Education
SERVICES LTD

Predicting Student Dropout Using Machine Learning

Veerpal Kaur CHATTHA

Education
SERVICES LTD

Goal:
Early identification of students at risk of dropping out using demographic, academic, and socioeconomic data for EducationServices Ltd..

Dataset:
4,424 students · 37 variables · Multi-class outcomes.

Data comes from This dataset is supported by program SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191, Portugal.

Education
SERVICES LTD

Problem Statement & Motivation

WHY THIS MATTERS

- Student dropout has academic, financial, and social consequences
- Early detection enables:
- Targeted academic support
- Financial intervention
- Improved student retention

KEY INSIGHTS

- Academic background and early performance are strong predictors of dropout
- Socioeconomic factors play a significant role in student success
- Certain courses and enrollment types show higher dropout rates
- Machine learning models can effectively identify at-risk students at enrollment

Education
SERVICES LTD

Dataset Overview (EDA – Structure)

Data Characterstics:

- Rows: 4,424 students
- Features: 36 input variables
- Target: Student outcome

Feature Types

- Demographic (age, gender, marital status)
- Academic (grades, curricular units)
- Financial (tuition status, scholarship)
- Socioeconomic (GDP, unemployment, inflation)

Class Distribution

- Graduate → ~50%
- Dropout → ~32%
- Enrolled → ~18%

Key Insight

- Dataset is moderately imbalanced
- Accuracy alone is misleading
- Required balanced evaluation metrics

Key Observations

- Academic performance strongly correlates with outcomes
- Students with:
  - Lower approved units
  - Unpaid tuition
  - Lower admission grades
  - show higher dropout risk
- Socioeconomic variables add context, not dominance

EDA Conclusion

Dropout is multi-factorial, not caused by a single variable.

Exploratory Data Analysis

Question

Are demographic characteristics associated with higher dropout rates?

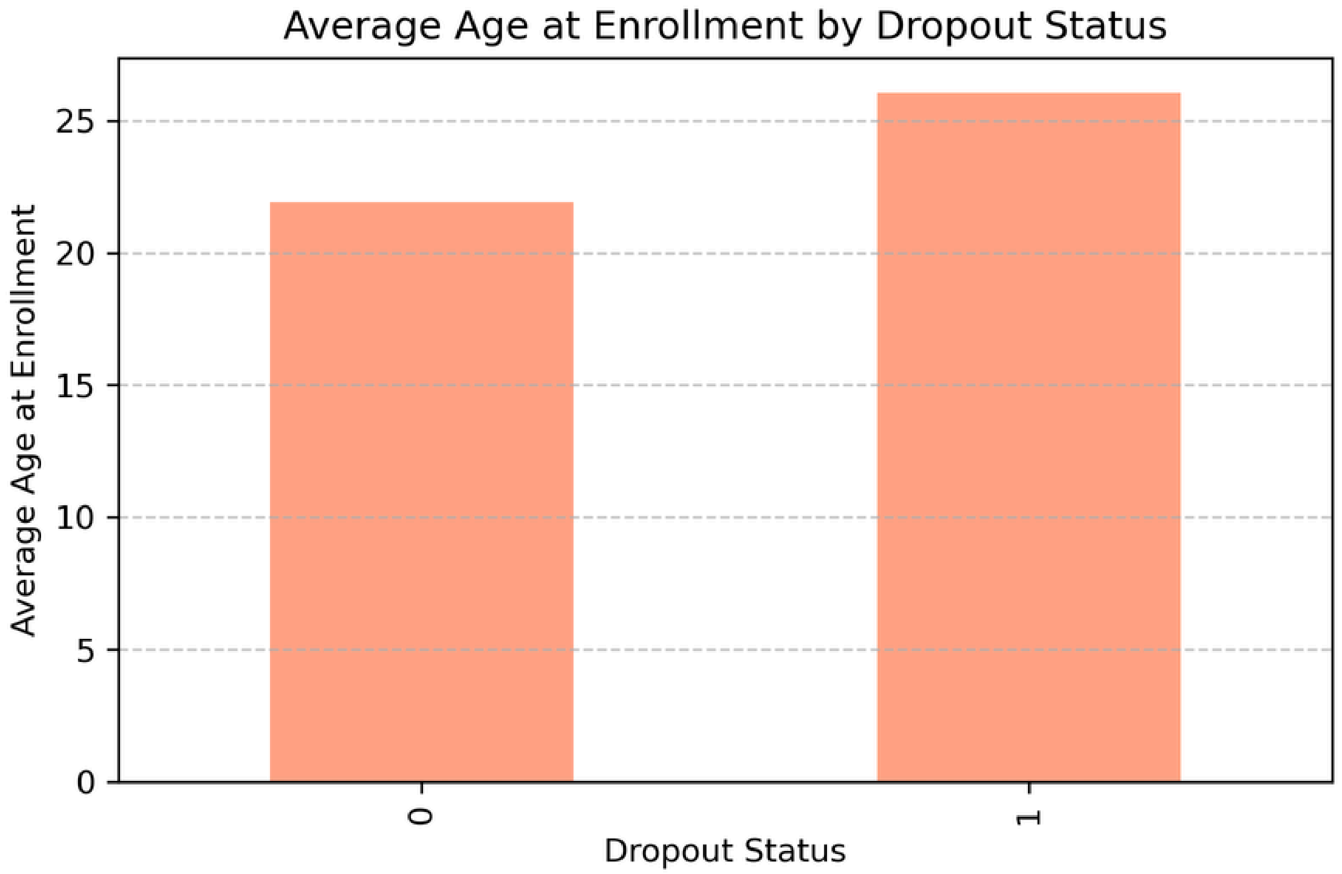
Hypotheses

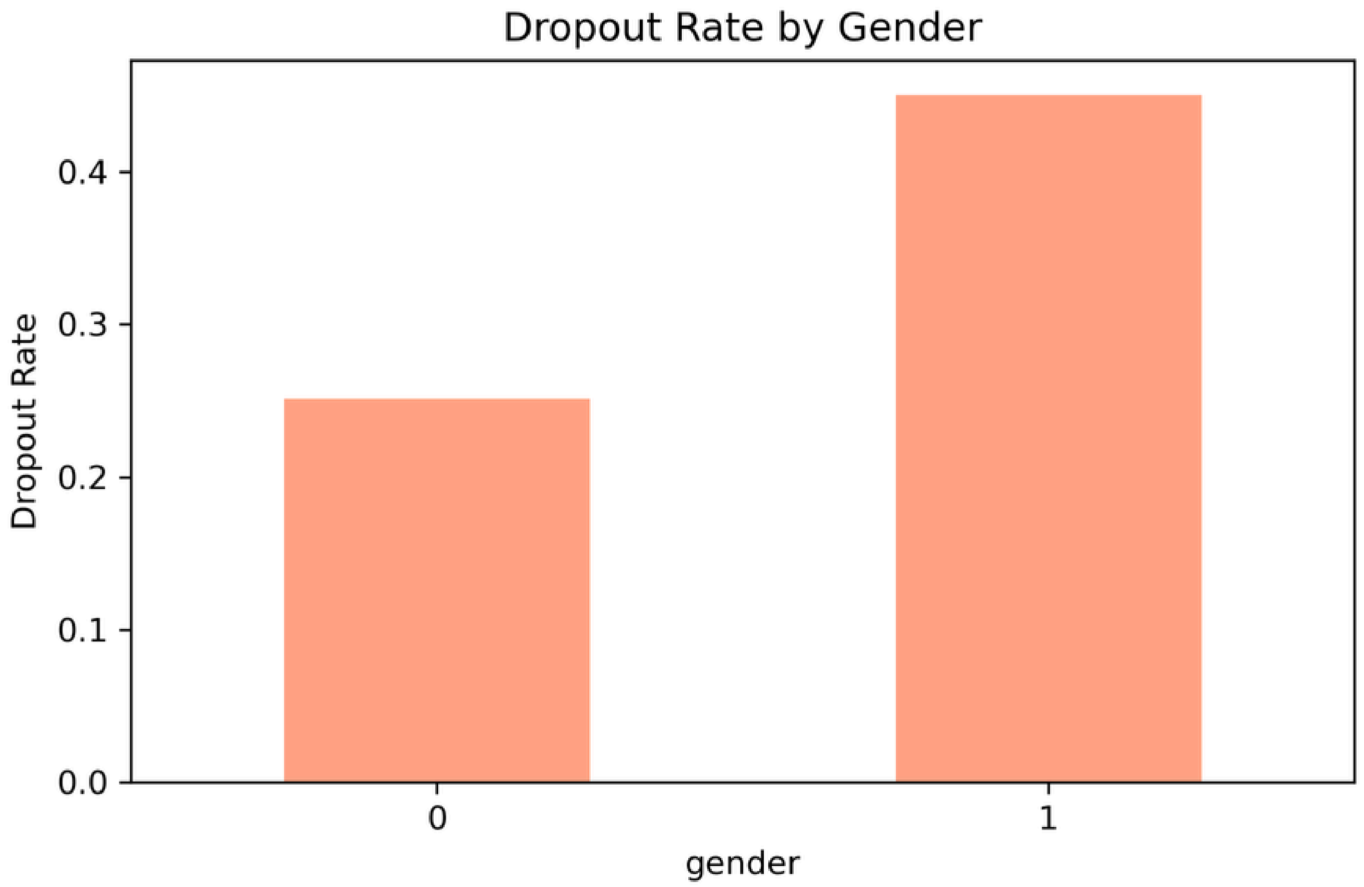
H_0: Dropout rates do not differ across age at enrollment.
H_1: Certain age groups have significantly higher dropout rates.

H_0: Dropout rates do not differ across genders.
H_1: Certain genders have significantly higher dropout rates.

Methods

Chi-square tests
Two-sample t-tests





Understanding the Problem

Supervised classification

FOCUSING ON:

Target variable:	Goal:	Features:
Target (Dropout = 0, Graduate = 1)	Predict if a student will drop out at enrollment	Academic, demographic, and socioeconomic indicators

Education
SERVICES LTD

Step-by-step Machine Learning Process

1. MODELS

This is a multi-class classification problem
We not only have Dropout vs Graduate —we also have Enrolled.
So we have two valid modeling choices:
Option A — 3-class classification

- Predict: Dropout, Enrolled, or Graduate
- Useful if we want to detect students at risk but still enrolled

Option B — Binary classification (very common in research)

- Merge Enrolled with Graduate
- Predict: Dropout vs Not-Dropout
- Useful if our main goal is early dropout prevention

2. CLASS BALANCE INSIGHT

This dataset is moderately imbalanced, but not extreme:

- Majority class: Graduate (~50%)
- Minority class: Enrolled (~18%)

Implications:

- Accuracy alone is NOT enough
- We must look at:
    - F1-score
    - Balanced accuracy
    - Recall for Dropout (very important in education)

Education
SERVICES LTD

Step-by-step Machine Learning Process

3. ML INTERPRETATION

From a modeling perspective:

- The model will naturally be better at predicting Graduate
- Without correction, it may miss Dropout cases
- This is a classic real-world bias-variance tradeoff:
  - ○ Bias → predicting majority class
  - ○ Variance → overfitting minority patterns

CHALLENGES

- Minority classes still misclassified despite SMOTE.
- Class 1 has lowest recall → model confuses it with classes 0 & 2.
- Categorical feature encoding impacts performance.

Model’s Performance

```
=== Cross-validated model comparison (train set) ===
 model accuracy balanced_acc macro_f1 weighted_f1
GaussianNB 0.24809 0.339897 0.242284 0.22563
```

Best by macro-F1: GaussianNB

```
=== Tuning RandomForest (RandomizedSearchCV, macro-F1) ===
Fitting 5 folds for each of 25 candidates, totalling 125 fits
Best CV macro-F1: 0.5601220105851845
Best params: {'clf__n_estimators': 300, 'clf__min_samples_split': 2, 'clf__min_samples_leaf': 2, 'clf__max_features': 'sqrt',
'clf__max_depth': 30}
```

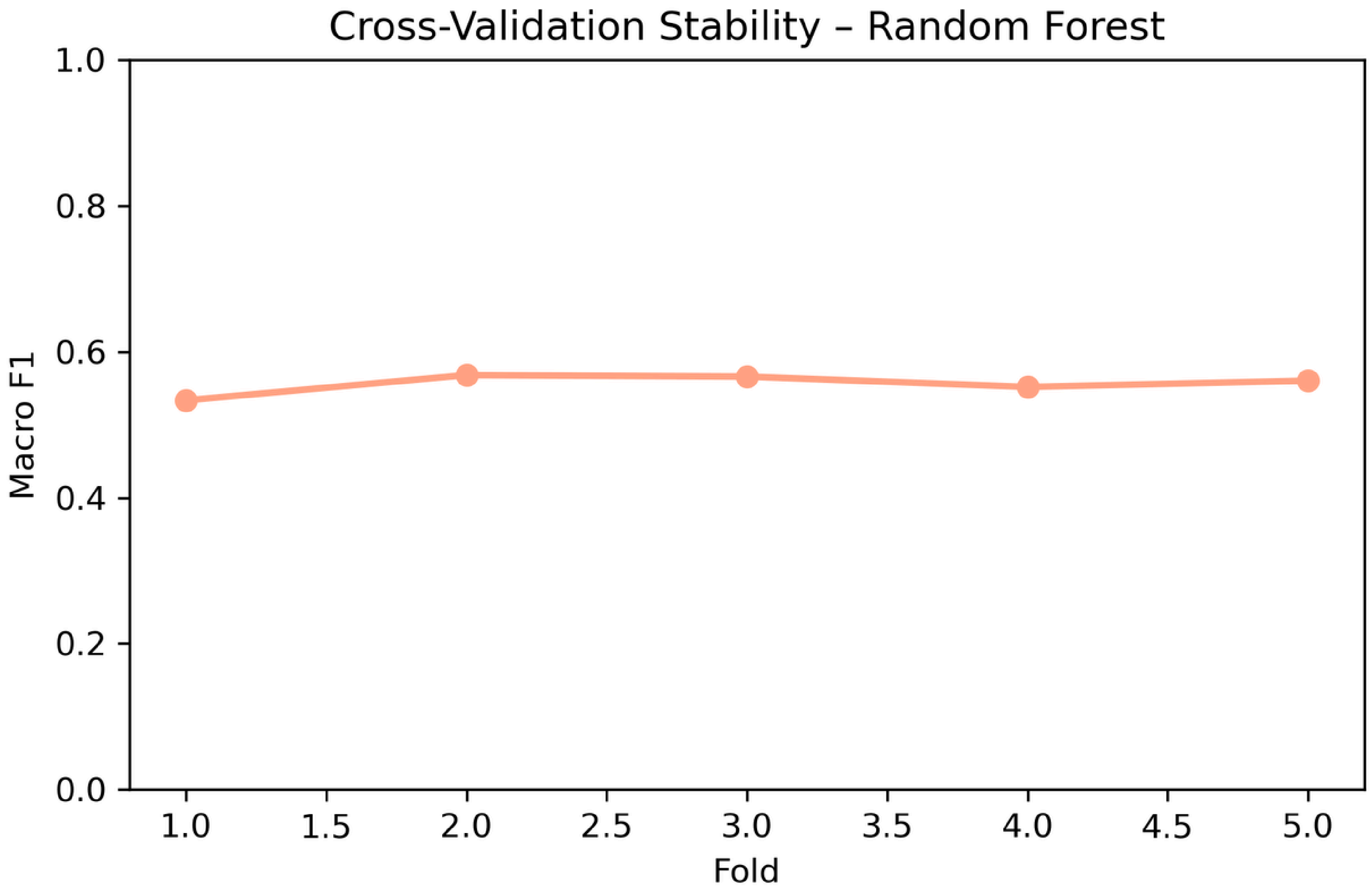
“Randomized hyperparameter tuning significantly improved model performance, achieving a macro-F1 score of 0.56. This indicates balanced predictive ability across all student outcome classes, particularly improving dropout detection. The results confirm that ensemble tree-based models are well-suited for capturing non-linear interactions in educational data.”

Key Observations

- About 57% of predictions are correct overall
- Macro F1 = average F1 across all classes equally
- 0.52 means overall performance across classes is moderate, with poor performance on Enrolled
- This metric adjusts for class imbalance
- 52.5% is just slightly better than random guessing
- Indicates the model is biased toward the majority classes, likely Graduate and Dropout

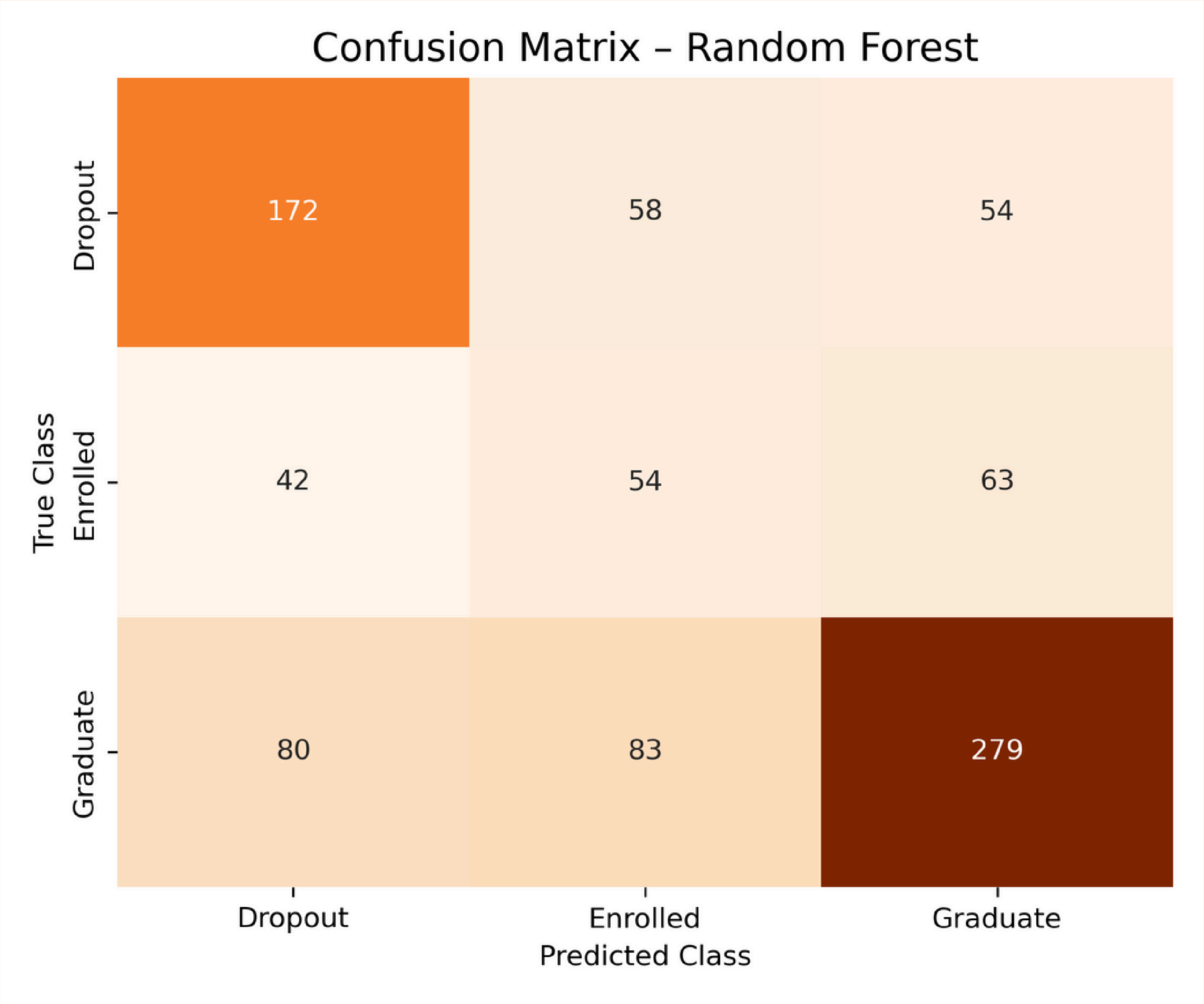
Model’s Perfomance

- “Random Forest model is stable across folds with consistent macro F1 (~0.55)”
- “The model predicts Dropout and Graduate better than Enrolled, as shown in the confusion matrix and class F1 scores”
- “Minority class (Enrolled) still challenging → potential for SMOTE, boosting, or feature engineering”



conclusions

- The model is doing best on the Dropout class (172 correct is high).
- It seems to confuse classes in the other categories (many values outside the diagonal).
- The bottom-right cell looks very dark, meaning the model may also do well on the last class.



The model can predict the majority class reasonably well, but minority classes—especially class 1—still underperform. Future work should focus on smarter feature engineering, improved categorical handling, advanced sampling, and robust model choice to increase macro F1 and balanced accuracy.

Thank You!

Veerpal Kaur CHATTHA