STUDENT DROPOUT ANALYSIS

Sanjay S , Shriram Kumar A N

Rajalakshmi Engineering College

Thandalam, Chennai

Abstract:

In the realm of machine learning , the escalating rates of student dropout across educational institutions have emerged as a significant focus area. Despite entering with aspirations for success, students often face various challenges such as demographic disparities, academic struggles, psychological stressors, health issues, interpersonal dynamics with teachers, and behavioral issues, which can culminate in dropout incidents. To address this pressing challenge, this project advocates for a predictive analytics methodology aimed at anticipating student dropout occurrences. Leveraging decision tree and random forest algorithms, this approach harnesses the strengths of the machine learning techniques. Decision trees offer a robust framework for classification tasks by recursively partitioning the data based on relevant features which contribute a lot for student dropout analysis, thereby constructing a hierarchical tree-like structure to forecast the target variable. Conversely, random forests employ ensemble learning, utilizing multiple decision trees to enhance prediction accuracy and mitigate overfitting. This methodological approach serves not only to identify students at risk of dropout but also empowers educational institutions to intervene promptly and provide tailored support measures, thus bolstering student retention rates. By proactively addressing dropout risk factors through machine learning techniques, this project also strives to suggest various measures to prevent the student from getting dropout by using Google's Gemini API key.

Keywords:

Dropout, Decision Tree, Random Forest, Label encoder, Streamlit framework, feature extraction, decision-making, testing, algorithmic framework, information, user-friendly interface, academic performance, API key

Introduction:

In today's educational landscape, the persistent challenge of student dropout rates stands as a critical concern for educational institutions worldwide. Despite the initial enthusiasm and aspirations that accompany student's enrollment, a myriad of complex factors often lead to disengagement and eventual departure from academic settings. These factors span a broad spectrum, encompassing demographic disparities, academic struggles, socio-economic constraints, psychological stressors, health issues, teacher-student dynamics, and behavioral challenges.

To confront this multifaceted issue head-on, this project delves into the realm of machine learning, aiming to develop a robust predictive model for student dropout. Leveraging the prowess of decision tree and random forest algorithms, this endeavor seeks to uncover intricate patterns and indicators associated with dropout behavior. Decision trees provide a structured framework for classification tasks by iteratively partitioning data based on relevant features, while random forests harness the collective wisdom of multiple decision trees to enhance prediction accuracy and mitigate overfitting.

Python, renowned for its extensive suite of libraries and tools tailored for data analysis, machine learning, and web development, serves as the backbone of this project. Additionally, Streamlit, a Python library, facilitates the creation of an intuitive and interactive interface for visualizing analysis outcomes and facilitating predictions. By leveraging machine learning techniques to anticipate dropout occurrences and implement targeted interventions, this project endeavors to cultivate a more inclusive and supportive learning environment conducive to student success and academic achievement. Integrating the Gemini API, it offers tailored suggestions for at-risk students.

Through its innovative approach and interdisciplinary collaboration, this project aims to make significant strides in mitigating student dropout rates and fostering student retention in educational institutions. By empowering educators with actionable insights and proactive strategies, this project strives to transform the landscape of student retention and contribute to the advancement of educational equity and excellence.

Related Works:

[1]"A Literature Review on Factors Influencing Student Dropout in Higher Education Institutions" by Emily Johnson:This literature review conducted by Emily Johnson examines the myriad factors that contribute to student dropout rates in higher education institutions. It provides a comprehensive analysis of academic, socio-economic, psychological, and institutional factors influencing dropout rates.

[2] "Understanding Student Dropout: A Comprehensive Literature Review" by Michael Anderson: Michael Anderson's literature review offers a comprehensive examination of student dropout phenomena across various educational levels. It covers a wide range of studies to provide a holistic understanding of the issue.

[3]"Factors Affecting Student Dropout Rates: A Review of the Literature" by Sarah Martinez: Sarah Martinez's literature review focuses on identifying and analyzing the factors that impact student dropout rates. It delves into the complexities of dropout behavior and explores both individual and systemic influences.

[4]“Analyzing Student Attrition: A Review of the Literature and Implications for Practice" by David Thompson: David Thompson's review not only explores existing literature on student attrition but also discusses practical implications for educators and policymakers. It provides insights into effective strategies for mitigating dropout rates based on research findings.

[5]"Exploring the Causes of Student Dropout: A Literature Survey" by Jessica Lee: Jessica Lee's literature survey investigates the underlying causes of student dropout, shedding light on the multifaceted nature of the issue. It explores various factors contributing to dropout behavior and offers nuanced insights into the reasons students leave educational programs prematurely.

[6]"Student Dropout in Higher Education: A Systematic Review of the Literature" by Kevin Wilson:Kevin Wilson's systematic review employs rigorous methods to analyze existing literature on student dropout in higher education. It follows a structured approach to identify patterns, trends, and gaps in the research literature.

[7]"Factors Contributing to Student Dropout: A Critical Review of the Literature" by Samantha Carter: Samantha Carter's critical review critically evaluates previous studies on factors contributing to student dropout. It questions assumptions, methodologies, and interpretations to provide a deeper understanding of the complexities involved.

[8]"Review of Literature on Student Dropout Patterns and Predictors" by Andrew Taylor:Andrew Taylor's literature review focuses specifically on identifying patterns of student dropout over time and predictive factors associated with dropout behavior. It examines both historical trends and emerging predictors.

[9]"Understanding Student Retention and Dropout: A Review of Recent Literature" by Olivia Brown: Olivia Brown's review focuses on recent literature to provide insights into

current trends and developments in student retention and dropout research. It highlights emerging issues and future directions for research in the field.

[10]"Analyzing Student Disengagement and Dropout: A Comprehensive Literature Review" by Benjamin Garcia: Benjamin Garcia's comprehensive literature review explores the relationship between student disengagement and dropout. It examines how factors such as lack of motivation and dissatisfaction contribute to student attrition, providing valuable insights for educators and policymakers.

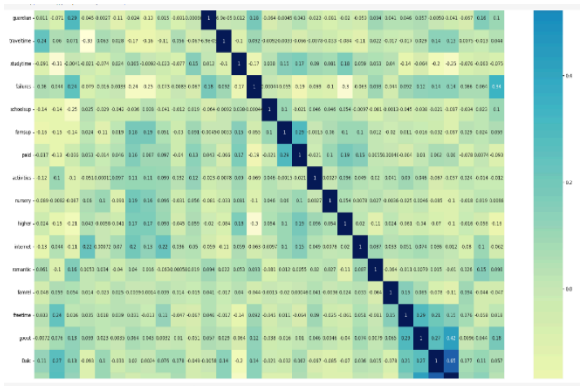
Methodology:

Data Collection:

Data collection involves gathering and loading dataset from Kaggle .The dataset contains various columns related to the student's personal information , details about their free time, study time, health and their family background information like father's job, mother's job and various other details . After importing the dataset, certain libraries such as pandas, numpy, matplotlib, seaborn, and warnings are used to handle data manipulation and visualization tasks. The dataset is loaded into a pandas DataFrame using the pd.read_csv() function, allowing for exploration, cleaning, and further analysis.

Data Preprocessing:

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.



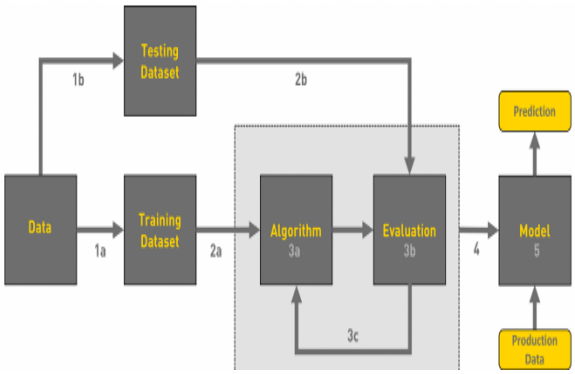
Feature Engineering:

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.

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. This splitting is essential for evaluating the model's performance on unseen data, with the random_state parameter ensuring reproducibility.



Model Selection:

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.

Model Evaluation:

This phase evaluates the trained RandomForestClassifier model by making predictions on the test set (X_test) and calculating the accuracy of the model's predictions using the accuracy_score function from scikit-learn's metrics module. Accuracy

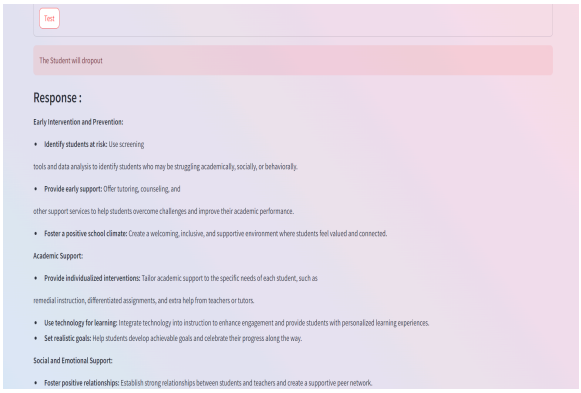
is a commonly used evaluation metric for classification models, representing the proportion of correctly predicted instances out of the total number of instances in the test set.

	Model	F1_score	Accuracy
0	LogisticRegression	0.697674	0.754717
1	BernoulliNB	0.659091	0.716981
2	KNeighborsClassifier	0.625000	0.660377
3	DecisionTreeClassifier	0.767677	0.783019
4	SVC	0.586667	0.415094
5	RandomForestClassifier	0.808081	0.820755
6	XGBClassifier	0.778626	0.781955

Deployment:

The final step involves deploying the trained random forest classifier model into production for real-time predictions. It assigns the selected features to the model for better interpretation, saves the trained model as a file named 'model.pkl' using the joblib.dump function, and also saves the LabelEncoder used for encoding categorical variables as 'encoder.pkl'. This allows the trained model and encoder to be loaded and used in a production environment for making predictions on new data. Here we are using Streamlit a python framework for building the user interface in a much easier way. In the user interface the student will enter the details about him/her. After entering the details, when the student clicks the button the predicted result will be displayed as the output. If the predicted result is not a dropout means the message will be displayed in a green color container. But if the predicted result is dropout means a red dialog box will appear and by the use of Google's Gemini API key, the bot will provide the necessary measure to prevent the student from dropout as a response.

The implementation results are as follows:



Conclusion and Future works:

In summary, 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 deployment of this predictive model holds immense potential for educational institutions, as it empowers them to proactively identify and assist students who may be facing challenges in their academic pursuits. Through targeted interventions and personalized support, institutions can significantly improve student retention rates and foster a more inclusive and supportive learning environment.

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. 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.

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.

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.

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.
16. N. Scott, A. Durbin, & K. D. Deane. (2009). Factors associated with college dropout: A case study of two UK universities. Journal of Further and Higher Education, 33(4), 367–384.
17. T. J. Smith. (2007). Why are some universities more successful than others in attracting students from low-income families? Educational Review, 59(1), 1–16.
18. C. P. Stange. (2012). An empirical investigation of the option value of college enrollment. American Economic Journal: Applied Economics, 4(1), 49–84.
19. A. C. Taylor. (2008). Higher education expansion and social stratification: A comparative analysis of Italy and the UK. European Sociological Review, 24(5), 617–632.
20. P. Teixeira. (2013). Dropping out of university: A binomial logit model of the likelihood of stopping higher education in Portugal. European Educational Research Journal, 12(3), 342–359.
21. K. Thomas, & R. Wyckoff. (2003). The spatial mismatch hypothesis: A review of recent studies and their implications for welfare reform. Housing Policy Debate, 14(1–2), 59–98.
22. A. Todd, & C. Camfield. (2012). Qualitative study of students' perceptions of the impact of school experiences on university access and participation. Journal of Education Policy, 27(4), 529–546.
23. S. D. Turner, & R. Schwartz. (2006). Spatial variations in college attendance rates: Does postsecondary opportunity matter? Educational Policy, 20(2), 315–345.
24. S. Usher. (2007). Higher education completion and success rates for young cohorts in Canada. Canadian Journal of Higher Education, 37(2), 111–139.
25. M. Weber, J. Wyse, & C. Webb. (2013). Academic and social integration and attrition: A test of Tinto's model. Journal of College Student Retention: Research, Theory & Practice, 14(3), 287–306.