ClassPass: Predicting Student Success with Explainable AI

Phil Akagu-Jones & Ariel Tyson

CMPT 310 - D100 Introduction to Artificial Intelligence, Fall 2025

Motivation, Purpose & Overview

ClassPass is an AI-driven prediction system designed to classify university students into three categories at enrollment: **Dropout**, **Enrolled**, or **Graduate**. By analyzing demographic and academic data, the system aims to identify at-risk students early, allowing for timely academic intervention.

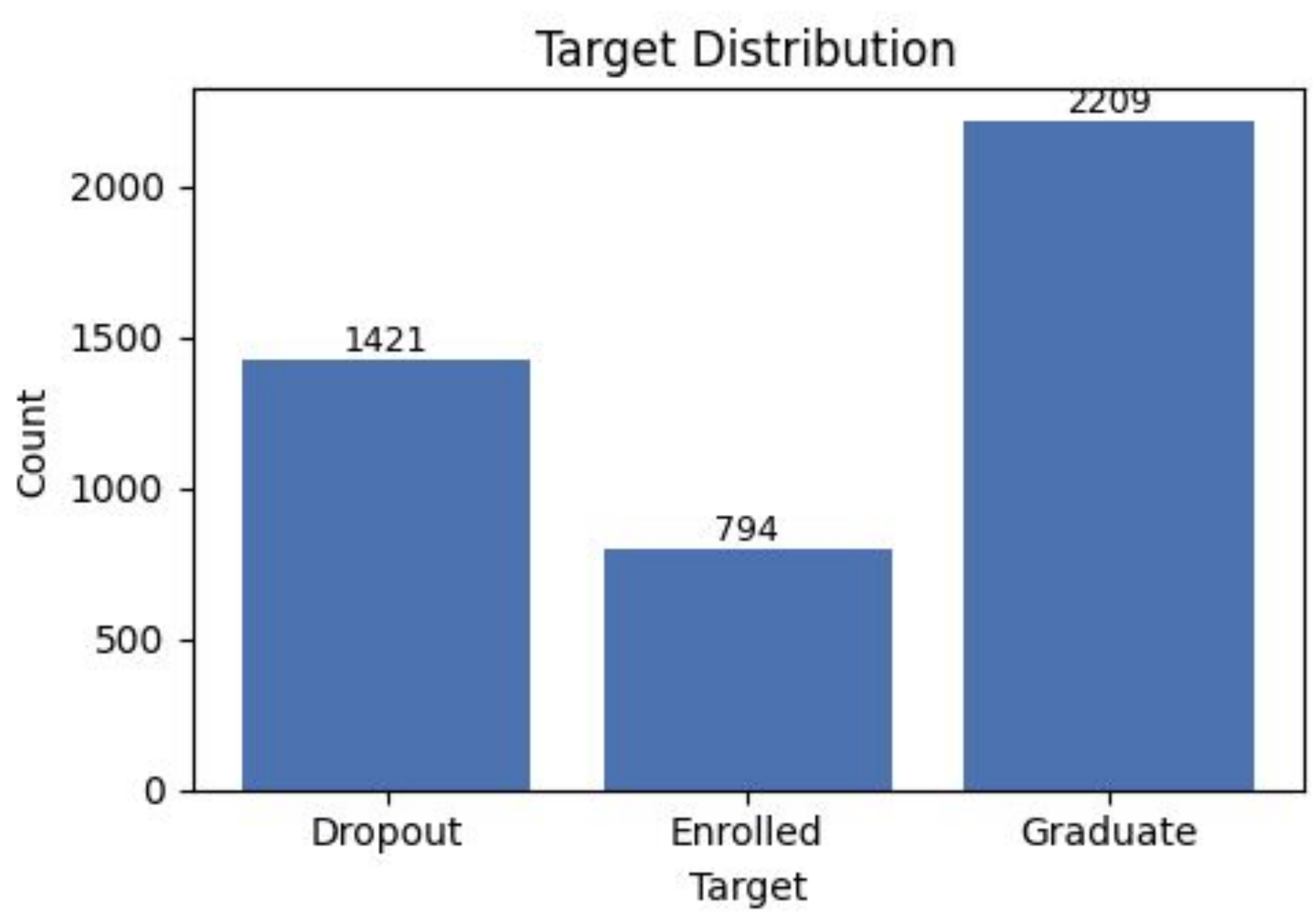
Motivation

- **Early Intervention:** Student retention is a critical challenge. Identifying risk factors before grades drop is essential for support.
- **Interpretability:** Unlike "black box" models, ClassPass focuses on explainable predictions (e.g., "Student X is at risk because they share characteristics with these 5 past dropouts").
- **Custom Implementation:** All core algorithms were implemented from scratch to demonstrate a deep understanding of AI fundamentals.

The Data

- **Source:** UCI Predict Students' Dropout and Academic Success Dataset.
- **Features:** Socio-economic factors, academic history (grades/failures), and enrollment details.
- **Challenge:** The dataset suffers from **Class Imbalance**, with "Dropout" cases being underrepresented compared to "Graduate"

Target Distribution



Methods & Technologies

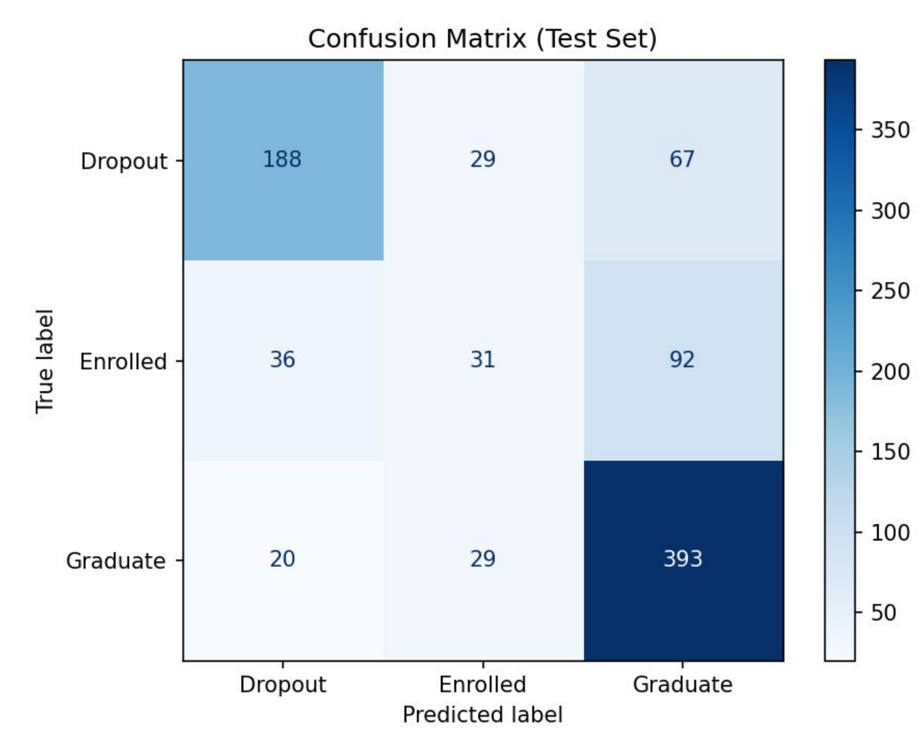
System Architecture The system follows a modular pipeline designed for reproducibility and scalability.

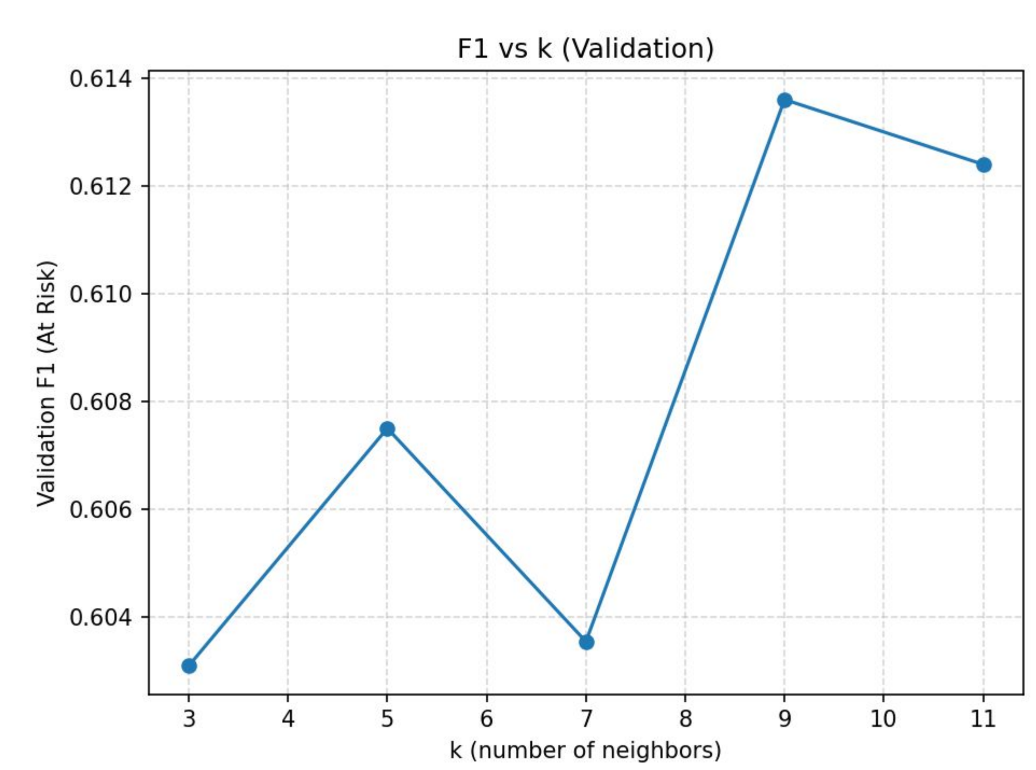
1. **Raw Data** (`data/raw`)
2. **Preprocessing** (`src/classpass/preprocess.py`): One-Hot Encoding -> Standardization -> Stratified Split.
3. **Modeling** (`src/classpass/knn.py`): Custom kNN Classifier.
4. **Evaluation** (`src/classpass/evaluation.py`): F1 Score, Confusion Matrix.
5. **Output**: Prediction + Neighbor Explanation.

Key Methods & Technologies

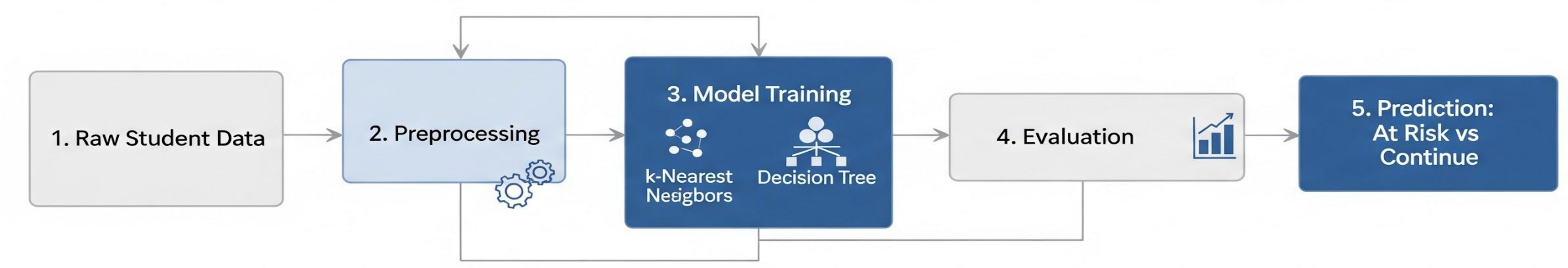
- **Custom k-Nearest Neighbors (kNN):**
  - Implemented from scratch (no `sklearn` for core logic).
  - Supports **Euclidean** and **Manhattan** distance metrics.
  - Features a **k-d tree** (planned/implemented) for efficient querying.
  - **Explainability:** Returns indices of nearest neighbors to justify predictions.
- **Data Preprocessing Pipeline:**
  - Automated cleaning and parsing of numerical statistics.
  - Handling of categorical features via one-hot encoding (e.g., occupations, qualifications).
  - Validation sets used for hyperparameter tuning (finding optimal *k*).

Reports & Figures





Data Pipeline



Conclusions

Challenges & Lessons Learned

- **High Dimensionality:** Extensive one-hot encoding of categorical features (like "Mother's Occupation") increased dataset size and complexity.
- **Class Imbalance:** Required careful validation to prevent the model from bias toward the majority class ("Graduate").
- **Verification:** Unit testing the custom kNN against standard implementations was time-consuming but ensured correctness.

References

1. Realinho, V., Machado, J., & Baptista, L. (2021). "Predict Students' Dropout and Academic Success". UCI Machine Learning Repository.