**PREDICTING STUDENT DROPOUT RATES**

**Team Name: Tech Titans**

**Team Member 1: Roshini Priya K**

**Team Member 2: Pooja Sree M**

**INTRODUCTION**

This project's goal is to use machine learning models to forecast the possibility of student dropout based on a variety of parameters. The ultimate objective is to create a prediction model that would enable educational establishments to take early action and lower the dropout rate of students.

**PROBLEM STATEMENT**

An educational institution wants to predict the likelihood of students dropping out before completing their course or degree program. By identifying at-risk students early, the institution can provide targeted interventions and support. Your task is to develop a machine learning model that predicts student dropout risk based on various factors such as academic performance, attendance, socio-economic background, and extracurricular participation.

**OBJECTIVE**

To assess several machine learning algorithms that are used to forecast student dropout rates.To determine which model, balancing precision and recall, offers the best accuracy and choose the top-performing model for deployment and adjust the others.

**DATA OVERVIEW**

The dataset contains features such as:

* StudentID, Age, Gender, AttendanceRate, GPA
* ParentalEducation, FamilyIncome, ExtracurricularParticipation
* SchoolType, DistanceToSchool, PreviousSuspensions
* SupportServicesUsed, Dropout (Target variable)

Key preprocessing tasks included handling missing values, converting categorical features to numerical, and normalizing the data.

**TRAINING/TEST DATA SPLIT**

The dataset was split into training and testing sets, adhering to a conventional 80/20 split. This ratio ensures sufficient data for training while reserving enough data for reliable model evaluation. This ensures that the models are trained on a significant portion of the data while being evaluated on unseen data.

**MACHINE LEARNING ALGORITHM USED**

The following machine learning models were trained and evaluated:

* Logistic Regression: Linear classification model.
* Decision Tree: A tree-based model.
* Random Forest: An ensemble model with grid search.
* Gradient Boosting: An iterative ensemble model.
* K-Nearest Neighbour (KNN): A distance-based classifier.
* Support Vector Machine (SVM): A hyperplane-based classifier.

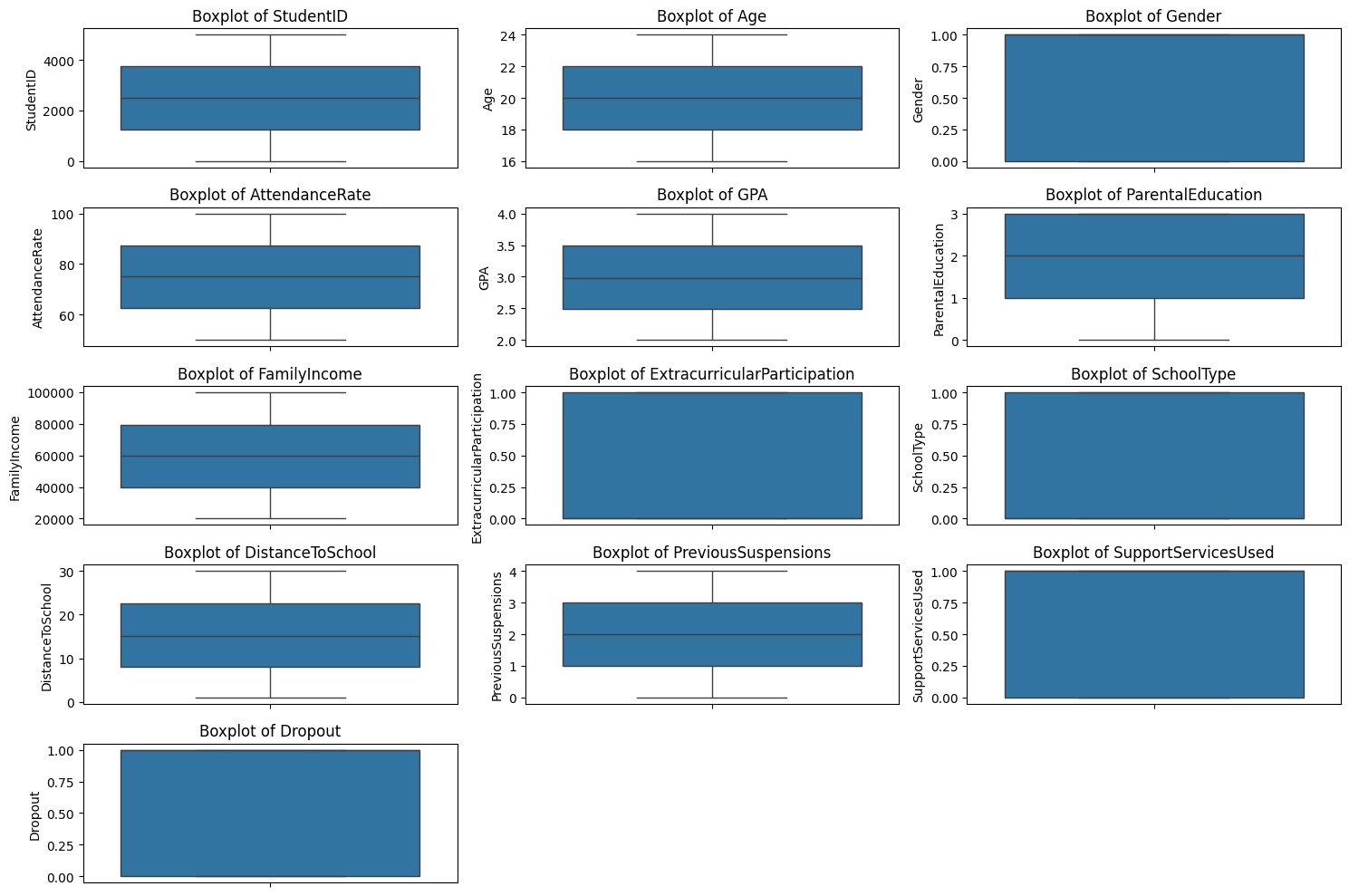
**PERFORMANCE METRICS**

The models were evaluated using the following metrics:

* Accuracy: Percentage of correct predictions.
* Precision: The proportion of positive identifications that were actually correct.
* Recall: The proportion of actual positives that were correctly identified.
* F1-Score: The harmonic mean of precision and recall.
* AUC-ROC: Area under the ROC curve, measuring classifier performance across all thresholds.

**OUTLIER REMOVAL**

Boxplots were generated for all numeric features to identify outliers. Based on this, no explicit outlier removal was done as the focus was on training robust models that could handle noise.



**DIMENSIONALITY REDUCTION**

To reduce dimensionality, Principal Component Analysis (PCA) was applied, retaining 95% of the variance in the data. The highly correlated features identified by the correlation matrix were dropped before applying PCA:

* FamilyIncome
* PreviousSuspensions
* SupportServicesUsed

**ACCEPTABLE ERROR RATES**

Given that student dropout prediction is a sensitive topic, an accuracy rate below 50% is considered unsatisfactory. Future work may explore tuning hyperparameters or gathering more data to improve model performance. The models displayed similar accuracy and AUC-ROC scores around 48-50%, indicating no model significantly outperformed others.

**MODEL VERSIONING**

The models were saved using Joblib and Pickle for versioning:

* KNN model: Saved as knn\_model.pkl
* Each version of the model will be saved with the training date to allow for future comparisons when models are retrained or updated.

**RESULT SUMMARY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MODEL | ACCURACY | PRECISION | RECALL | F1-SCORE | AUC-ROC |
| Logistic Regression | 49.50% | 0.50 | 0.59 | 0.54 | 0.49 |
| Decision Tree | 50.00% | 0.50 | 0.47 | 0.48 | 0.50 |
| Gradient Boosting | 49.40% | 0.50 | 0.92 | 0.65 | 0.49 |
| KNN | 50.30% | 0.50 | 0.50 | 0.50 | 0.50 |
| SVM | 47.80% | 0.48 | 0.54 | 0.51 | 0.48 |
| Random Forest | 48.30% | 0.48 | 0.46 | 0.47 | 0.48 |

**CONCLUSION**

The current models show limited predictive power, with accuracy hovering around 50%. Future work should focus on gathering more relevant features, fine-tuning models, and possibly using advanced techniques like deep learning or ensemble methods to improve the prediction of student dropouts.  
  
 