1. Introduction

The goal of this project is to predict the likelihood of a student dropping out of a course based on various features related to their academic and personal data. The dataset includes various numerical and categorical features, such as grades, attendance, age, and socio-economic factors. We used multiple machine learning models to train on this data, evaluate their performance, and select the best model for predicting student dropout.

---

2. Methodology

2.1 Data Preprocessing

- **Data Loading:** The dataset was loaded from a CSV file, and the features were separated using a semicolon delimiter (;).
- **Feature Selection:** The most relevant features for the dropout prediction task were selected based on mutual information scores. This resulted in the selection of the top 10 most influential features:
  - Curricular units (approved, grade, evaluations, enrolled) for both the 1st and 2nd semesters
  - Course type
  - Tuition fees status
  - Age at enrollment
- **Data Cleaning:** No missing values were found in the dataset. Outliers were detected using boxplots and the IQR method. Outliers were removed from the dataset to ensure model robustness.
- **Feature Encoding:** Categorical variables were encoded using label encoding to make them suitable for machine learning algorithms that require numerical input.
- **Feature Scaling:** Standardization was applied to the numerical features to ensure that all features had a mean of 0 and a standard deviation of 1. This is particularly important for models like SVM that are sensitive to feature scaling.

2.2 Model Training and Evaluation

Eight different machine learning models were trained individually, with hyperparameter tuning done using GridSearchCV to optimize their performance. The following models were used:

1. **Logistic Regression**
2. **Random Forest**

3. **XGBoost**
4. **SVM**
5. **Gradient Boosting**
6. **AdaBoost**
7. **Naive Bayes**
8. **Decision Tree**

Each model was evaluated using cross-validation (5-fold) and trained using the validation set. The models were then assessed based on their **accuracy**, **precision**, **recall**, and **F1-score**.

2.3 Model Evaluation

After training each model separately, their performance was compared using validation accuracy. The following steps were carried out to evaluate the models:

- **Accuracy Calculation:** The accuracy score for each model was obtained using the score() method on the validation set.
- **Model Comparison:** A comparison of models' performance was made based on their accuracy scores.
- **Feature Importance:** SHAP and LIME were used for model interpretability. The most influential features for each model were identified.

2.4 Model Deployment

The best-performing model, which was **Random Forest**, was selected for deployment in a simple interactive web application using **Streamlit**. Users can input data through the app, and the model will predict the likelihood of a student dropping out based on the input features.

---

3. Results

3.1 Model Performance Comparison

The following table displays the validation accuracy scores for each model:

Model	**Validation Accuracy**
Logistic Regression	0.8010
Random Forest	0.8972
XGBoost	0.9037
SVM	0.8538
Gradient Boosting	0.9016
AdaBoost	0.8770
Naive Bayes	0.7373
Decision Tree	0.8646

From this comparison, **XGBoost** (0.9037), **Random Forest** (0.8972), and **Gradient Boosting** (0.9016) performed the best, with **XGBoost** achieving the highest accuracy.

3.2 Model Selection

Although **XGBoost** performed slightly better than **Random Forest** in terms of accuracy, **Random Forest** was selected as the final model for deployment due to its better balance between precision, recall, and F1-score, especially for predicting "Dropout" and "Graduate" students. It achieved an **F1-score** of 0.9148, which was the highest across all models.

3.3 Feature Importance

Using **SHAP** (Shapley Additive Explanations), the top 10 most influential features for the **Random Forest** model were identified:

1. Curricular units 2nd sem (approved)
2. Curricular units 1st sem (approved)
3. Curricular units 2nd sem (grade)
4. Curricular units 1st sem (grade)
5. Course
6. Curricular units 2nd sem (evaluations)

7. Tuition fees up to date
8. Curricular units 1st sem (evaluations)
9. Curricular units 2nd sem (enrolled)
10. Age at enrollment

These features are critical indicators of whether a student is likely to drop out or graduate, with **Curricular units** and **Grade**-related features being the most important.

3.4 Model Deployment

- **Streamlit App:** The selected model (Random Forest) was deployed in an interactive web application using Streamlit. Users can input details like **curricular units**, **grades**, **tuition fees status**, and **age**. Based on these inputs, the app predicts whether the student will be **Enrolled**, **Graduate**, or **Dropout**.

---

4. Insights and Recommendations

- **Feature Insights:** The features that most influence the dropout prediction are related to academic performance, specifically grades and curricular units completed. This highlights the importance of academic support and tracking for students who may be at risk of dropping out.
- **Model Interpretation:** The Random Forest model, with its ensemble approach, is more robust and interpretable compared to others like SVM or Naive Bayes. It can provide valuable insights into feature importance, making it easier to understand the factors affecting student retention.
- **Deployment:** The Streamlit app is an easy and interactive way to deploy the model, allowing for real-time predictions. Future improvements could include adding additional features like socio-economic factors or attendance, which could provide more accurate predictions.

---

5. Future Work

- **Additional Features:** More features such as student attendance, parental support, and extracurricular activities could be incorporated to improve the model's prediction accuracy.
- **Model Fine-tuning:** Further tuning of hyperparameters and testing with other ensemble methods like **LightGBM** or **CatBoost** could lead to better performance.

- **Integration with School Systems:** The model can be integrated into school management systems to monitor students in real-time and intervene if a student is predicted to drop out.

---

6. Conclusion

This project successfully built, evaluated, and deployed a machine learning model to predict student dropouts using various classification algorithms. Among the models tested, **LGB** emerged as the best choice, providing high accuracy and interpretability. The deployment in Streamlit enables easy access for users to predict the dropout status based on student data. This solution can help educational institutions identify at-risk students and take proactive measures to ensure student retention.