VIETNAM NATIONAL UNIVERSITY,
HO CHIMINH CITY
UNIVERSITY OF INFORMATION TECHNOLOGY
FACULTY OF COMPUTER SCIENCE

FINAL PROJECT:

STUDENTS' DROPOUT AND ACADEMIC SUCCESS PREDICTION

Course: CS114 – Machine Learning
Lecturer: PhD. Vo Nguyen Le Duy

Project Members:
1. Truong Thien Phu 23521190
2. Nguyen Xuan Phuc 23521213
3. Tang Hoang Phuc 23521219
4. Phan Thuy Phuong 23521248

I. Introduction and Related Work

Student dropout has become an increasingly common issue in higher education, affecting not only individual academic outcomes but also the overall quality and reputation of educational institutions. The ability to predict whether a student will graduate, drop out, or continue their studies (remain enrolled) is crucial for early intervention and support, helping institutions improve student retention and success rates.

In this study, we utilize a cleaned dataset sourced from Kaggle, containing academic and demographic information of over 75,000 students. The main objective is to build a predictive model that classifies students into one of three academic outcomes: graduate, dropout, or enrolled. By applying machine learning techniques, we aim to uncover key patterns and factors influencing student success, thereby supporting data-driven decisions in academic management.

Our primary objectives and methodology in this final report are:

1. **Data Exploration & Preprocessing:** To examine the distribution and interrelationships among features, handle missing values, address class imbalance, and engineer relevant features that may reveal hidden patterns in student academic performance.

2. **Model Development:** To train and compare multiple classification algorithms including Logistic Regression, K-Nearest Neighbors, Random Forest, and Support Vector Machine – in order to identify the most effective model for predicting academic outcomes (Graduated, Dropped Out, or Enrolled).

3. **Evaluation:** To assess model performance through cross-validation using a range of evaluation metrics, including accuracy, precision, recall, and F1-score. Among these, F1-score is considered the primary metric due to its ability to provide a balanced evaluation of precision and recall, especially in cases of class imbalance.

By combining statistical analysis with modern machine learning techniques, this study aims not only to build an accurate prediction system but also to generate practical insights that can support educational institutions in improving student retention and success strategies.

II. Methodologies

A. Problem Setup

The primary objective of this study is to develop a predictive model that can accurately classify students into one of three academic outcomes: *Graduated*, *Dropped Out*, or *Enrolled*. This task is framed as a multiclass classification problem, where each student is represented by a set of features derived from academic records, demographic attributes, and enrollment history.

Given a dataset consisting of labeled student records, the goal is to learn a function:

$$f : X \to Y$$

where:

- $X \subseteq \mathbb{R}^n$ denotes the feature space, representing student characteristics (e.g., age, gender, financial aid status, admission grade, etc.),
- $Y = \{\textit{Graduated}, \textit{Dropped Out}, \textit{Enrolled}\}$ represents the set of academic outcomes.

Each student instance $x_i \in X$ is associated with a ground-truth label $y_i \in Y$, and the objective is to train a model f that generalizes well to unseen data—i.e., accurately predicts the academic outcome of new student records.

The dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ is partitioned into training and test subsets. A variety of machine learning algorithms—including Support Vector Machines (SVM), Multinomial Logistic Regression, Random Forests, and K-Nearest Neighbors (KNN)—will be evaluated to identify the most effective model for classification.

Model performance is assessed using standard multiclass metrics such as accuracy, precision, recall, and F1-score. Special consideration is given to class imbalance, ensuring fair evaluation across all outcome categories.

Ultimately, the proposed model is designed to support academic institutions in the early identification of at-risk students, thereby enabling timely and targeted interventions to enhance student retention and academic success.

B. Data Exploration

The dataset provides detailed information on students, encompassing demographic attributes, academic performance, economic status, and macroeconomic context. The primary objective is to predict each student's academic status—*Graduated*, *Dropped Out*, or *Enrolled*—based on these features.

The dataset is well-structured with minimal presence of outliers. In terms of class distribution, approximately 47% of students are labeled as *Graduated*, 33% as *Dropped Out*, and 20% as *Enrolled*, indicating a moderate class imbalance. Although not perfectly balanced, this distribution still provides a reasonable representation of each class, which is beneficial for training classification models. A relatively balanced target variable helps reduce model bias toward majority classes and contributes to more robust and generalizable predictive performance across all outcome categories.

Most features are categorical or ordinal with clear semantic meaning, suggesting that proper encoding and preprocessing will be essential to preserve their predictive value during the modeling phase. While some features—such as academic performance metrics and financial status indicators—are expected to play a significant role in determining student outcomes, others may have minimal or negligible impact. Identifying and retaining the most informative variables, while filtering out irrelevant or redundant ones, is therefore a critical step to enhance model efficiency, reduce noise, and improve overall predictive accuracy.