UNIVERSITY of HOUSTON

CULLEN COLLEGE of ENGINEERING

Predicting Student Dropout and Academic Success Using Machine Learning

Team:
AKSHAY MANOJ (Student ID: 2154017)
NIKITA ANIL YADAV (Student ID: 2349683)
SAAD RAJA MUHAMMAD(Student ID: 2313786)
SHREYA SUNIL TELAVANE (Student ID: 2398529)
VAMSHI KRISHNA MADHAVAN (Student ID: 2344937)

Submission Date: March 9, 2025

UNIVERSITY of
HOUSTON
CULLEN COLLEGE of ENGINEERING

Contents

List of Tables

List of Figures

1 Introduction

The ability to predict student dropout and academic success is crucial for educational institutions to implement early interventions, improve student retention rates, and enhance academic performance. Identifying at-risk students early can lead to personalized support strategies, reducing dropout rates and fostering better educational outcomes.

Challenges in this research include the following:

- Handling missing or imbalanced data, as dropout cases may be fewer compared to successful students.
- Selecting the most relevant features that contribute to accurate predictions.
- Ensuring model interpretability so that educational stakeholders can effectively use the insights.
- Addressing ethical concerns such as bias in predictions based on socioeconomic status or demographic attributes.

2 Problem Statement

How accurately can machine learning models predict student dropout and academic success based on demographic, academic, and socioeconomic factors?

3 Data Description

The dataset *Predict students' dropout and academic success* is obtained from the UCI Machine Learning Repository: UCI Machine Learning Repository.

This dataset contains demographic, academic, and socioeconomic data about students, along with their final status (dropout, enrolled, or graduated). It includes features such as age, gender, course enrollment, family support, financial situation, academic performance, and more.

Variable Name	**Data Type**	**Description**
Marital Status	Integer	1 - Single, 2 - Married, 3 - Widowed, 4 - Divorced, 5 - Common-law, 6 - Separated
Application mode	Integer	Various application phases and contingents (e.g., 1st phase, international, transfer)
Application order	Integer	Application order (0 - first choice, 9 - last choice)
Course	Integer	Course code (e.g., 33 - Biofuel Production, 171 - Animation Design, etc.)
Daytime/evening attendance	Integer	1 - Daytime, 0 - Evening

Variable Name	Data Type	Description
Previous qualification	Integer	Highest previous education (e.g., 1 - Secondary, 2 - Bachelor's, etc.)
Previous qualification (grade)	Continuous	Grade of previous qualification (0-200)
Nationality	Integer	Nationality code (e.g., 1 - Portuguese, 41 - Brazilian, etc.)
Mother's qualification	Integer	Mother's highest education (e.g., 1 - Secondary, 2 - Bachelor's, etc.)
Father's qualification	Integer	Father's highest education (e.g., 1 - Secondary, 2 - Bachelor's, etc.)
Mother's occupation	Integer	Mother's occupation (e.g., 0 - Student, 2 - Scientist, 5 - Personal Services)
Father's occupation	Integer	Father's occupation (e.g., 0 - Student, 2 - Scientist, 5 - Personal Services)
Admission grade	Continuous	Admission grade (0-200)
Displaced	Integer	1 - Yes, 0 - No
Educational special needs	Integer	1 - Yes, 0 - No
Debtor	Integer	1 - Yes, 0 - No
Tuition fees up to date	Integer	1 - Yes, 0 - No
Gender	Integer	1 - Male, 0 - Female
Scholarship holder	Integer	1 - Yes, 0 - No
Age at enrollment	Integer	Age of student at enrollment
International	Integer	1 - Yes, 0 - No
Curricular units 1st sem (credited)	Integer	Number of units credited in 1st semester
Curricular units 1st sem (enrolled)	Integer	Number of units enrolled in 1st semester
Curricular units 1st sem (evaluations)	Integer	Number of evaluations for 1st semester units
Curricular units 1st sem (approved)	Integer	Number of approved units in 1st semester
Curricular units 1st sem (grade)	Integer	Grade average for 1st semester (0-20)
Curricular units 1st sem (without evaluations)	Integer	Number of units without evaluations in 1st semester
Curricular units 2nd sem (credited)	Integer	Number of units credited in 2nd semester
Curricular units 2nd sem (enrolled)	Integer	Number of units enrolled in 2nd semester
Curricular units 2nd sem (evaluations)	Integer	Number of evaluations for 2nd semester units
Curricular units 2nd sem (approved)	Integer	Number of approved units in 2nd semester
Curricular units 2nd sem (grade)	Integer	Grade average for 2nd semester (0-20)
Curricular units 2nd sem (without evaluations)	Integer	Number of units without evaluations in 2nd semester
Unemployment rate	Continuous	Unemployment rate (%)
Inflation rate	Continuous	Inflation rate (%)
GDP	Continuous	Gross Domestic Product
Target	Categorical	Classification task (dropout, enrolled, graduate)

Table 1: Variable Names and Corresponding Definitions

A simple examination of the data shows class imbalance: fewer students drop out compared to those who graduate or remain enrolled. Addressing this imbalance will be essential for developing robust predictive models.

Variable Name	Value 1	Value 2
Marital Status	1	1
Application mode	17	15
Application order	5	1
Course	171	9254
Daytime/evening attendance	1	1
Previous qualification	1	1
Previous qualification (grade)	122	160
Nationality	1	1
Mother's qualification	19	1
Father's qualification	12	3
Mother's occupation	5	3
Father's occupation	9	3
Admission grade	127.3	142.5
Displaced	1	1
Educational special needs	0	0
Debtor	0	0
Tuition fees up to date	1	0
Gender	1	1
Scholarship holder	0	0
Age at enrollment	20	19
International	0	0
Curricular units 1st sem (credited)	0	0
Curricular units 1st sem (enrolled)	0	6
Curricular units 1st sem (evaluations)	0	6
Curricular units 1st sem (approved)	0	6
Curricular units 1st sem (grade)	0	14
Curricular units 1st sem (without evaluations)	0	0
Curricular units 2nd sem (credited)	0	0
Curricular units 2nd sem (enrolled)	0	6
Curricular units 2nd sem (evaluations)	0	6
Curricular units 2nd sem (approved)	0	6
Curricular units 2nd sem (grade)	0	13.66666667
Curricular units 2nd sem (without evaluations)	0	0
Unemployment rate	10.8	13.9
Inflation rate	1.4	-0.3
GDP	1.74	0.79
Target	Dropout	Graduate

Table 2: Student Data Points

4 Method

To analyze and predict student dropout and academic success, we will use a combination of the following techniques:

4.1 Dimensionality Reduction

- Principal Component Analysis (PCA) to reduce feature redundancy and improve computational efficiency.

- Feature selection techniques such as Recursive Feature Elimination (RFE) to retain the most informative variables.

4.2 Prediction Models

- Supervised learning models such as Logistic Regression, Random Forest, and XGBoost for classification.

4.3 System Modeling and Evaluation

- Model evaluation using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.
- Cross-validation to ensure model generalizability.
- Addressing class imbalance using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to enhance prediction performance for minority classes.

By integrating these methods, we aim to develop an effective predictive framework to assist educational institutions in identifying and supporting at-risk students before they drop out.

4.4 Project Timeline

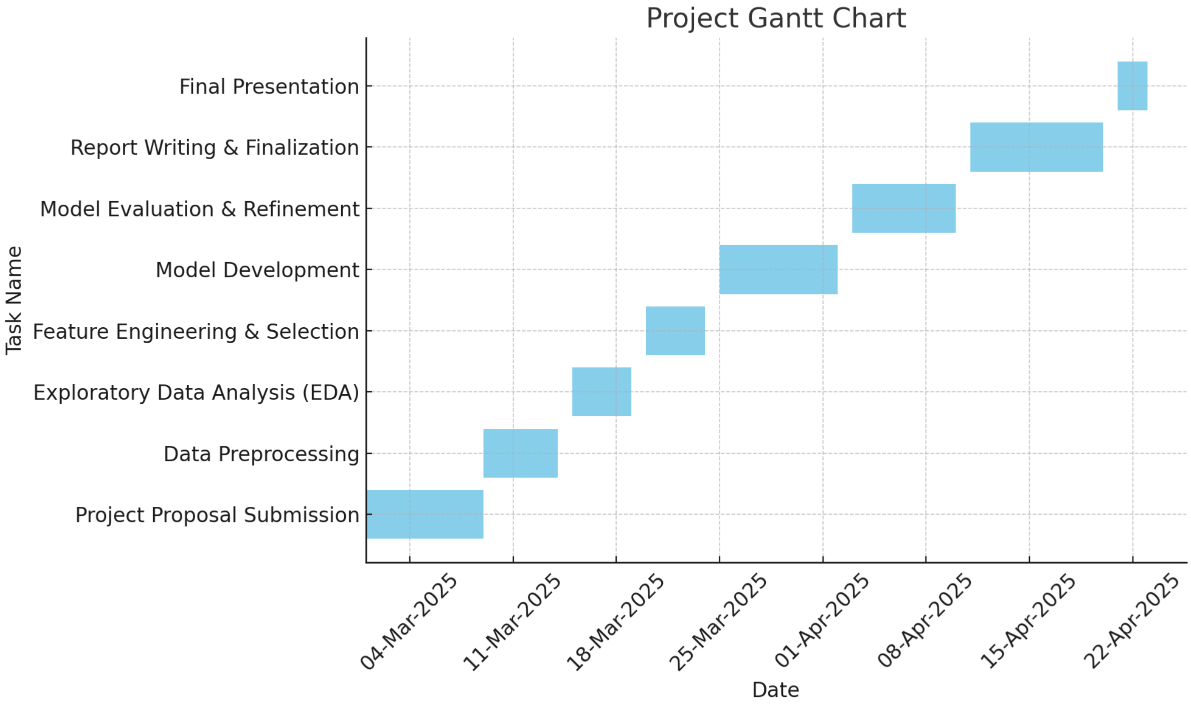

Figure 1: Task Gantt Chart

5 References

UCI Machine Learning Repository. (n.d.). Predict students' dropout and academic success dataset. Retrieved from http://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success.