STUDENT DROPOUT ANALYSIS

MINI PROJECT REPORT

Submitted by

Sanjay S **210701231**

Shriram Kumar A N **210701247**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

PROGRESS THROUGH KNOWLEDGE

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY, CHENNAI 600 025

MAY 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**Student Dropout Analysis**" is the bonafide work of **"Sanjay S (210701231) and Shriram Kumar A N (210701247)"** who carried out the work under my supervision.Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. M. RAKESH KUMAR, M.E., Ph.D.,
Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on __________________

Internal Examiner **External Examiner**

ABSTRACT

The escalating rates of student dropout across educational institutions have become a significant focus area in machine learning projects. Students face various challenges, such as demographic disparities, academic struggles, psychological stressors, health issues, teacher-student dynamics, and behavioral problems, leading to dropouts. To address this issue, this project employs predictive analytics to anticipate dropout occurrences using decision tree and random forest algorithms in Python. Decision trees provide a robust framework for classification by recursively partitioning data based on relevant features, creating a hierarchical tree structure to predict the target variable. Random forests enhance this approach by using multiple decision trees to improve prediction accuracy and reduce overfitting. This methodology not only identifies at-risk students but also enables educational institutions to intervene early with tailored support measures, thereby improving retention rates. Furthermore, the project integrates the Gemini API, offering personalized suggestions to students identified as at risk of dropping out. The API provides targeted recommendations and resources to address specific student challenges, fostering a supportive learning environment. By proactively addressing dropout risk factors through machine learning and the Gemini API, this project aims to enhance student retention and promote academic success.

ACKNOWLEDGEMENT

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

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Dr. M. RAKESH KUMAR, M.E., Ph.D.,** Assistant Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Sanjay S-210701231
Shriram Kumar A N-2107012347

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

EDA	Exploratory Data Analysis
SVM	Support Vector Machines
XGB	Gradient Boosting
KNN	K Nearest Neighbours
SVC	Support Vector Classifier
API	Application Programming Interface

CHAPTER 1
INTRODUCTION

1.1 GENERAL

Student dropout is a critical issue in educational institutions, leading to significant personal and societal consequences. Students often face challenges such as demographic disparities, academic struggles, psychological stress, health issues, teacher-student relationship problems, and behavioral issues, which can culminate in dropping out.

1.2 OBJECTIVE

The primary objective of this project is to predict student dropout occurrences using machine learning techniques, thereby enabling timely interventions. By leveraging decision tree and random forest algorithms in Python, this project aims to identify at-risk students early on. The system provides educational institutions with actionable insights to implement support measures, improving student retention rates.

1.3 EXISTING SYSTEM

Existing systems in educational analytics primarily focus on tracking academic performance and demographic information. While some use basic predictive models to flag potential dropouts, these systems often lack accuracy and fail to provide actionable insights. Most current approaches do not fully utilize the potential of advanced machine learning techniques to address the dropout issue comprehensively.

1.4 PROPOSED SYSTEM

The proposed system leverages decision tree and random forest algorithms in Python to predict student dropout occurrences by analyzing factors like family’s background, academic performance and health. Integrating the Gemini API, it offers tailored suggestions for at-risk students. This predictive analytics approach not only enhances accuracy in identifying at-risk students but also enables educational institutions to provide timely, personalized support, thereby improving student retention rates and fostering a supportive learning environment.

CHAPTER 2

LITERATURE SURVEY

This paper [1] discusses the factors influencing university dropout rates, highlighting the critical role of socio-economic and institutional variables. It emphasizes the need for advanced analytical methods to identify at-risk students and proposes ensemble learning methods for improved prediction accuracy. The study provides a comprehensive analysis of how demographic and academic factors interplay to influence dropout rates, offering valuable insights for predictive models.

Arulampalam, Naylor, and Smith [2] examine the effects of in-class variation and student rank on dropout probabilities. Their study employs a cross-sectional and time-series analysis, revealing that both individual academic performance and relative positioning in class significantly affect withdrawal decisions. These findings underscore the importance of personalized interventions based on academic data.

Breier [3] explores the financial considerations behind student dropout, suggesting that poverty plays a more significant role than previously thought. By re-conceptualizing financial difficulties, this study provides a nuanced understanding of economic barriers, which can be integrated into dropout prediction models to enhance their sensitivity to financial distress indicators.

Chimka's work [4][5] utilizes proportional hazards models to analyze graduation data, offering a statistical approach to understand dropout timelines. This method could be beneficial in developing dynamic prediction models that adjust based on students' progress over time, providing early warnings for potential dropouts.

Guimarães, Sampaion, and Sampaino [6] apply a bivariate probability model to investigate university dropout decisions in Brazil. Their research highlights the influence of both individual and systemic factors, suggesting that predictive models must account for diverse socio-cultural contexts to be effective.

Johnson [7] focuses on commuter college students, identifying specific factors that determine persistence versus dropout. This study highlights the unique challenges faced by non-residential students, which can inform tailored prediction models and intervention strategies.

Min et al. [8] utilize nonparametric survival analysis to study student retention, emphasizing the impact of self-efficacy and personality traits on academic success. Their longitudinal approach provides valuable data for developing more accurate and personalized dropout prediction systems.

Zeng, Hu, and Zhou [9] contribute to understanding how personality traits and self-efficacy beliefs affect academic achievement. Their findings suggest that incorporating psychological assessments into predictive models can improve their accuracy and effectiveness.

Smith and Naylor [10] analyze the impact of schooling on university performance, demonstrating that prior educational experiences significantly influence dropout rates. This study supports the inclusion of detailed educational histories in predictive analytics to better identify at-risk students.

Werblow [11] examines the effect of high school size on math achievement and dropout rates. His findings suggest that smaller school environments might reduce dropout rates, indicating that predictive models should consider school size and related variables.

De Witte and Haelermans [12] investigate the effect of institutional culture on dropout rates, highlighting how organizational factors contribute to student retention. Their research suggests that predictive models should include institutional variables to capture the broader educational environment's impact.

Dynarski [13] evaluates the effect of student aid on college attendance and completion, providing evidence that financial support can significantly reduce dropout rates. This study underscores the importance of financial variables in predictive models.

Heckman and LaFontaine [14] analyze trends in high school graduation rates, providing historical data that can inform dropout prediction. Their research suggests that long-term trends and demographic changes should be incorporated into predictive analytics.

Rumberger [15] offers a comprehensive analysis of why students drop out of high school, proposing actionable strategies to mitigate dropout rates. His insights into the root causes of dropout behavior are crucial for developing effective intervention measures within predictive models.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre-installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

FRONT END	HTML, CSS
BACK END	Python
FRAMEWORKS	Streamlit
SOFTWARES USED	Visual Studio, Jupyter Notebook

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

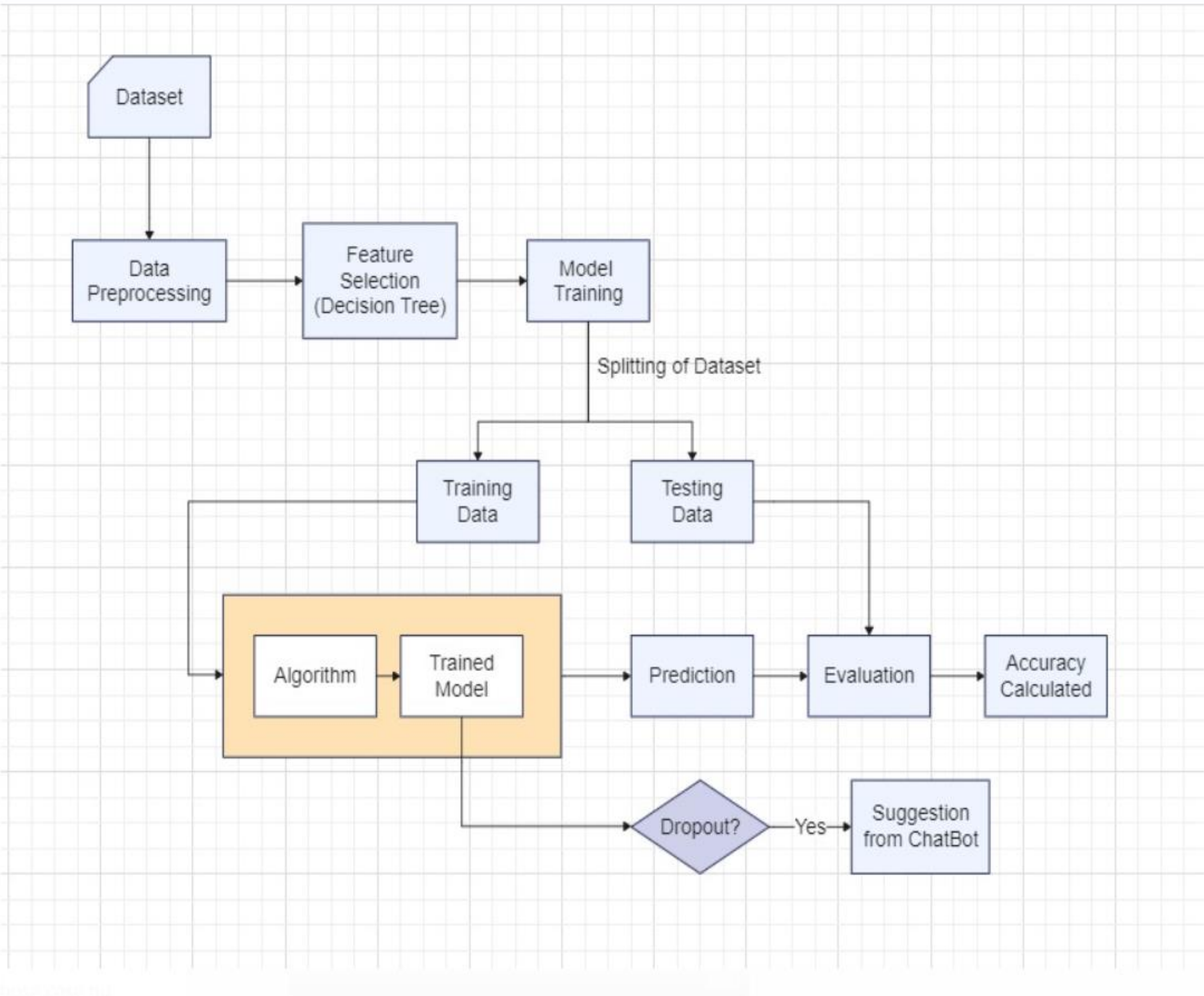

Fig 3.2.1 Architecture Diagram

PRE-PROCESSING:

Here, the emphasis is on preprocessing the dataset, specifically utilizing scikit-learn's LabelEncoder to encode categorical factors like gender, internet accessibility, and many more. Using LabelEncoder, it converts categorical values into numerical labels as iteratively traverses each column in the dataset, identifying categorical variables by their 'object' data type. This encoding makes machine learning methods that depend on numerical inputs easier to use and increases their accuracy when applied to categorical data. We also eliminate the dataset's null values in this stage. Additionally, we are utilizing the corr() method to determine the correlation between the features in the dataset. This correlation is then used to plot the classes on a heatmap so that patterns can be seen between them.

TRAINING SET:

In the domain of student dropout prediction, the training set undergoes meticulous data preparation and model training to ensure accurate predictions. The dataset includes various features encompassing students' personal information, study habits, free time, health status, and family background. These features are thoroughly analyzed and cleaned to ensure data quality, encompassing information such as demographic details, academic performance, extracurricular activities, parental occupation, and socioeconomic status. Upon importing the dataset and leveraging libraries like pandas, numpy, matplotlib, seaborn, and warnings for data manipulation and visualization, the dataset is loaded into a pandas DataFrame via the pd.read_csv() function. Subsequent steps involve comprehensive data cleaning, including handling missing values, addressing outliers, and encoding categorical variables for compatibility with machine learning algorithms. Advanced techniques such as logistic regression, decision trees, random forests are then employed for model training. The dataset is split into a training set and a validation set, with the former utilized for training the model. Throughout this process, the model learns the patterns and relationships indicative of dropout behavior, with algorithm experimentation and hyperparameter tuning being integral. Following model training, evaluation using the validation set assesses the model's accuracy in predicting dropout cases, computed through performance metrics like accuracy and precision

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA PRE-PROCESSING:

Here, the focus is on preprocessing the dataset, particularly encoding categorical variables like the gender, availability of the internet and many more using scikit-learn's LabelEncoder. It iterates over each column in the dataset, identifying categorical variables by their 'object' data type, and uses LabelEncoder to transform categorical values into numerical labels. This encoding facilitates machine learning algorithms that require numerical inputs, enhancing their effectiveness with categorical data for better accuracy. In this step we also remove the null values present in the dataset. Also we are using the corr() method for finding the correlation between the features present in the dataset which is later used for plotting the heatmap of the classes to find patterns among them.

4.1.2 FEATURE ENGINEERING:

Feature selection is performed using scikit-learn's SelectFromModel with a DecisionTreeClassifier. This process involves splitting the dataset into training and testing sets, initializing a SelectFromModel object with a DecisionTreeClassifier, fitting the model on the training data, and retrieving the selected features based on the information gain of the classes present in the dataset. By leveraging a decision tree model to identify important features, this approach aims to enhance the predictive power of the model by focusing on relevant features for classification tasks. The selected features are student free time, study hours, total amount of failures faced by the student in each subject, and number of days he/she is absent for school or college.

4.1.3 MODEL TRAINING:

This involves preparing the dataset for training a machine learning model to predict student dropout. It separates the dataset into features (X) and the target variable (y), splits the data into training and testing sets using scikit-learn's train_test_split function, with 80% for training and 20% for testing with random_state of 40.

4.1.4 MODEL SELECTION AND EVALUATION:

RandomForestClassifier from scikit-learn's ensemble module is used to train a machine learning model on the training data as it had better efficiency as compared to various algorithms like Logistic Regression, XGB classifier and KNeighbours classifier and various classifiers. RandomForestClassifier fits multiple decision tree classifiers on various sub-samples of the dataset and uses averaging to improve predictive accuracy and control overfitting. The model is trained on the features (X_train) and target variable (y_train) using the fit method.

CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

The code imports pandas (as pd), numpy (as np), matplotlib.pyplot (as plt), seaborn (as sns), and warnings, each serving essential functions. Pandas is used for data manipulation and analysis, particularly for DataFrames. Numpy provides support for arrays and mathematical functions. Matplotlib.pyplot is a plotting library, while seaborn is for statistical data visualization. Warnings are suppressed to avoid cluttering output. The 'dropout.csv' dataset contains student information relevant to dropout analysis, including student IDs, demographics, and academic performance metrics. These records offer insight into potential dropout predictors, guiding further analysis.

The exploratory data analysis (EDA) on the dropout dataset starts with loading the data and examining its initial rows to understand its structure. The dataset comprises 395 entries and 31 columns. A thorough inspection reveals that there are no missing values, but some columns contain non-integer values that need to be converted to numerical categorical data. This conversion is essential because machine learning models cannot be trained on categorical data directly. By using the LabelEncoder from the sklearn.preprocessing library, all object-type columns are transformed into integer values, which facilitates further analysis.

An essential aspect of the analysis is to understand the distribution of the target variable, 'dropout'. The dataset shows that 265 students did not drop out, while 130 did, leading to a dropout rate of around 32.91%. This distribution is visualized using seaborn's countplot, which confirms the slightly skewed nature of the dataset. This distribution is suitable for classification tasks. To ensure that the features are not highly correlated, the correlation matrix of the dataset is computed. The resulting matrix indicates that no features have a correlation higher than 0.65, which means all features can be retained for model training.

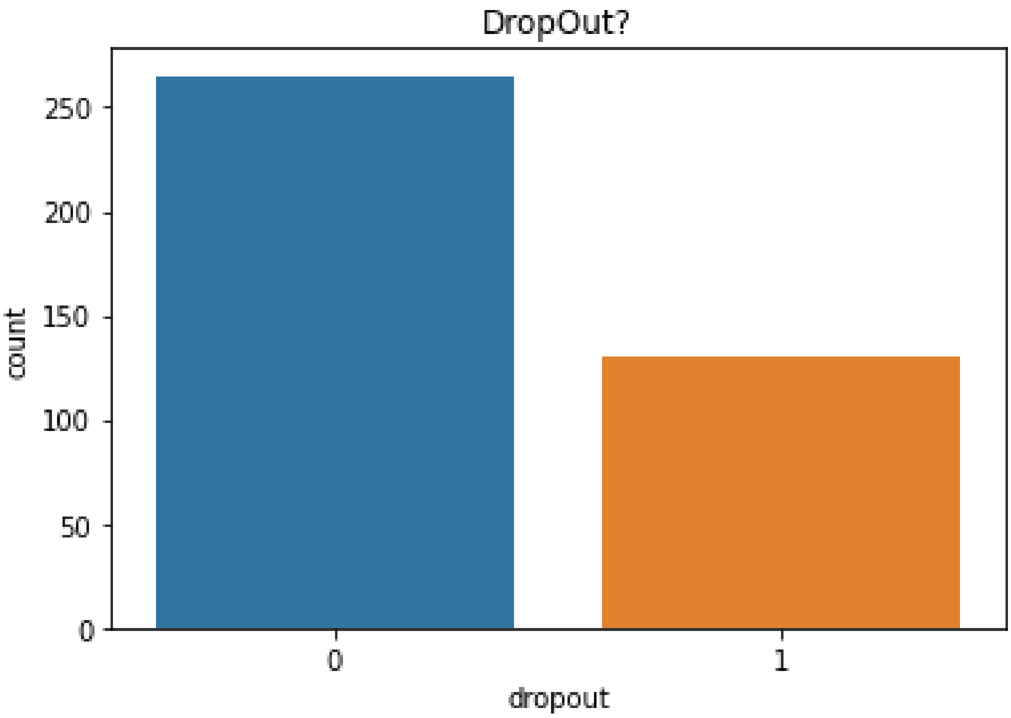

Fig 5.1.1 Count of Dropouts and Non-Dropouts

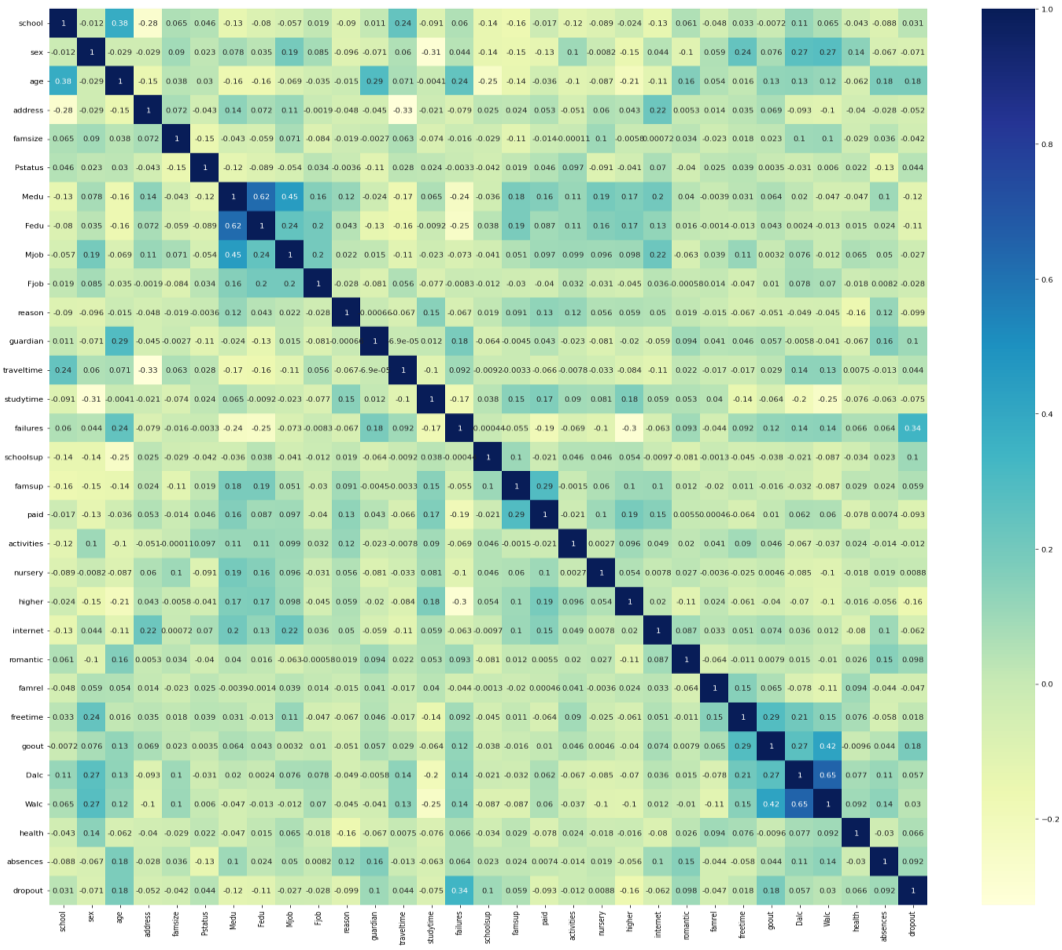

Fig 5.1.2 Heat Map

Feature selection is conducted using a DecisionTreeClassifier to pinpoint the most influential features on the target variable. Thirteen features are identified, including 'Mjob', 'Fjob', 'reason', and 'guardian', among others related to students' background and behavior. These selected features are used to form a new dataset, which is split into training and testing sets. To address class imbalance, the RandomOverSampler technique from the imblearn library is applied, ensuring an equal representation of both classes in the training data.

Various classification models are trained and evaluated using the training data. The models include Logistic Regression, Bernoulli Naive Bayes, K-Nearest Neighbors, Decision Tree, Support Vector Classifier, Random Forest, and XGBoost. These models are assessed using metrics such as accuracy, precision, recall, and F1 score. Among the models, the Random Forest Classifier performs the best, achieving an accuracy of 81.13% and an F1 score of 0.79. The model's high performance can be attributed to its ability to manage complex interactions between features effectively.

	Model	F1_score	Accuracy
0	LogisticRegression	0.564706	0.650943
1	BernoulliNB	0.568421	0.613208
2	KNeighborsClassifier	0.620000	0.641509
3	DecisionTreeClassifier	0.754717	0.754717
4	SVC	0.586667	0.415094
5	RandomForestClassifier	0.791667	0.811321
6	XGBClassifier	0.731707	0.751880

Fig 5.1.3 Accuracy Of Models

The evaluation of the Random Forest model is shown by a confusion matrix that is produced by Yellowbrick. The model's performance is examined in depth in the confusion matrix, which demonstrates that it correctly predicts the majority of both classes. The high recall and precision for each class show that the model does a good job of differentiating between students who will and won't drop out. This visual tool makes it easier to see where the model might be improved and to comprehend its advantages.

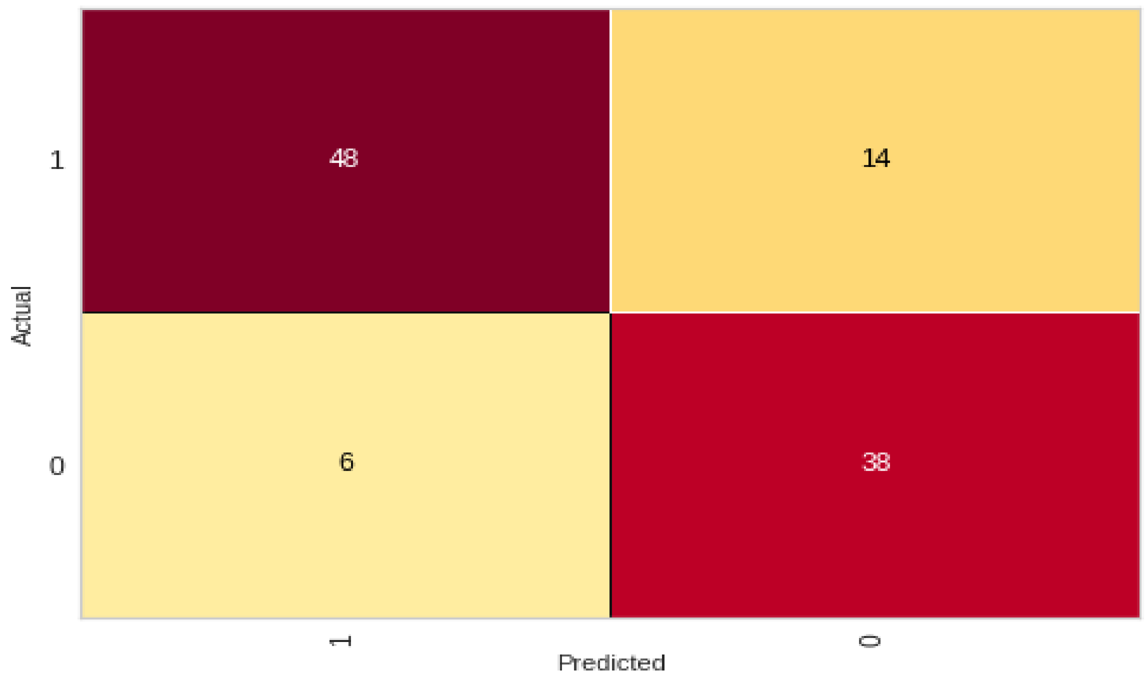

Fig 5.1.4 Confusion Matrix

The analysis is concluded with an assessment of the Random Forest model's feature relevance. Factors including "mother's job," "father's job," "reason" for selecting school, and "guardian" are important in predicting dropout. Teachers and legislators can use these data to identify pupils who are at-risk and carry out focused interventions. Joblib is used to save the LabelEncoder and the model's parameters, making deployment and forecasting easier.

In general, the EDA and model training emphasize how crucial it is to comprehend the underlying data and choose the appropriate features in order to produce precise predictions. The Random Forest model performs better than other models, which highlights its versatility and suitability for real-world situations. A thorough approach for handling comparable categorization issues in educational datasets is provided by the analysis.

5.2 OUTPUT SCREENSHOTS

Fig 5.2.1 Form for collecting details

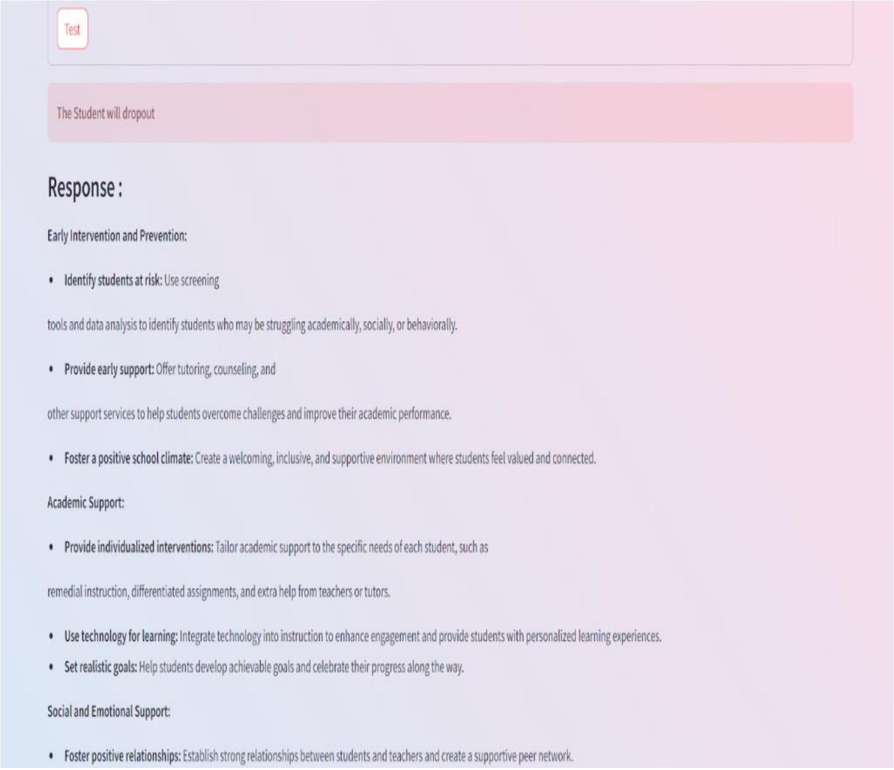

Fig 5.2.2 Response of Chatbot

CHAPTER 6
CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

This project underscores the power of machine learning techniques in addressing the pressing issue of student dropout rates within educational institutions. By meticulously collecting and analyzing data on various student demographics, academic performance metrics, and socio-economic factors, we have successfully developed a robust predictive model using sophisticated algorithms like decision trees and random forests. The model's exceptional accuracy on test data validates its effectiveness in identifying students at risk of dropout and facilitating timely interventions to support their academic journey.

The implementation of this predictive model has significant opportunities for educational establishments, as it enables them to proactively detect and support students who could be experiencing difficulties in their academic endeavors. Educational institutions can enhance their student retention rates and cultivate a more welcoming and encouraging learning atmosphere by implementing focused interventions and providing individualized support. This initiative essentially acts as a testament to the revolutionary power of machine learning and data-driven analysis in tackling intricate societal issues like student dropout rates. We can build a more equal and encouraging educational environment that enables every kid to flourish by utilizing innovation and technology.

6.2 FUTURE ENHANCEMENTS

Looking ahead, there are numerous opportunities for further advancement and refinement of this project. Future endeavors could explore the integration of additional data sources and features to enhance the model's predictive capabilities further. By incorporating more granular data, such as detailed behavioral insights and real-time academic performance, the model can become even more precise in identifying at-risk students.

Additionally, ongoing research and development efforts could focus on leveraging advanced machine learning algorithms and ensemble techniques to improve model performance and generalization across diverse student populations. Techniques such as deep learning, ensemble learning, and transfer learning could be explored to handle the complexities of dropout prediction in various educational contexts.

Furthermore, longitudinal studies and continuous monitoring of student progress could provide valuable insights into the long-term effectiveness of intervention strategies and the overall impact on student outcomes. By iteratively refining and optimizing the predictive model, educational institutions can continuously adapt and evolve their support mechanisms to meet the evolving needs of students and promote their academic success.

REFERENCES

[1] A. Araque, C. Roldán, & A. Salguero. (2009). Factors influencing university dropout rates. Computers & Education, 53(2), 563-574.

[2] W. Arulampalam, R. A. Naylor, & J. P. Smith. (2005). Effects of in-class variation and student rank on the probability of withdrawal: cross-section and time-series analysis for UK university students. Economics of Education Review, 24(3), 251–262.

[3] M. Breier. (2010). From “financial considerations” to “poverty”: towards a reconceptualisation of the role of finances in higher education student drop out. Higher Education, 60(6), 657–670.

[4] J. R. Chimka. (2002). Joint Statistical Meetings - Section on Quality & Productivity (Q&P). Proportional Hazards Models of Graduation (pp. 526–527). Retrieved-from: http://www.amstat.org/sections/SRMS/proceedings/y2002/files/JSM2002-001075.pdf

[5] J. R. Chimka, & L. H. Lowe. (2008). Interaction and survival analysis of graduation data. Educational Research and Review, 3(1), 29–32. Retrieved from http://www.academicjournals.org/ERR

[6] J. Guimarães, B. Sampaion, & Y. Sampaino. (2010). What is behind University Dropout Decision in Brazil ? A Bivariate Probability Model. The Empirical Economics Letters, 9(June), 601–608.

[7] J. Johnson. (1997). Commuter college students: What factors determine who will persist or who will drop out? College Student Journal, 31(3), 323.

[8] A. Tamir, E. Watson, B. Willett, Q. Hasan, and J.-S. Yuan, “Crime Prediction and Forecasting using Machine Learning Algorithms,” 2021. [Online]. Available: https://www.researchgate.net/publication/355872171

[9] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, “Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions,” *IEEE Access*, vol. 11, pp. 60153–60170, 2023, doi: 10.1109/ACCESS.2023.3286344.

[10] X. Zhang, L. Liu, L. Xiao, and J. Ji, “Comparison of machine learning algorithms for predicting crime hotspots,” *IEEE Access*, vol. 8, pp. 181302–181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

[11] R. Iqbal *et al.*, “An Experimental Study of Classification Algorithms for Crime Prediction.” [Online]. Available: www.indjst.org

[12] A. Krysovatyy, H. Lipyanina-Goncharenko, S. Sachenko, and O. Desyatnyuk, “Economic Crime Detection Using Support Vector Machine Classification.”

[13] S. Dynarski. (2003). Does aid matter? measuring the effect of student aid on college attendance and completion. The American Economic Review, 93(1), 279–288.

[14] J. J. Heckman, & P. A. LaFontaine. (2010). The American high school graduation rate: Trends and levels. The Review of Economics and Statistics, 92(2), 244–262.

[15] R. W. Rumberger. (2011). Dropping out: Why students drop out of high school and what can be done about it. Harvard University Press.