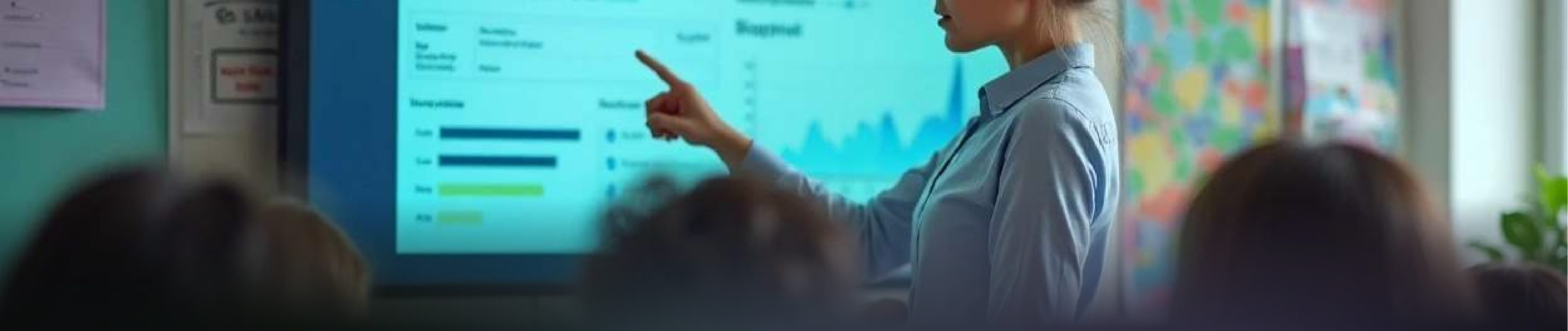


# Predicting Student Dropout Rates for Early Intervention


V

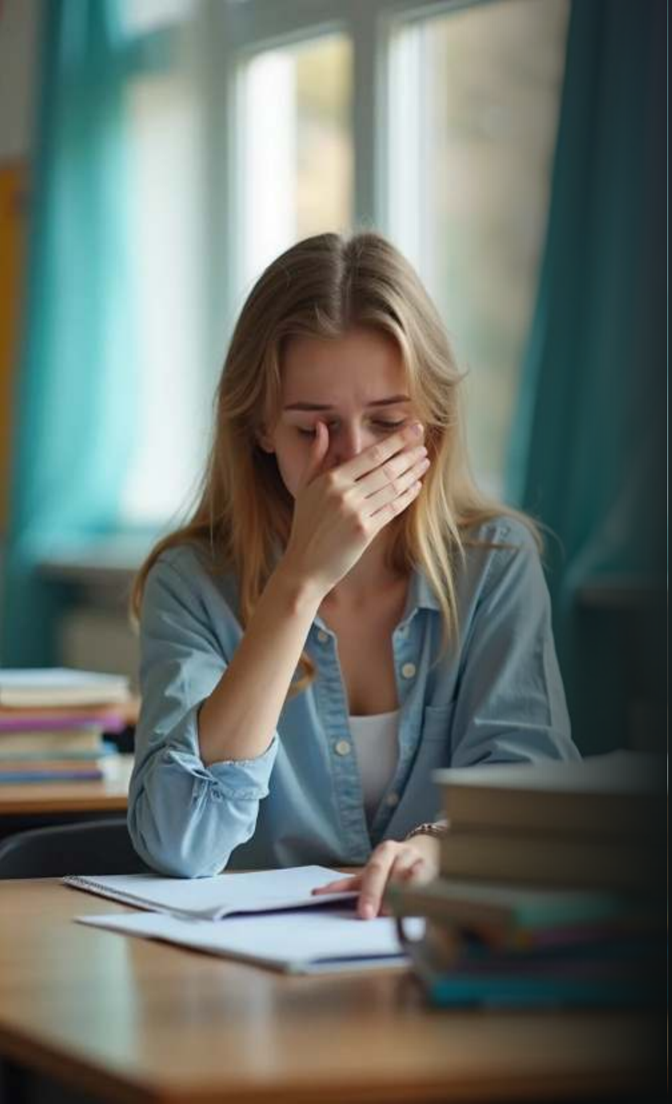
**Author: Vincent Bwoga**

Created with Pi

# CONTENTS

01

# Objectives



# Objective Details

1. Identify students at high risk of dropping out by the end of the first semester.
2. Provide actionable insights to academic advisors for targeted interventions.
3. Improve overall student retention rates by 10% within two years.

02

# Stakeholders

# Stakeholder List

University administration (to allocate resources and monitor program success).

Academic advisors (to implement interventions based on predictions).

03

# Key Performance Indicator (KPI)

# KPI Details

Retention Rate Improvement: Percentage increase in student retention after implementing interventions, targeting a 10% improvement.

04

# Data Collection & Preprocessing

# Data Sources

Student Information System (SIS): Academic records, grades, and enrollment status.

Learning Management System (LMS): Engagement metrics like assignment submissions and login frequency.

# Potential Bias

Historical academic records may reflect socioeconomic biases, as students from lower-income backgrounds may have lower grades due to external factors (e.g., part-time work), skewing risk predictions.

# Preprocessing Steps

1. Handling Missing Data: Impute missing grades or attendance records using median values for numerical data or mode for categorical data to maintain dataset integrity.
2. Normalization: Scale numerical features (e.g., GPA, attendance) to a 0-1 range to ensure equal weighting in model training.
3. Feature Encoding: Convert categorical variables (e.g., major, enrollment status) into numerical formats using one-hot encoding to make them model-compatible.

05

# Model Development

# Model Development

### Model Choice

- Random Forest
  - Justification: Random Forest handles non-linear relationships and mixed data types (numerical and categorical) well, is robust to outliers, and provides feature importance for interpretability, which is valuable for understanding dropout factors.

### Data Splitting

- Split data into 70% training, 15% validation, and 15% test sets. The training set trains the model, the validation set tunes hyperparameters, and the test set evaluates final performance to ensure unbiased assessment.

### Hyper parameters to Tune

- Number of Trees (n_estimators): Increasing