

CMPT 310 Project Proposal

Group Name: **ClassPass**
Phil Akagu-Jones — paa46@sfu.ca
Ariel Tyson — ajt11@sfu.ca

October 26, 2025

Project Idea

We're building a **multi-technique student-success prediction system** that classifies students at enrollment into **Dropout / Enrolled / Graduate**. The goal is to **flag at-risk students early** using enrollment features (e.g., admission grade, attendance type, prior failures, financial-aid indicators), with **local** and **global** explainability.

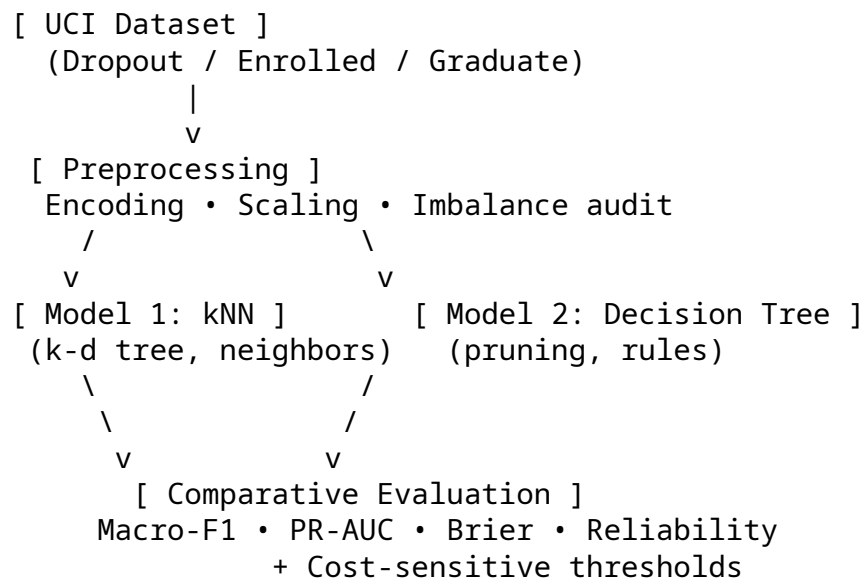
Techniques (implemented by us):

- **Custom k-Nearest Neighbors (kNN)**: Euclidean/Manhattan distances, feature scaling, **k-d tree** acceleration, **neighbor-based local explanations** (top-k exemplars).
- **Custom Decision Tree (ID3/CART-style)**: information gain/GINI, **pre/post-pruning**, depth/leaf constraints, **rule-based global explanations** + feature importance.

Evaluation & rigor: Nested cross-validation, **macro-F1**, per-class **PR-AUC**, **calibration** (Brier + reliability curves), and **cost-sensitive thresholds** emphasizing Dropout.

Data challenges: 3-class **imbalance** + mixed types; we use stratified splits, class-weighted metrics, and a short **resampling audit** if needed.

System Overview (visual)



Tools and Resources

- **Language / Libraries:** Python 3; NumPy, pandas; matplotlib (plots); scikit-learn *utilities* (metrics/plots only); optional imbalanced-learn.
- **Dataset:** UCI *Predict Students' Dropout and Academic Success* (public, tabular).
<https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>

Project Plan / Timeline

Milestone 1 (Oct 26):

- Data audit & preprocessing (encodings, scaling, **stratified** splits).
- **Custom kNN (brute-force) baseline** on 3-class target.
- Metrics: **macro-F1** + confusion matrix; a few **neighbor-based explanations**.
- Scaffold for nested-CV + plotting.

Milestone 2 (Nov 16):

- **Custom Decision Tree** + pruning; add **k-d tree** to kNN.
- Nested CV; per-class **PR curves**, **calibration** (Brier/reliability), **cost-sensitive thresholds**.
- Ablations: distance/scaling (kNN), depth/pruning (DT), optional feature subsets.

Final / Demo (Dec 2):

- Polished comparison; local/global explanations.
- Reproducible code + README + env file; **How-To Guide**; live/recorded demo.

Timeline (visual):

```
Weeks:  0    2    4    6    8    10   12   14
M1      [====]
M2           [=====]
Final/Demo                [===]
\end{T}
```

Minimal Viable System

Load & clean data; encode categoricals; scale numerics; **stratified** train/val/test split.

- Implement **custom kNN (brute-force)**; tune **k**; report **macro-F1** & confusion matrix.
- Show **2-3 neighbor-based explanations** for sample predictions. (*Anchors Milestone 1.*)

Metric templates (visuals)

Confusion Matrix (template, 3x3)

```
+-----+-----+-----+
|       | P1  | P2  | P3   (Predicted)
+-----+-----+-----+
| T1    |  •   |  •   |  •
| T2    |  •   |  •   |  •
| T3    |  •   |  •   |  •
```

(True)

PR Curve (template)

