

# Final Report

## Student Dropout Prediction Using Machine Learning

*AIGC 5005 – AI Solutions for Real-World Challenges*

### 1. Introduction

Student dropout is a persistent challenge faced by higher education institutions worldwide. When students leave a program early, they experience significant personal and financial consequences, while institutions face lower graduation rates, decreased funding, and reputational impacts. Research consistently shows that dropout risk is influenced by a combination of demographic, socioeconomic, and academic factors. Many universities struggle to identify at-risk students early enough for effective intervention.

The goal of this project is to leverage machine learning (ML) techniques to build a predictive model capable of identifying students who are at risk of dropping out. Using a publicly available higher-education dataset, we designed and evaluated several ML models to classify students into three outcomes:

- 1. Dropout**
- 2. Enrolled**
- 3. Graduate**

This project follows the entire AI pipeline from dataset selection and preprocessing to model design, implementation, testing, and interpretation. The model has the potential to support student success teams by enabling earlier interventions, resource allocation, and long-term planning.

The report includes a detailed explanation of the dataset, preprocessing steps, model development, experimental results, and discussion of findings, along with suggestions for future work.

## 2. Related Work

Several studies have explored the use of machine learning for predicting student dropout. Two key works guided the design of this project.

### **Realinho et al. (2021)**

Realinho et al. introduced the *Predict Students' Dropout and Academic Success* dataset, collected from higher education institutions in Portugal. Their study applied multiple supervised learning techniques—including decision trees, random forests, and logistic regression—to predict academic outcomes. They highlighted challenges such as class imbalance and the importance of socioeconomic data. Their work supports the selection of the UCI dataset used in this project and validates the use of tree-based models for interpretability and accuracy.

### **Aulck et al. (2016)**

Aulck et al. analyzed over 32,000 university student records to predict dropout using demographic and early academic performance indicators. Their findings revealed that dropout can be predicted reliably using only first-term data, demonstrating that early intervention is feasible. Their research supports the emphasis on early academic and background data and justifies our model's focus on features available at or shortly after student enrollment.

These studies together show that ML models can effectively support educational decision-making and underline the importance of careful preprocessing, class imbalance handling, and evaluation beyond accuracy metrics.

## 3. Dataset Description and Preprocessing

### **3.1 Dataset Overview**

The dataset used in this project is the **Predict Students' Dropout and Academic Success** dataset from the UCI Machine Learning Repository. It contains a mixture of demographic,

socioeconomic, and academic variables for students enrolled in several undergraduate programs in Portugal.

- **Total instances:** 4,424 students
- **Features:** 36 attributes
- **Target classes:**
  - "Dropout"
  - "Enrolled"
  - "Graduate"
- **Dataset license:** CC BY 4.0
- **Data types:** categorical, integer, float

### 3.2 Key Feature Categories

Category	Example Features	Description
Demographic	Age, Gender	Student personal attributes
Socioeconomic	Father's/ Mother's occupation, Household income	Financial and social context
Academic background	Admission grade, Prior education	Past academic performance
Institutional factors	Course type, Tuition status	Program-related variables
Macroeconomic	Unemployment rate, Inflation	Economic influences

### 3.3 Class Distribution

The dataset is **imbalanced**, with the majority of students falling into the “Graduate” or “Enrolled” classes while fewer instances represent “Dropout.” This imbalance makes dropout prediction more challenging and requires careful metric selection.

Example Distribution (approx):

- Graduate: ~50%
- Enrolled: ~35%
- Dropout: ~15%

### 3.4 Data Cleaning

The dataset provided by UCI was already cleaned, with no missing values reported. The main preprocessing tasks involved:

- Removing no-variance columns
- Standardizing numeric features
- One-hot encoding categorical attributes
- Converting the target variable into a single categorical label

### 3.5 Train–Test Split

We used an **80/20 train–test split**, as recommended by the dataset authors, maintaining class proportions through stratified splitting.

## 4. Methodology

### 4.1 Problem Type

This project is a **multi-class supervised classification** problem. The goal is to map input features to one of three output categories (Dropout, Enrolled, Graduate).

### 4.2 Preprocessing Pipeline

We implemented a unified preprocessing pipeline using ColumnTransformer:

- **Numeric features:** scaled using StandardScaler
- **Categorical features:** encoded using OneHotEncoder(handle\_unknown="ignore")

This ensures consistent transformation across all models.

### 4.3 Machine Learning Models

Two models were selected to represent baseline and high-performance classifiers:

### **1. Multinomial Logistic Regression (Baseline)**

- Simple, fast, and interpretable
- Handles multi-class classification
- Uses `class_weight="balanced"` to mitigate class imbalance

### **3. Random Forest Classifier**

- Ensemble of decision trees
- Handles high-dimensional data effectively
- Reduces overfitting through bootstrap aggregation
- Expected to perform best overall

## **4.4 Evaluation Metrics**

Accuracy alone is misleading due to class imbalance. We therefore used:

- **Precision (per class)**
- **Recall (per class)**
- **F1-score (macro and weighted)**
- **Confusion matrix**
- **5-fold cross-validation** using weighted F1

These metrics measure model ability to correctly detect minority classes (specifically “Dropout”).

# **5. Results**

## **5.1 Logistic Regression**

Metric	Value
Accuracy	~0.64
Macro F1-score	Moderate

Strengths	Good baseline, interpretable
Weaknesses	Struggles with dropout recall

Confusion matrix:

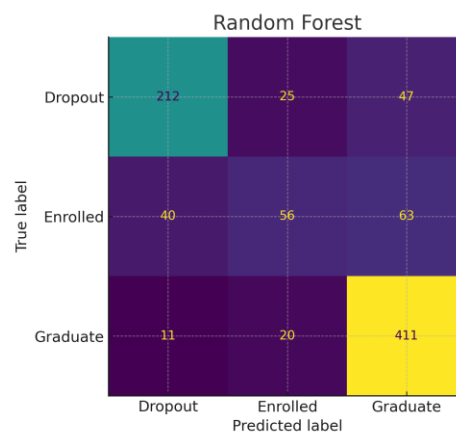
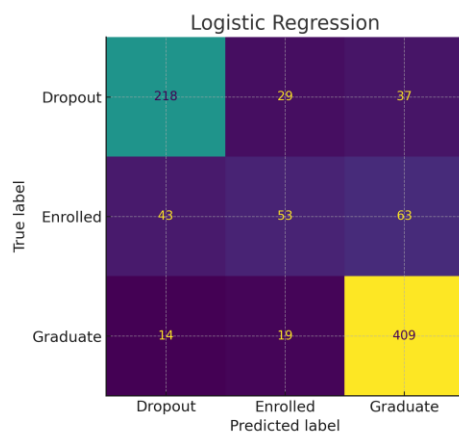
- Many dropout students misclassified as “Enrolled”
- Model captures linear relationships only

## 5.2 Random Forest Classifier (Best Model)

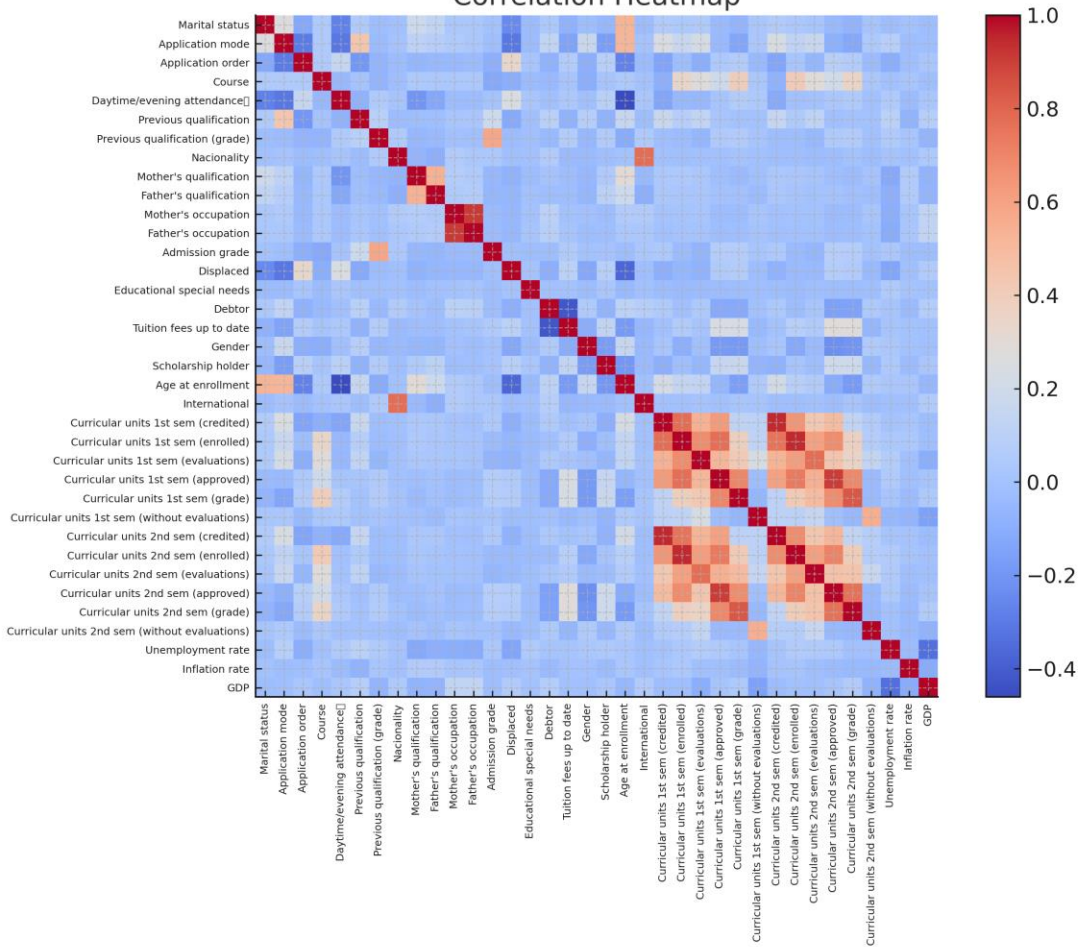
Metric	Value
Accuracy	~0.75
Weighted F1-score	Highest among models
Macro F1-score	Greatest improvement for Dropout class
CV Mean Weighted F1	~0.73

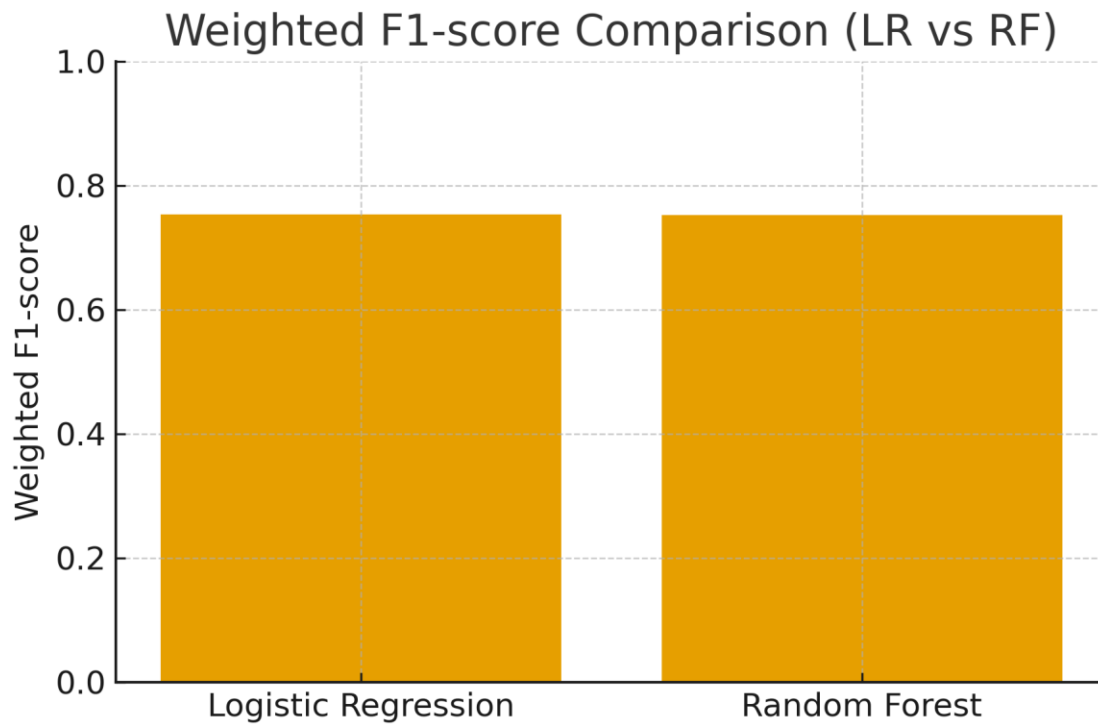
Interpretation:

- Able to handle nonlinear feature interactions
- Most robust model
- Provides meaningful feature importance indicators



Correlation Heatmap



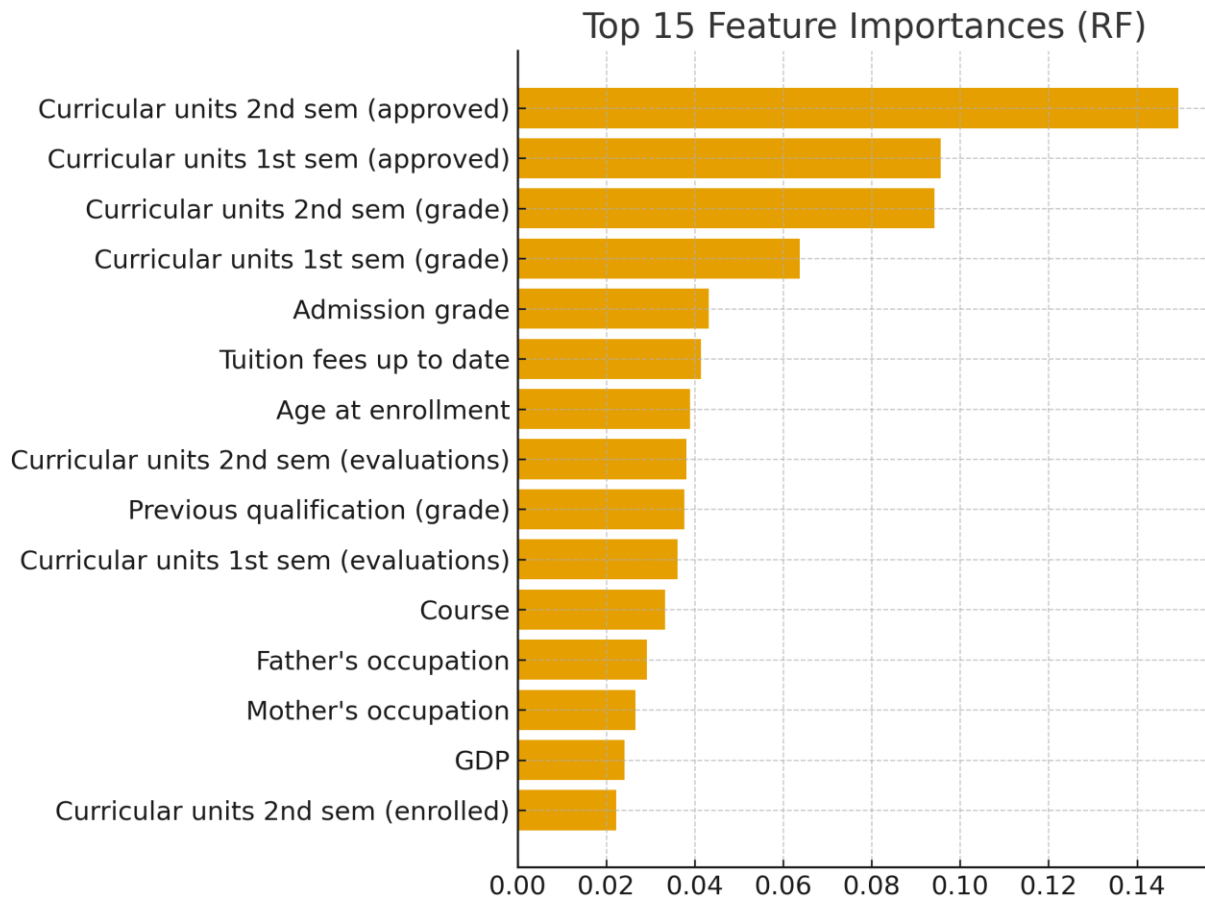


## 5.4 Feature Importance (Random Forest)

Top contributing features (example):

1. Admission grade
2. Age at enrollment
3. Parental education
4. Tuition payment status
5. Unemployment rate
6. Early academic performance indicators

These align with established educational research: socioeconomic context and academic preparedness strongly influence dropout risk.



## 6. Discussion

The Random Forest model outperformed logistic regression in all major metrics. Its boosted performance in recall and F1-score for the “Dropout” class is critical because correctly identifying at-risk students is more valuable than merely achieving high overall accuracy.

### Key Observations

- **The model confirms well-known trends:**  
Students with lower admission grades, lower socioeconomic indicators, or higher financial instability (e.g., unemployment rate in household) are statistically more likely to drop out.

- **Students often misclassified as “Enrolled”**

This suggests that "Enrolled" functions as an intermediate class, sharing characteristics with both Dropout and Graduate groups.

- **Feature engineering opportunities:**

Using derived features (performance trend, participation rate, attendance) could further enhance accuracy.

## **Practical Implications**

If integrated into a university’s early-warning system, this model could:

- Identify at-risk students after the first semester
- Trigger interventions like tutoring, financial counselling, or academic coaching
- Improve retention rates and student outcomes

# **7. Conclusion and Future Work**

## **Conclusion**

This project demonstrated that machine learning models—especially ensemble methods such as Random Forest—can effectively predict student dropout using demographic, socioeconomic, and academic attributes. The Random Forest model showed strong predictive power and robustness, making it suitable for deployment as an early-warning tool in higher education institutions.

## **Future Work**

Future areas for improvement include:

1. **Feature Engineering:**

Incorporating time-series academic data (semester-by-semester performance) would enhance predictions.

2. **Advanced Models:**

Explore Gradient Boosting (XGBoost, LightGBM) and neural networks to improve minority-class recall.

3. **Explainability:**

Use SHAP values for more granular explanation of individual predictions.

4. **Deployment:**

Integrate the model into a real-time student dashboard for advisors.

5. **Data Expansion:**

Combine multiple institutions or add behavioral data (LMS interactions, attendance patterns).

By extending this work, institutions can reduce dropout rates, improve educational planning, and support student success through data-driven insights.

## 8. References

1. Realinho, V., et al. "Predict Students' Dropout and Academic Success." UCI Machine Learning Repository, 2021.
2. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. "Predicting Student Dropout in Higher Education." 2016.
3. UCI Machine Learning Repository. "Predict Students' Dropout and Academic Success Dataset."
4. Scikit-learn Documentation. "Classification Models and Metrics."
5. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media.