SUPERVISED MACHINE LEARNING MODEL FOR PREDICTING STUDENT DROPOUT

Mr. SAGAR CHOUDHARY, AYUSHI
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
QUANTUM SCHOOL OF TECHNOLOGY, ROORKEE, UTTARAKHAND

ABSTRACT

This research delves into the comparative performance of several machine learning algorithms in predicting student academic outcomes. By analyzing a rich dataset with attributes related to academic history, socio-economic status, and personal demographics, we employ Logistic Regression, Decision Trees, Random Forests, k-Nearest Neighbors, SVM, AdaBoost, and XGBoost. The study emphasizes preprocessing strategies such as scaling and the use of SMOTE to handle imbalanced data. Our findings reveal critical predictors of academic success, offering valuable guidance for educational policymakers to enhance student support and performance monitoring systems.

Keywords: Dropout, Prediction, Machine Learning , Supervised Learning

INTRODUCTION

Education is foundational to the development of individuals and nations alike. As **Nwabueze (2011)** articulated, education is a critical sector of the economy that produces the workforce needed for socioeconomic, political, and cultural advancement. It encompasses the systematic cultivation of intellect, skills, and character through structured learning and study. The significance of education extends beyond personal growth; it serves as a catalyst for national development and societal progress. Over the past six decades, various government initiatives have successfully increased school enrolment rates through diverse educational programs. However, these efforts have not consistently translated into high completion rates, leading to a persistent issue of school dropout. This phenomenon is particularly concerning among youth, as it hinders their professional and personal development. The dropout issue is a major concern in both developed and developing nations, including Nigeria (**Udomah et al., 2020**).

The decision to drop out of school is influenced by a myriad of factors. Critical determinants include academic aspirations, self-assessment capabilities, and academic performance (**Robbins et al., 2004**). Additional factors such as institutional commitment, social support networks, involvement in school activities, and financial constraints also play significant roles. The repercussions of dropping out extend beyond the individual, adversely affecting educational institutions and the broader economy (**Nurmalitasari et al., 2023**).

Individuals who drop out of school face numerous challenges, including reduced earning potential, higher unemployment rates, and increased susceptibility to criminal behavior. Societal impacts include increased social costs and decreased economic productivity (**Real et al., 2018**).Predictive models offer a strategic approach to mitigating the dropout crisis. By identifying at-risk students early, these models enable targeted interventions that can help keep students on track (**Jay et al., 2020**). The insights gained from predictive analytics can inform policy decisions, optimize resource allocation, and ultimately reduce dropout rates, fostering a more educated and capable population. Developing an effective predictive model requires careful selection and preparation of the dataset. This involves choosing relevant features that reflect both academic and non-academic factors influencing dropout rates, ensuring data quality, and addressing potential biases. Ethical considerations are paramount to prevent unintended consequences (**Nurdaulet et al., 2021**).

MATERIALS AND METHODS

This study utilized a dataset with various student attributes, including demographic information, academic performance, social involvement, and institutional factors. Data preprocessing involved imputing missing values, encoding categorical variables, and normalizing numerical features. To address class imbalance various model were trained and evaluated based on accuracy, and confusion matrix.The system chart is shown in figure 1 below. It shows the system chart and procedures utilized in this investigation are shown below:

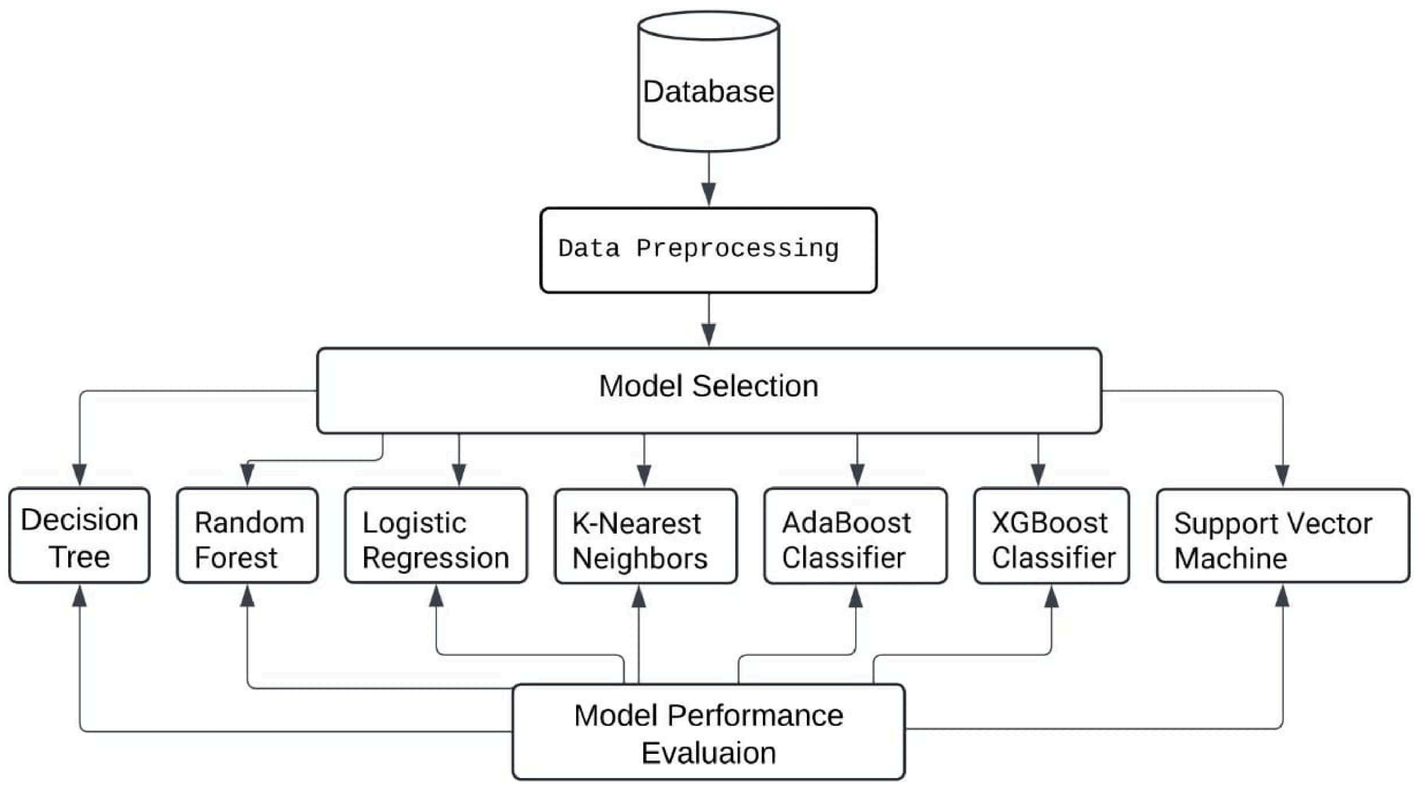

Figure 1: System Chart

Dataset Description

This dataset provides a comprehensive view of students enrolled in various undergraduate degrees offered at a higher education institution. It includes demographic data, social-economic factors and academic performance information that can be used to analyze the possible predictors of student dropout and academic success. This dataset contains multiple disjoint databases consisting of relevant information available at the time of enrollment, such as application mode, marital status, course chosen and more. Additionally, this data can be used to estimate overall student performance at the end of each semester by assessing curricular units credited /enrolled /evaluated/ approved as well as their respective grades. Finally, we have unemployment rate, inflation rate and GDP from the region which can help us further understand how economic factors play into student dropout rates or academic success outcomes. This powerful analysis tool will provide valuable insight into what motivates students to stay in school or abandon their studies for a wide range of disciplines such as agronomy, design, education nursing journalism management social service or technologies.

#	Column	Non-Null Count	Dtype
0	Marital status	4424 non-null	int64
1	Application mode	4424 non-null	int64
2	Application order	4424 non-null	int64
3	Course	4424 non-null	int64
4	Daytime/evening attendance	4424 non-null	int64
5	Previous qualification	4424 non-null	int64
...			
33	GDP	4424 non-null	float64
34	Target	4424 non-null	object

Table 1: Dataset Description

Data Preprocessing

Data processing is a crucial step in preparing the dataset for machine learning models. It includes several key tasks to ensure the data is clean, integrated, and in a suitable format for analysis.

Data Cleaning

It involves removing unnecessary columns and rows, correcting inaccuracies, and handling missing values. This ensures the data is accurate and reliable for the models.

S.No.	MaritalStatus	ApplicationMode	ApplicaOrder	...	UnemploymentRate	InflatioRate	GDP	Target
0	1	8	5	...	10.8	1.4	1.74	0
1	1	6	1	...	13.9	-0.3	0.79	2
2	1	1	5	...	10.8	1.4	1.74	0
3	1	8	2	...	9.4	-0.8	-3.12	2
4	2	12	1	...	13.9	-0.3	0.79	2

Table 2: Transformed dataset

After transforming the data to an analysis-ready format, we used a heat map to visualize the correlation between different features in the dataset. A heat map is a graphical representation that uses color coding to represent the values of correlations, making it easier to identify patterns and relationships between variables. The heat map revealed significant correlations between various features, helping us understand which factors might influence student dropouts the most.The correlation heatmap shown in Figure 2 depicts the correlation between the variables.

Exploratory Data Analysis

Once the data was transformed and correlations were examined, we proceeded with Exploratory Data Analysis (EDA). EDA is crucial for understanding the underlying patterns and distributions within the dataset. During EDA, we generated histograms for numerical features to visualize their distributions. Histograms provide a clear view of how data points are spread across different values and help identify any skewness or anomalies in the data. . Figure 3 shows the histogram for each numerical attribute.

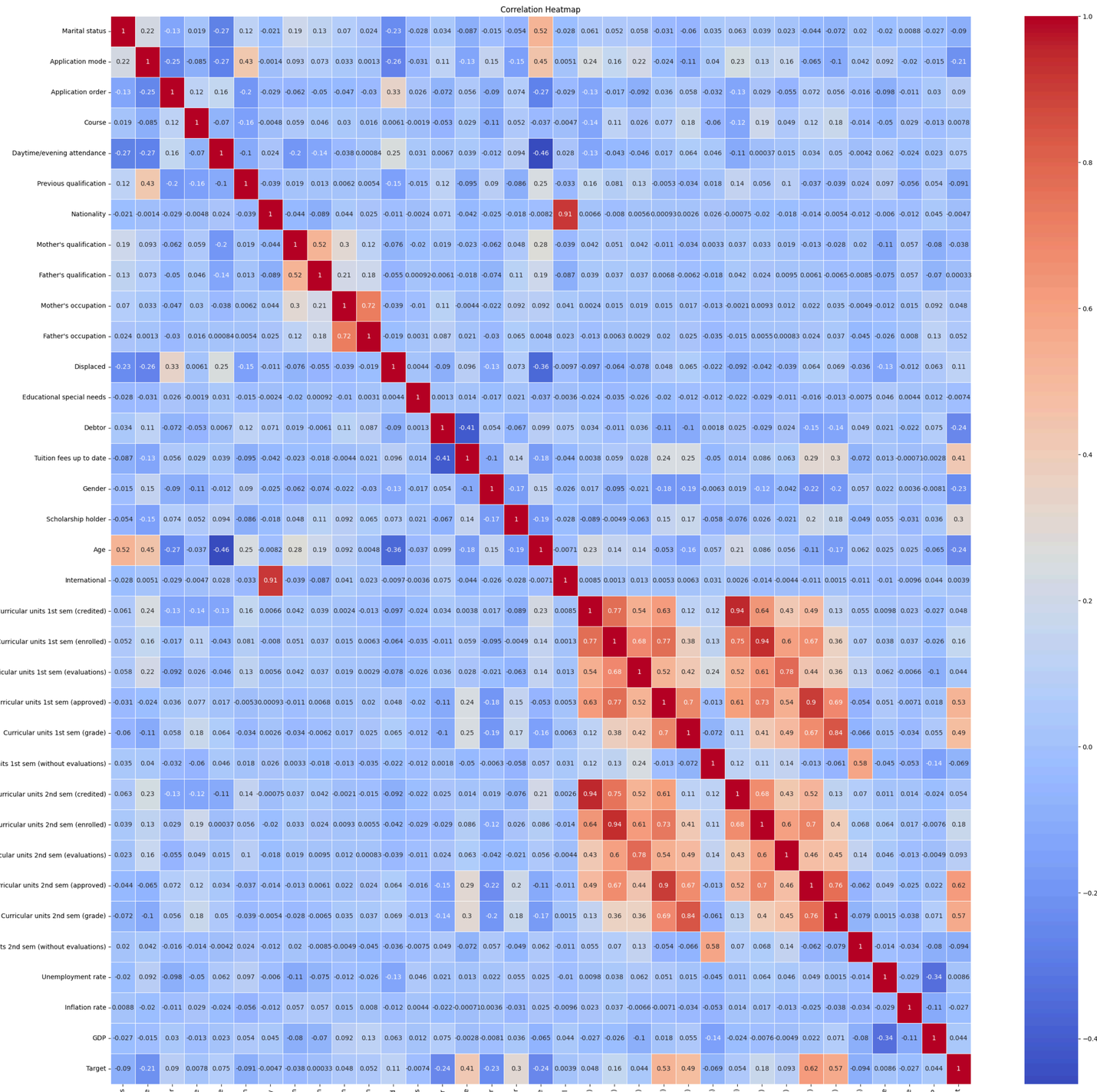

Figure 2: Correlation among the variables

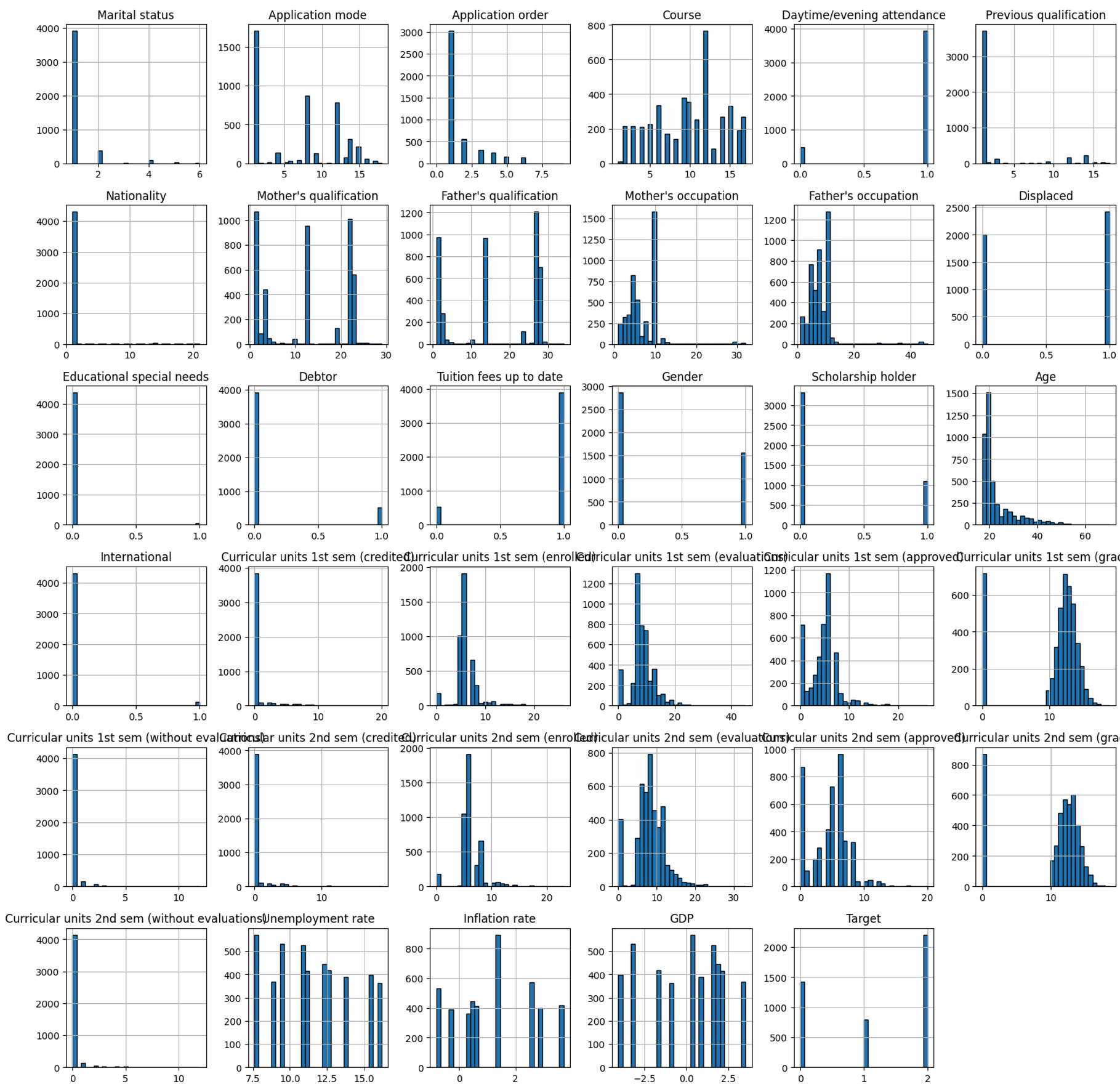

Figure 3: A histogram for each numerical attribute

Model Selection

It is a crucial step in the ml pipeline, where the goal is to identify the most suitable algorithm to accurately predict the target variable based on the given data. This involves evaluating various models, considering their performance & choosing the best one for the task.
In our study, we evaluated several machine learning models to predict student dropouts. Here is a brief overview of each model used:

1. **Decision Tree Classifier (dtree):** A non-parametric model that splits the data into subsets based on the feature values, creating a tree-like structure.
2. **Random Forest Classifier (rfc)**: An ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.
3. **Logistic Regression (lr)**: A statistical model that predicts binary outcomes by estimating probabilities using a logistic function.
4. **K-Nearest Neighbors (knn)**: A simple, instance-based learning algorithm that classifies a sample based on the majority class among its k-nearest neighbors.
5. **AdaBoost Classifier (abc)**: An ensemble method that combines multiple weak classifiers to create a strong classifier, improving prediction accuracy.
6. **XGBoost Classifier (xbc)**: An optimized gradient boosting framework that builds additive models in a sequential manner to enhance performance.
7. **Support Vector Machine (svm)**: A powerful model that finds the optimal hyperplane to separate classes in the feature space.

Model Evaluation

Model evaluation is a critical phase in machine learning, essential for assessing the performance of different models. In our study, we divided the dataset into training and testing sets with a 70-30 ratio. The training set, representing 70% of the data, was used to train the models, while the testing set, accounting for the remaining 30%, was utilized to evaluate their performance. This evaluation process allows us to gauge how well each model generalizes to unseen data, helping us understand their strengths and weaknesses. Ultimately, the aim is to select the most effective algorithm for predicting student dropouts, ensuring reliable and accurate predictions.

Methodology & Evaluation Metrics

To assess the performance of each machine learning model in predicting student dropout, we employed a standard methodology involving the division of the dataset into training and testing sets as mentioned. Additionally, accuracy a commonly used metric in classification tasks, was employed to evaluate the performance of each model. It measures the proportion of correctly classified instances out of the total instances in the testing set. The accuracy score is calculated using the formula:

- Accuracy = (Number of correct predictions) / (Total number of predictions)

RESULTS AND DISCUSSION

Table 3: Dataset table in the study

Data Set	Dropout (0)	Successful (1)	Total
Train	101	423	524
Test	43	186	229
Sum	144	609	753

Seven hundred and fifty-three 753 of the 800 data points collected between 2018 and 2023 were used in the experiments that were conducted. Learning on Decision Tree, Random Forest, Logistic Regression, KNN, AdaBoost, XGBoost, and SVM was achieved by dividing the collected and pretreated data into learning and test datasets in a 70-30 ratio. We apply a tester to the trained model in order to evaluate the accuracy of the prediction. Each model was trained and tested, and their performances were measured using appropriate metrics to determine the most accurate model for predicting student dropouts.

The accuracy scores of the models vary across the board. Gradient Boosting Machine achieved the highest accuracy of 82.15%, indicating its superior performance in making accurate predictions. Meanwhile, Decision Tree and Logistic Regression both attained an accuracy of 69.38%, suggesting relatively weaker predictive capabilities compared to other models. Moreover, the accuracy of LightGBM is not available in the provided data.Overall, the findings suggest that the ensemble methods, particularly Soft Voting, are well-suited for predicting student dropout based on the provided dataset.

Table 4: Summary of the Performance Evaluation

Model	Accuracy (%)
Decision Tree	69.38
Random Forest	80.56
Logistic Regression	78.08
K-Nearest Neighbors (KNN)	69.38
AdaBoost	77.18
XGBoost	79.89
Support Vector Machine (SVM)	77.06
Ensemble (Soft Voting)	82.15
Ensemble (Hard Voting)	81.02

Table 5: Execution time in seconds

Model	Execution Time
Decision Tree	43.5 seconds
K-Nearest Neighbors (KNN)	43.7 seconds
Naive Bayes	43.8 seconds
Logistic Regression	44.3 seconds
Artificial Neural Network	44.4 seconds
Support Vector Machine	44.4 seconds
Random Forest	44.6 seconds
Gradient Boosting Machine	45.8 seconds
XGBoost	46.6 seconds
LightGBM	47.7 seconds

CONCLUSION

In conclusion, the early prediction of student dropout is crucial for academic institutions to provide timely intervention and support, ultimately improving students' success rates. This study utilized various machine learning techniques to predict academic dropout among students, including Decision Trees, Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, and Artificial Neural Networks. Through the implementation of these predictive models, course advisors, organizations, and universities can proactively assess students' performance and implement effective interventions to enhance academic outcomes.
Among the models evaluated, the Logistic Regression Model emerged as the most effective in predicting student dropouts. Its superior performance underscores its potential utility as a tool for early identification and intervention. However, it's essential to note that the accuracy of the proposed model could be further enhanced by re-evaluating it with additional datasets, potentially drawn from extensive academic repositories or Big Datasets.
In summary, the findings of this study highlight the significance of leveraging machine learning techniques for early prediction of student dropout, enabling institutions to tailor interventions and support mechanisms to meet students' needs effectively. By continuing to refine and optimize predictive models, academic institutions can make significant strides in enhancing student retention and success.

REFERENCES

1. **Real, A. C., Oliveira, C. B., & Borges, J. L. (2018). "Using Academic Performance to Predict College Students Dropout: A Case Study".** https://doi.org/10.1590/S1678-4634201844180590

2. **Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). "Do psychosocial and study skill factors predict college outcomes? A meta-analysis". Psychological Bulletin, 130(2), 261–288.** https://doi.org/10.1037/0033-2909.130.2.261

3. **Nwabueze, A. I. (2011). "Achieving MDGs through ICTs Usage in Secondary Schools in Nigeria: Developing Global Partnership with Secondary Schools." Germany: Lambert Academic Publishing.**

4. **Udomah, N. G., & Archibong, U. I. (2020). "Psychological Factors and Dropout Tendency of Year One Students in Schools of Nursing, Akwa Ibom State, Nigeria". International Journal of Research in Education and Management Science, 3(2).**

5. **Nurmalitasari, N., Zalizah, A. L., & Faizuddin, M. N. (2023). "Factors Influencing Dropout Students in Higher Education". Education Research International, 2023, 1-13.** https://doi.org/10.1155/2023/7704142.

6. **Jay, S. G., Allemar, J. P., & Ramcis, N. V. (2020). "Predicting Students' Dropout Indicators in Public School using Data Mining Approaches". International Journal of Advanced Trends in Computer Science and Engineering, 9(1). Available online:** http://www.warse.org/IJATCSE/static/pdf/file/ijatcse110912020.pdf https://doi.org/10.30534/ijatcse/2020/110912020

7. **Nurdaulet, S., Alibek, O., Yershat, S., & Shirali, K. (2021). "Prediction of Student's Dropout from a University Program". International Conference on Electronics Computer and Computation (ICECCO).** https://doi.org/10.1109/ICECCO53203.2021.9663763