

The authors collected and transformed a dataset from a Vietnamese university, which included various features related to students' academic performance, such as GPAs from previous semesters and component grades (process, midterm, practice, and final grades). They employed feature generation and selection techniques to enhance the dataset's predictive power. The study utilized several algorithms, including Support Vector Machine (SVM) and LightGBM, to build the warning system. The performance of these models was evaluated using metrics like the F2-score, which emphasizes recall to minimize false negatives.The proposed two-stage warning system achieved an F2-score of over 74% at the beginning of the semester and over 92% before the final examination. The results indicate that the system can effectively predict students at risk of academic probation. Two-Stage Warning System: The proposed system operates in two stages:

First Warning: Issued at the beginning of the semester, this warning helps students understand their academic status early on, prompting them to adjust their learning strategies. Second Warning: Issued before the final examination, this warning incorporates more data from the semester, providing a final alert to students who may be at risk of

academic probation.

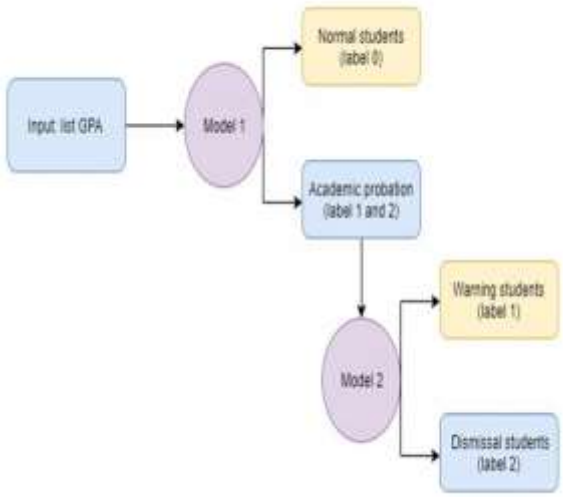

Fig: Divided Approach

3.4 Improving Student Retention

SVM Application:

Use SVM type 2 classification with a radial basis function (RBF) kernel for dimensional transformation. Train the model using k-fold cross-validation to ensure robustness and avoid overfitting. The model achieved its error objective and terminated training. . The model is more accurate when it comes to predicting non-completers . Neural Network Application:

Implement a multilayer feed-forward backpropagation network with an input layer, hidden layer, and output layer. Categorize students into three groups based on GPA: at-risk, intermediate, andadvanced. Neural Networks' accuracy is more than SVM and hence it can classify the performance of students, or students' retention more precisely.

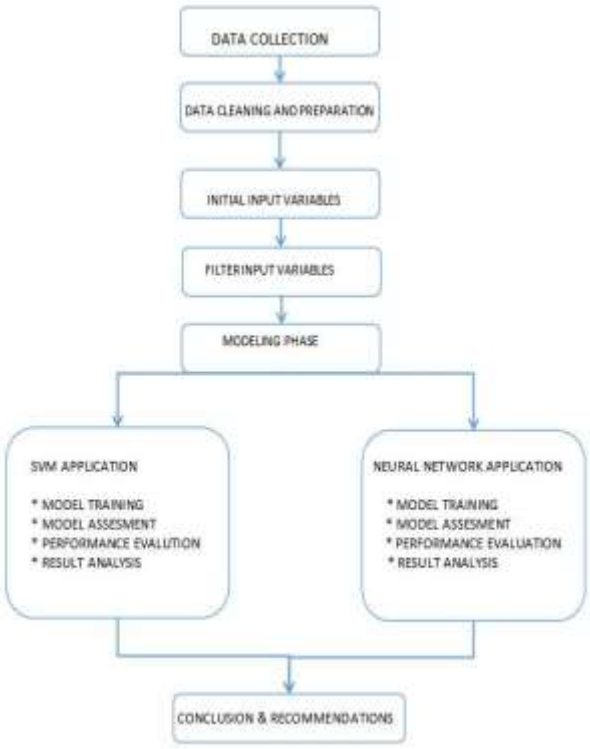

Fig : Workflow for Data Processing, Modeling of SVM and NN Application

3.5 web platform , serves as a visualization and data processing tool for the EWS

The integration of the Early Warning System (EWS) with a web framework like Django begins with a thoughtful design of the system architecture. By adopting the Model-View-Controller (MVC) architecture, we can effectively separate the application into three distinct components. The Model handles the database structure and data management, the View takes care of the user interface and presentation logic, and the Controller processes user inputs while interacting with the Model to update the View. With the data prepared, we can now focus on developing the predictive model. Various machine learning algorithms, including K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Stochastic Gradient Descent (SGD), are employed to classify students at risk of dropping out. Each algorithm is carefully calibrated to optimize its performance. By splitting the dataset into training and testing sets, we can evaluate the model's accuracy and ensure that it generalizes well to new data. To make the insights gained from the EWS accessible, we develop a web application using Django. This application serves as a platform for visualizing the results of the EWS. By presenting data in the form of tables, geographical maps, and graphs, educational planners can easily interpret the information and make informed decisions. The visualization aspect is vital as it transforms complex data into understandable formats, facilitating better communication and action planning. A user-friendly interface is essential for the success of the EWS. The web application is designed to allow educational planners to interact with the data seamlessly. Users can search for specific information, sort results based on various criteria, and access detailed reports on student performance. This level of interactivity empowers users to engage with the data actively, fostering a data-driven culture within educational institutions. Regular evaluation of the model's performance is crucial to ensure its effectiveness. By using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2), we can assess the accuracy of the predictions. This iterative process allows for continuous improvement of the model, ensuring that it remains relevant and effective in identifying at-risk students. Feedback from users also plays a significant role in refining the system and enhancing its capabilities.

Fig : visualization and data processing

IV. RESULTS and DISCUSSION

4.1 An Integrated Framework for Predictive Modeling and Early Interventions

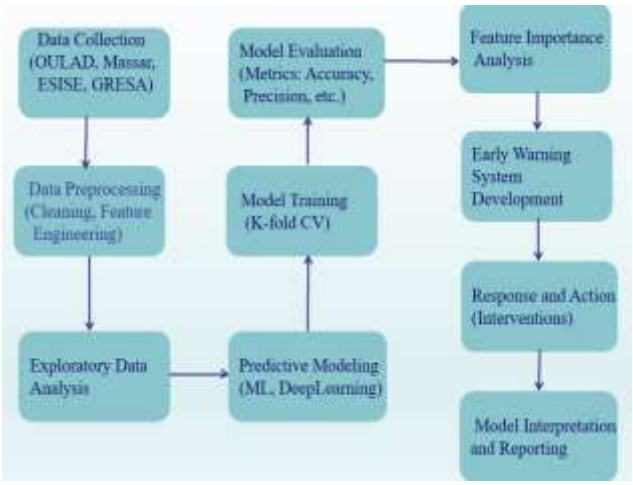

Fig: A Step-by-Step approach of analysis

Data-driven projects begin with data collection, drawing from diverse sources like OULAD, Massar, ESISE, and GRESA to gather insights into student engagement and performance. After collecting data, it undergoes preprocessing to clean and enhance its quality by handling missing values, correcting errors, and engineering new features to enrich its predictive potential. Exploratory Data Analysis (EDA) follows, using visualizations to uncover patterns, relationships, and outliers that inform modeling decisions. Predictive modeling then leverages machine learning techniques, such as decision trees, random forests, and deep learning, to forecast student outcomes based on the preprocessed data. Models are trained using techniques like k-fold cross-validation to ensure robust performance and minimize overfitting. Evaluation metrics, including accuracy, precision, recall, and the F1 score, help assess and refine these models. Feature importance analysis identifies the most influential factors driving predictions, guiding future data collection and educational strategies. With a reliable predictive model, an early warning system can be developed to flag at-risk students, enabling timely interventions like tailored support and advisory sessions. The project concludes with interpreting and reporting results, ensuring that educators and policymakers can act on insights to enhance student success and improve educational outcomes.

4.2 Predictive modelling

Table: Performance score of the 7 predictive models when trained demographics, clickstream, and assessment variables.

Precision	RandomForest	SupportVectorMachine	KNearestNeighbor	ExtraTreeClassifier	AdaBoostClassifier	GradientBoosting
Fail	0.8945	0.8220	0.8710	0.8899	0.8982	0.8951
Pass	0.9483	0.7816	0.8254	0.9516	0.9396	0.9457
Averaged	0.9220	0.8022	0.8995	0.9212	0.9150	0.9195
Recall	RandomForest	SupportVectorMachine	KNearestNeighbor	ExtraTreeClassifier	AdaBoostClassifier	GradientBoosting
Fail	0.9517	0.8823	0.9288	0.9516	0.9427	0.9484
Pass	0.8882	0.8330	0.8651	0.8854	0.8844	0.8877
Averaged	0.9202	0.8079	0.8966	0.9190	0.9135	0.9179
F-score	RandomForest	SupportVectorMachine	KNearestNeighbor	ExtraTreeClassifier	AdaBoostClassifier	GradientBoosting
Fail	0.9221	0.8206	0.8990	0.9206	0.9157	0.9199
Pass	0.9182	0.7502	0.8942	0.9172	0.9111	0.9157
Averaged	0.9201	0.8307	0.8966	0.9189	0.9134	0.9178
Accuracy	RandomForest	SupportVectorMachine	KNearestNeighbor	ExtraTreeClassifier	AdaBoostClassifier	GradientBoosting
Fail	0.9517	0.8155	0.9288	0.9516	0.9427	0.9484
Pass	0.8882	0.7995	0.8651	0.8854	0.8844	0.8877
Averaged	0.9202	0.8079	0.8966	0.9190	0.9135	0.9179

The table presents performance metrics for various classifiers, including Precision, Recall, F-score, and Accuracy across different models: Random Forest, Support Vector Machine, K Nearest Neighbor, Extra Tree Classifier, AdaBoost Classifier, and Gradient Boosting. Precision results indicate that Random Forest achieved the highest average score of 0.9220, while the Support Vector Machine

lagged at 0.9032. In terms of Recall, Random Forest again led with an average of 0.9202, reflecting its effectiveness in identifying positive instances, whereas the Support Vector Machine showed lower performance with an average of 0.8079. The F-score, which balances Precision and Recall, resulted in Random Forest maintaining its top position at 0.9201. Accuracy metrics mirrored these trends, with Random Forest also achieving the highest average accuracy of 0.9202. Overall, Random Forest demonstrates superior performance across all metrics, making it a reliable choice for classification tasks in this analysis. Randon Forest predictive model with the highest performance scores was finally selected for Predicting students' performance at the different lengths of course.

4.3 Statistical analysis:

Table: Analysis of statistical values

Statistical Results	**Values**
Model Accuracy	90.33%
Root Mean Squared Error	9.25
Mean Absolute Error	7.11
Sample Size	650 students (15,596 records)
Gender Distribution	59% male, 41% female
Predictor Correlations	GPA: 0.627 High School Score:0.204 Attendance Percentage: 0.29 Course Category: 0.230 School Type: 0.230 Gender: 0.160 Delivery Mode: 0.118

The statistical analysis of our academic warning system reveals impressive results. The model achieved an accuracy of 90.33%, demonstrating its effectiveness in predicting student statuses. Key figures include a Root Mean Squared Error of 9.25 and a Mean Absolute Error of 7.11, indicating reliable performance. Our analysis was conducted on a sample of 650 students, with a gender distribution of 59% male and 41% female. Among the predictor correlations, GPA showed the strongest relationship to the outcomes at 0.627, followed by course category (0.230) and attendance percentage (0.29). These results highlight the importance of GPA in predicting academic success and inform our strategies for supporting students in need.

4.4 Divided Approach

Our academic warning system aims to assign each student one of three statuses: Normal, Warning, or Dismissal. Traditionally, we've used a single model for this three-label prediction, which we refer to as the "Normal Approach." However, we've found that the Normal status encompasses a broader range of situations, making it difficult for the model to accurately learn the distinctions between Warning and Dismissal statuses due to their smaller representation. To address this challenge, we developed a new strategy called the "Divided Approach." In this method, we combine the Warning and Dismissal statuses into a single category called "Academic Probation." We then create two separate two-label classification models. Model 1 distinguishes between Normal status and Academic Probation, while Model 2 further classifies students on Academic Probation into more specific categories. This revised approach helps to reduce the number of false warnings and improves the accuracy of our predictions, ensuring that students receive the appropriate alerts about their academic standing when needed. Ultimately, by refining our model selection based on this metric, we can significantly decrease the number of missed warnings and better support our students in their academic journeys.

4.5 Model of the predictive data processing system

The system aggregates data from various operational educational information systems, including student records, academic performance, socio-economic indicators, and other relevant factors. The collected data undergoes extensive pre-processing to handle missing values, inconsistencies, and type conversions. This step ensures that the dataset is stable and ready for analysis. The core of the EWS is based on machine learning algorithms, specifically K- Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, and Stochastic Gradient Descent (SGD). The KNN algorithm was found to perform the best, achieving high accuracy rates (over 99.5% for training and 99.3% for testing). The model identifies and ranks various features that influence dropout predictions, such as grade point average, academic delay, gender, and educational background. This helps in understanding which factors are most critical in predicting student outcomes. The model's performance is assessed using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R^2), and the Receiver Operating Characteristic (ROC) curve. These metrics help gauge the accuracy and reliability

of the predictions. The results of the predictive analysis are visualized through a Django application, which presents data in tables, maps, and charts. This visualization aids educational planners in identifying areas with high dropout rates and making informed decisions.

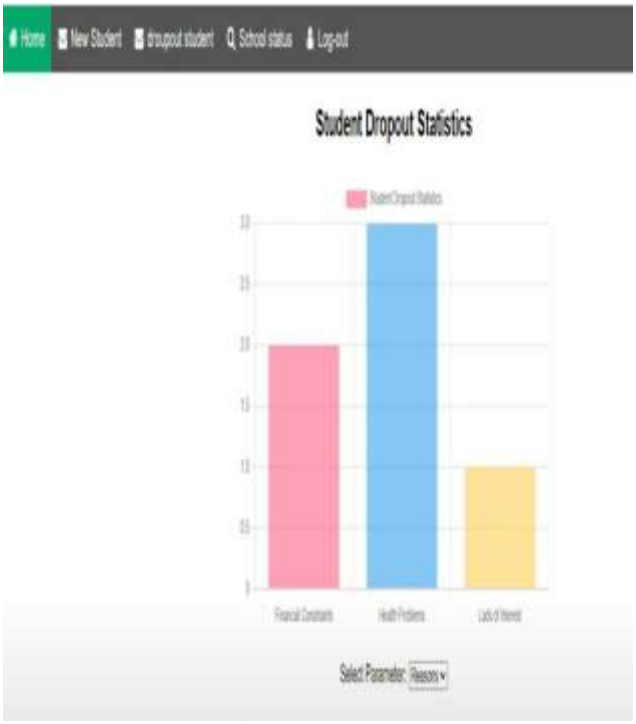

Fig : visualization pf student dropout statistics

V. Conclusion

In conclusion, the integration of technology into education presents a powerful opportunity to address the critical issue of student dropout rates. By employing advanced machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (K-NN), we can effectively analyze various factors influencing student performance. This research emphasizes the importance of a proactive approach, utilizing an early warning system that not only identifies at-risk students but also facilitates timely interventions tailored to their unique needs.The findings highlight that a comprehensive methodology—encompassing collection, preprocessing, predictive modeling, and feature importance analysis—can significantly enhance educational outcomes. By leveraging diverse data sources and robust analytical techniques, we can create a supportive learning environment that fosters student success. Additionally, incorporating behavioral models, like the Fogg Behavior Model, can further motivate students by providing them with timely alerts and personalized support. Looking ahead, it is essential for future research to refine these technological solutions and explore additional features that could enhance predictive capabilities. By continuously improving our understanding of the factors that contribute to student retention, we can ensure that every student has the opportunity to thrive academically. Ultimately, this initiative aims not just to reduce dropout rates but to empower students on their educational journeys, helping them realize their full potential and achieve their goals.

References

[1] Muhammad Adnan “Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models”, IEEE January 13, 2021.

[2] Mirna Nachouki , Elfadil A. Mohamed, Riyadh Mehdi, Mahmoud Abou Naaj "Student course grade prediction using the random forest algorithm: Analysis of predictors’ importance" Trends in Neuroscience and Education 33 (2023) 100214.

[3] Hanh Thi-Hong Duong1,Linh Thi-My Tran1,Huy Quoc To1,Kiet Van Nguyen1,2 "Academic performance warning system based on data driven-for higher education " Neural Computing and Applications (2023) 35:5819–5837.

[4] Improving students' retention using machine learning: Impacts and implications”Sandeep Trivedi,ScienceOpen Preprints, 2022.

[5] M. Skittou, M. Merrouchi and T. Gadi, "Development of an Early Warning System to Support Educational Planning Process by Identifying At-Risk Students," in IEEE Access, vol. 12, pp. 2260-2271, 2024.

[6] Lin, J. J., P. K. Imbrie, and Kenneth J. Reid. "Student retention modelling: An evaluation of different methods and their impact on prediction results." Research in Engineering Education Sysmposium (2009): 1-6.

[7] Guanin-Fajardo, J.H.; Guaña-Moya, J.; Casillas, J. Predicting Academic Success of College Students Using Machine Learning Techniques. Data 2024, 9, 60.

[8] Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. (2022). Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning Approach: A Case of Tanzanian's Secondary Schools. Applied Artificial Intelligence, 36(1). https://doi.org/10.1080/08839514.2022.2071406

[9] Kunchala, Vikas. Predicting Undergraduate Student Dropout Using Artificial Intelligence, Big Data and Machine Learning. MS thesis. University of Georgia, 2021

[10] Shia o, Yi-Tzone, et al. "Reducing dropout rate through a deep learning model for sustainable education: long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021." Smart Learning Environments 10.1 (2023): 55.

[11] Matz, Sandra C., et al. "Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics." Scientific Reports 13.1 (2023): 5705.

[12] Samuel, Ademola Olaide Adenubi1and Nathaniel. "Revolutionizing Education with Artificial Intelligence and Machine Learning: Personalization, Retention, and Resource Optimization."

[13] Freitas, Francisco A. da S., et al. "IoT system for school dropout prediction using machine learning techniques based on socioeconomic data." Electronics 9.10 (2020):1613.

[14] "Review on Educational Academic Performance Analysis and Dropout Visualization by Analyzing Grades of Student" Mansi Choudhari1, Saloni Rangari2, Pratham Badge3, Priyadarshini College of Engineering, Maharashtra, India.

[15] Mingyu Z, Sutong W, Yanzhang W, Dujuan W (2021) An interpretable prediction method for university student academic crisis warning. Complex Intell Syst 8(1):323–336.