Dự Đoán Sinh Viên Bỏ Học Và Thành Công Học Tập

Phát triển mô hình machine learning để dự đoán kết quả học tập của sinh viên và hỗ trợ can thiệp kịp thời

THÀNH VIÊN NHÓM

1.	Huỳnh Trần Học Đăng	2210731
2.	Doãn Phương Hùng Cường	2310381

NỘI DUNG

1. **Giới thiệu dự án**

Tổng quan về mục tiêu và phạm vi nghiên cứu dự đoán kết quả học tập

2. **Khám phá dữ liệu**

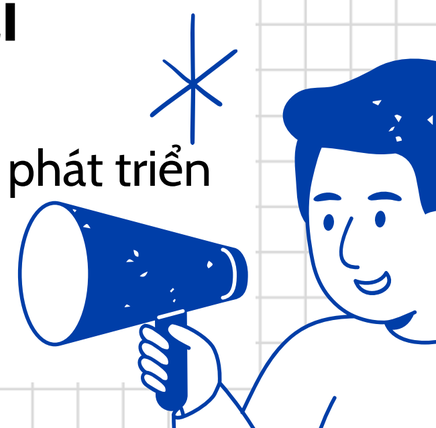
Phân tích dữ liệu sinh viên và các yếu tố ảnh hưởng đến kết quả học tập

3. **Mô Hình Machine Learning**

Tổng quan về mục tiêu và phạm vi nghiên cứu dự đoán kết quả học tập

4. **Kết Luận & Tương Lai**

Tổng kết kết quả và định hướng phát triển trong tương lai



Giới Thiệu Dự Án

Mục đích chính:

Dự án nhằm phát triển mô hình machine learning để dự đoán kết quả học tập của sinh viên: tốt nghiệp, vẫn đang học, hoặc bỏ học, sử dụng bộ dữ liệu "Predict Students' Dropout and Academic Success" từ UCI Machine Learning Repository.

Bộ dữ liệu này bao gồm các yếu tố nhân khẩu học, kinh tế xã hội, và học thuật, giúp mô hình dự đoán sớm để các trường đại học can thiệp, giảm tỷ lệ bỏ học và nâng cao hiệu quả giáo dục.

Giới Thiệu Dự Án

Tầm quan trọng:

Áp dụng kỹ thuật phân loại có giám sát vào dữ liệu giáo dục thực tế, nhấn mạnh các yếu tố đạo đức như tránh rò rỉ dữ liệu mục tiêu và xử lý mất cân bằng lớp.

Trong giáo dục, dự đoán sớm giúp hỗ trợ sinh viên gặp khó khăn, tránh lãng phí nguồn lực, và dự án này chứng minh cách dữ liệu có thể tạo ra insights để cải thiện tỷ lệ giữ chân sinh viên.

Phạm vi:

Tập trung vào dữ liệu đăng ký và học kỳ đầu tiên để đảm bảo dự đoán công bằng, không sử dụng dữ liệu sau đăng ký để tránh rò rỉ.

Mục tiêu cụ thể

Khám phá và tiền xử lý

Phân tích phân bố, tương quan và thiên kiến trong dữ liệu, tiền xử lý với kỹ thuật tạo đặc trưng không rò rỉ

Chọn Lọc Đặc Trưng

Xác định và loại bỏ đặc trưng gây rò rỉ mục tiêu, tập trung vào dữ liệu thời điểm đăng ký

Phát Triển Mô Hình

Xây dựng mô hình cơ bản và tinh chỉnh sử dụng Random Forest và Gradient Boosting, với kỹ thuật chia dữ liệu phân tầng và cân bằng lớp.

Tối Ưu & Đánh Giá

Sử dụng tìm kiếm ngẫu nhiên với kiểm chứng chéo để tối ưu siêu tham số, đánh giá bằng accuracy, F1-score macro, precision, recall, ROC-AUC.

Giải thích và insights

Phân tích tầm quan trọng đặc trưng để đưa ra khuyến nghị cho giáo dục.

Khám Phá Dữ Liệu - Tổng Quan

4424	36	0	50%
Instances	Features	Missing value	Gradute

Phân Bố Mục Tiêu

Graduate: 50%, Dropout: 32%, Enrolled: 18% - cho thấy mất cân bằng lớp. Có thể làm mô hình thiên vị về lớp đa số, nên cần kỹ thuật cân bằng như class weight để cải thiện dự đoán lớp thiểu số.ằn

Loại Đặc Trưng

20 đặc trưng số và 16 đặc trưng phân loại

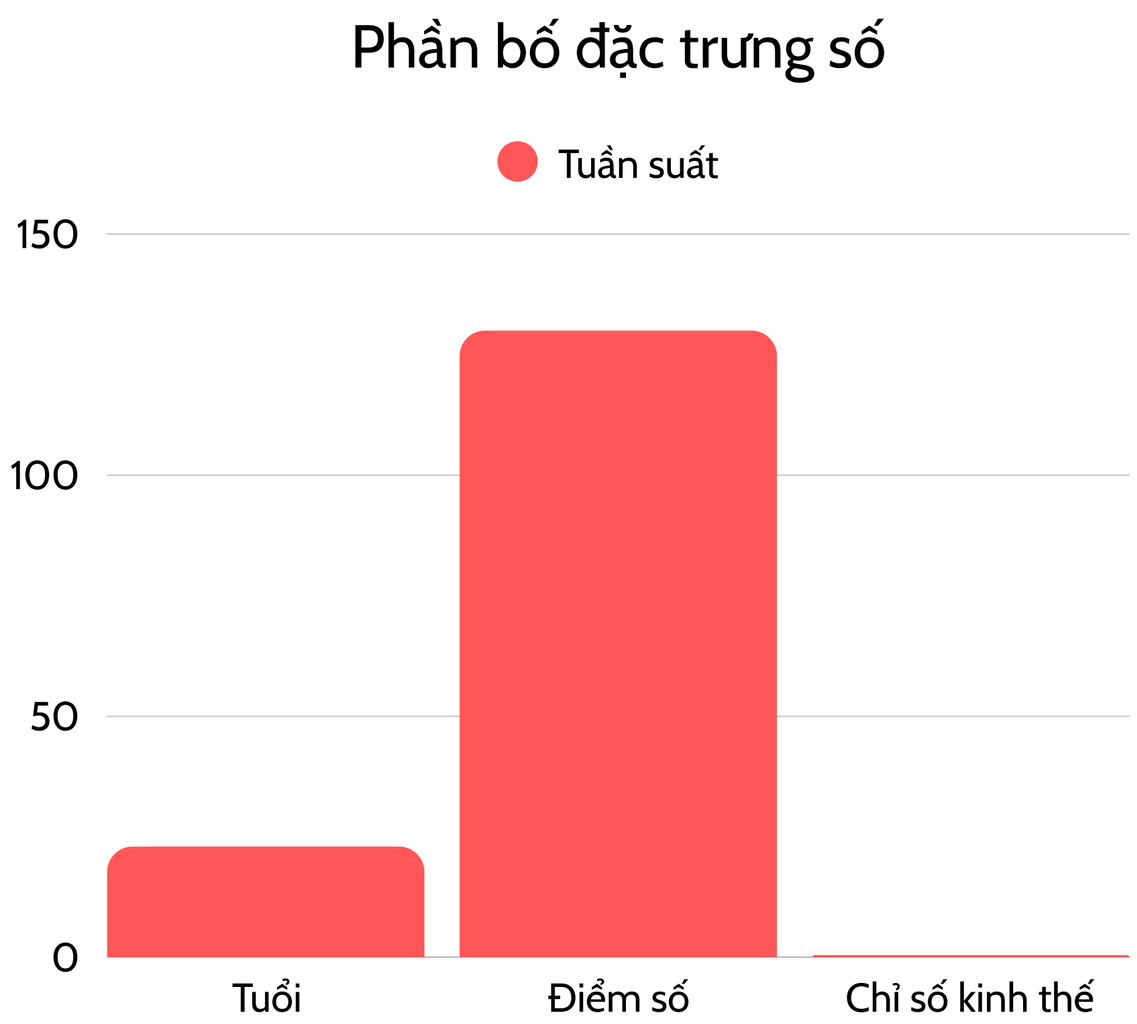
- Điểm số và tuổi
- Hôn nhân và khóa học
- Kinh tế xã hội

Chất Lượng Dữ Liệu

Không có giá trị thiếu hoặc trùng lặp

- Dữ liệu sạch
- Kiểm tra bất thường
- Chuẩn bị tốt cho ML

Khám Phá Dữ Liệu - Phân Bố & Tương Quan

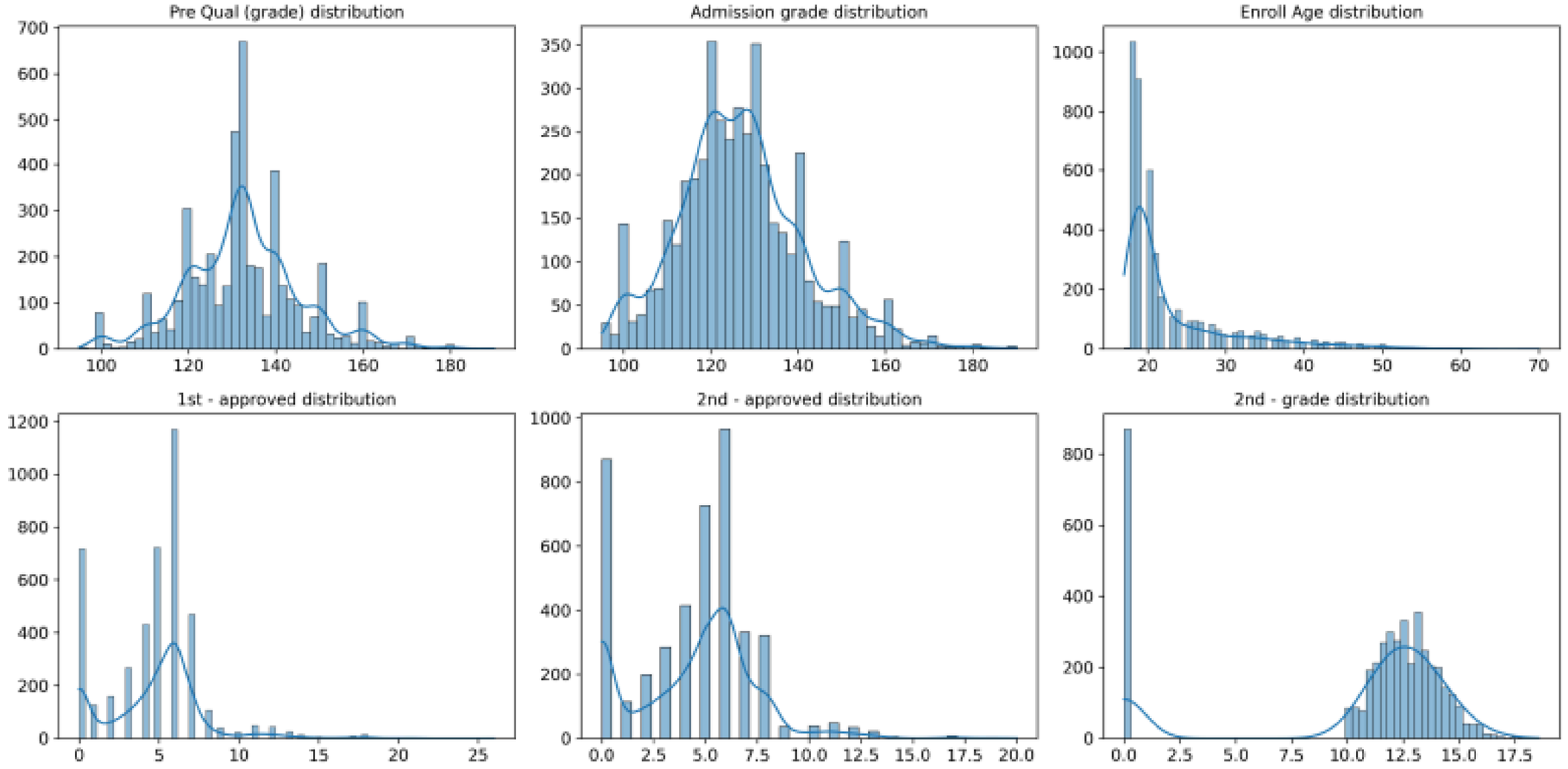


Phân bố đặc trưng

Tuổi lệch phải, điểm số đa đỉnh, cần biến đổi để gần chuẩn hơn

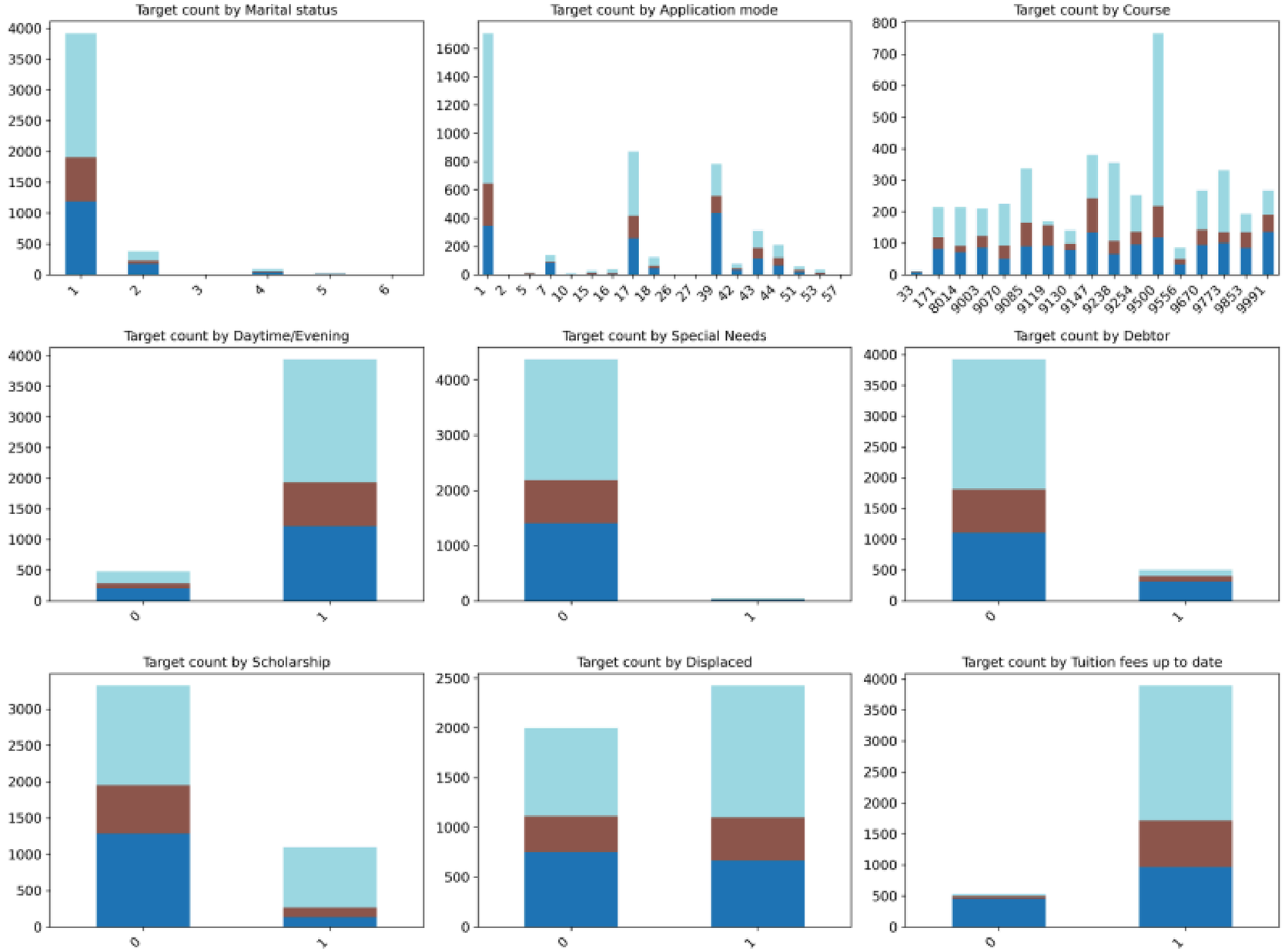
Tương quan

Mạnh giữa điểm học kỳ 1-2 (>0.7), yếu với chỉ số kinh tế (<0.1)

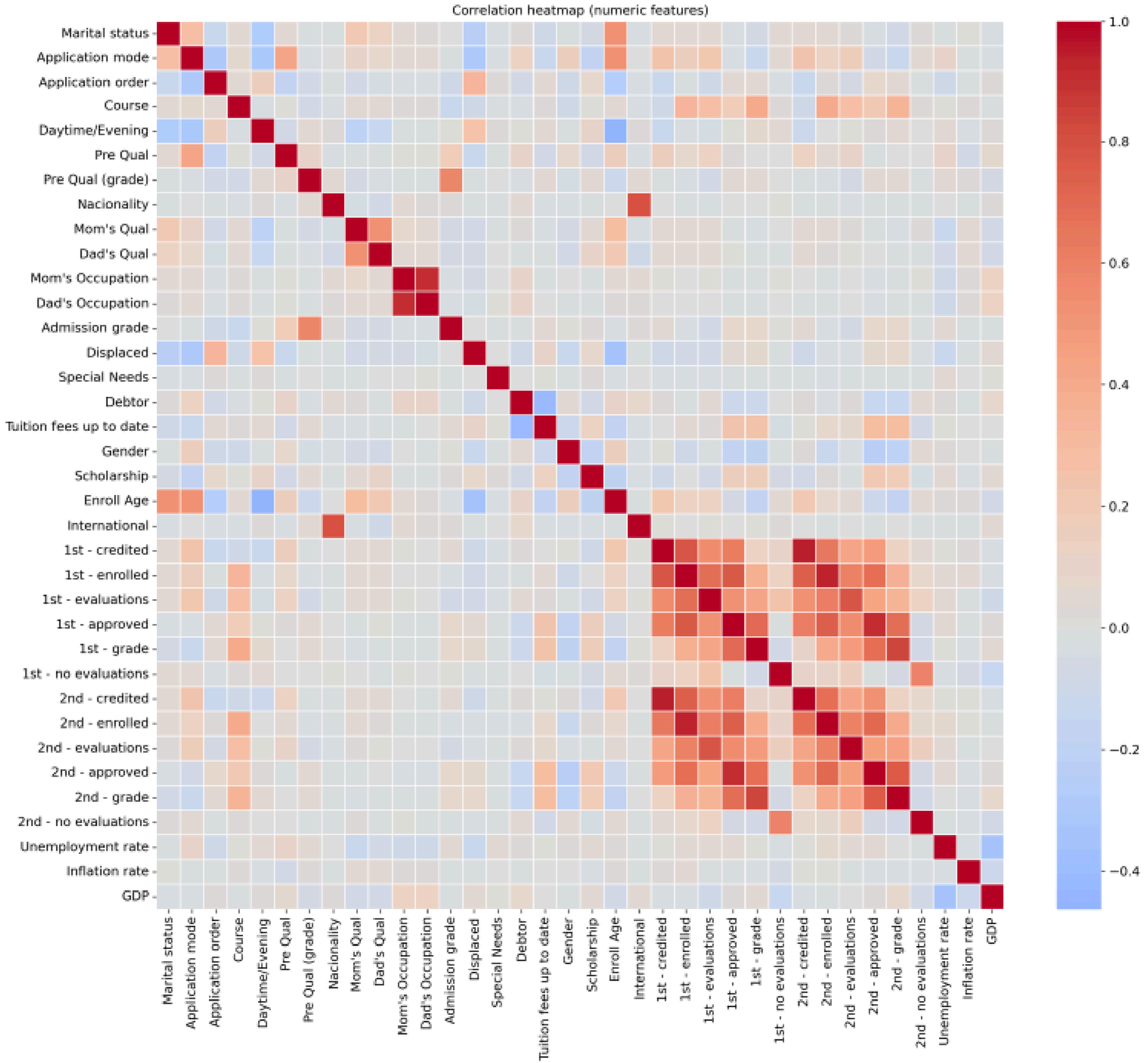


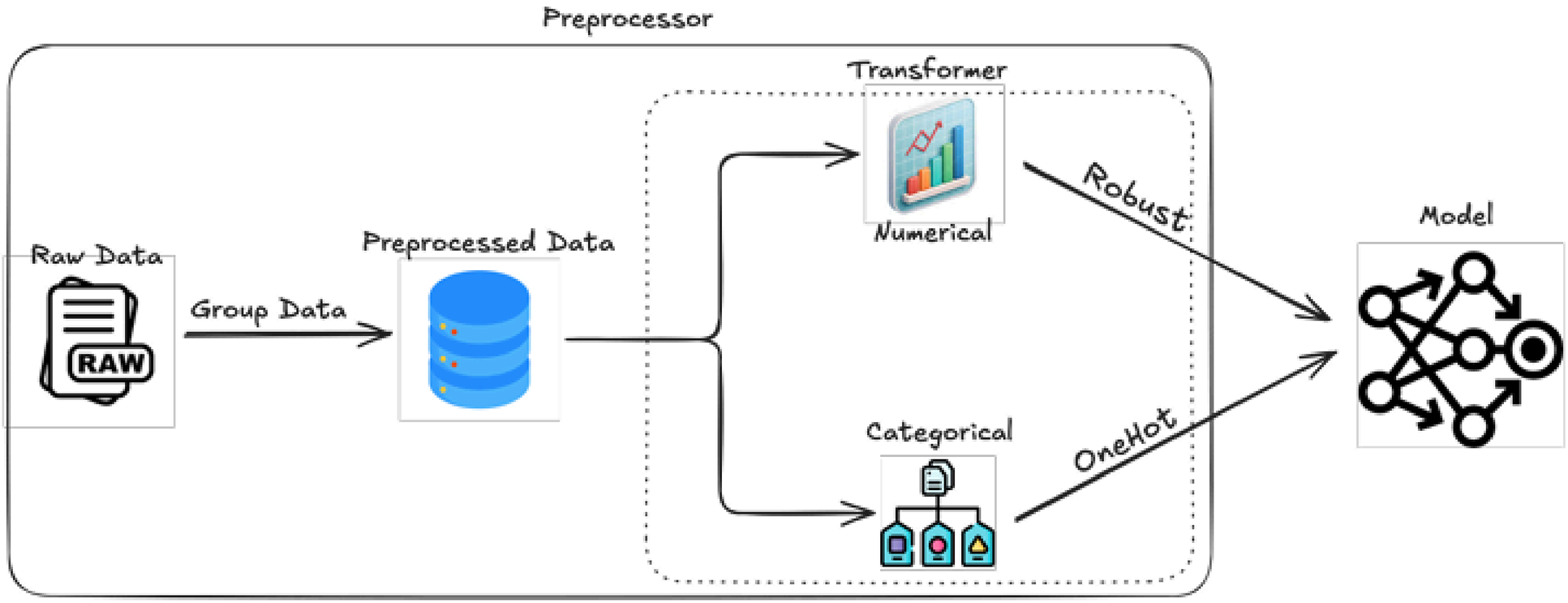
Numerical Feature Distributions

Key Categorical Features' Distributions



Correlation Heatmap





Pretrain Pipeline

Tiền Xử Lý & Tạo Đặc Trưng

Phát Hiện Bất Thường

Loại bỏ tuổi <17 hoặc >70, điểm ngoài 95-190, kiểm tra chỉ số học kỳ để đảm bảo dữ liệu hợp lý

Nhóm Đặc Trưng

Nhóm hôn nhân, khóa học, quốc tịch để giảm chiều và cải thiện hiệu quả mô hình

Tạo Đặc Trưng Mới

Tỷ lệ phê duyệt, điểm trung bình, delta cải thiện, nhóm tuổi để chuẩn hóa và nắm bắt xu hướng

Mô Hình Cơ Bản - Lựa Chọn

Random Forest

Giảm overfitting qua bagging, phù hợp dữ liệu hỗn hợp và cung cấp tầm quan trọng đặc trưng

S

W

Gradient Boosting

Sửa lỗi tuần tự, nắm bắt mẫu phức tạp, hiệu suất tốt với dữ liệu mất cân bằng

Đánh Giá Ban Đầu

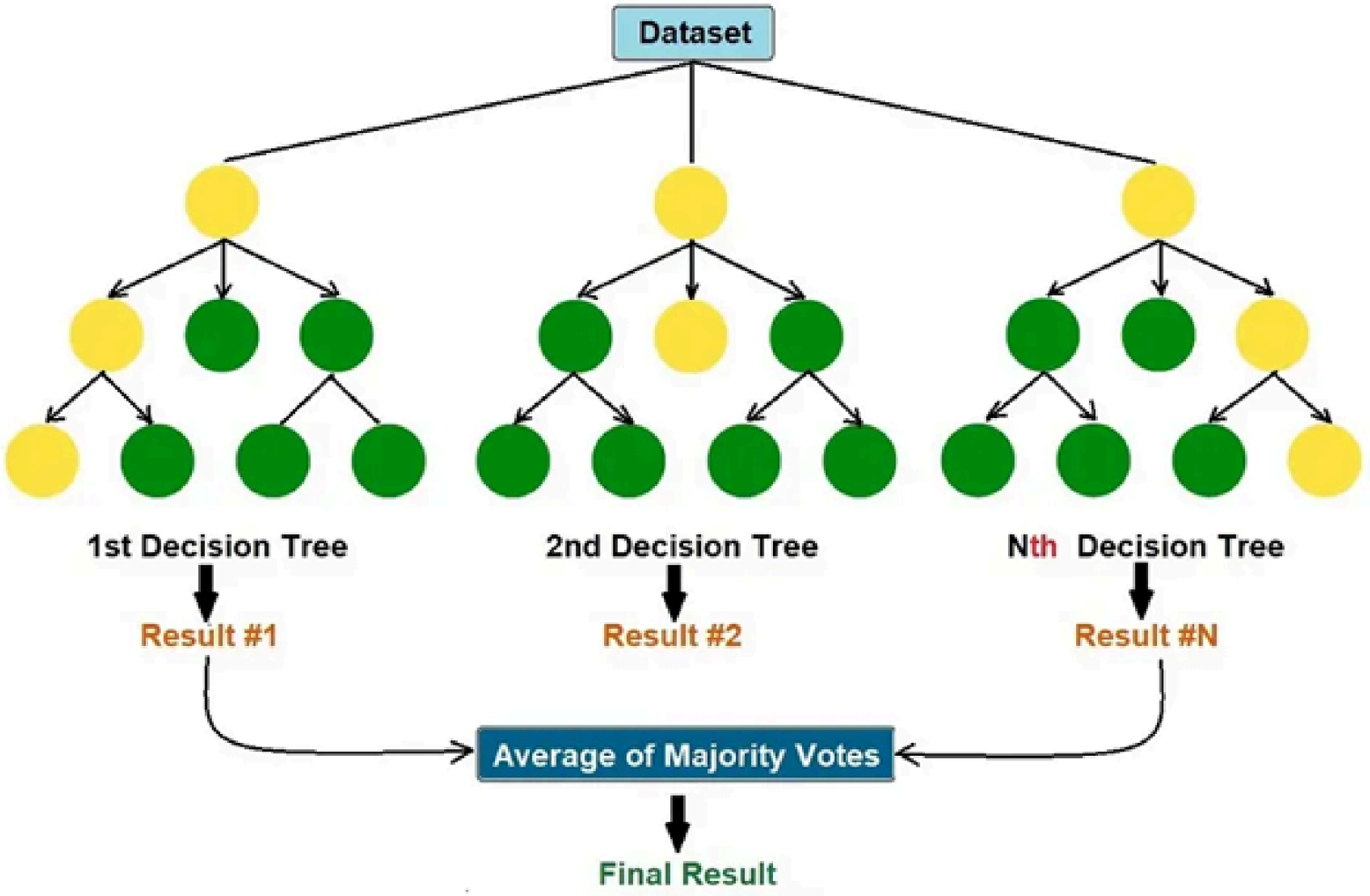
Accuracy ~0.78, F1_macro ~0.71, ROC-AUC ~0.88.

T

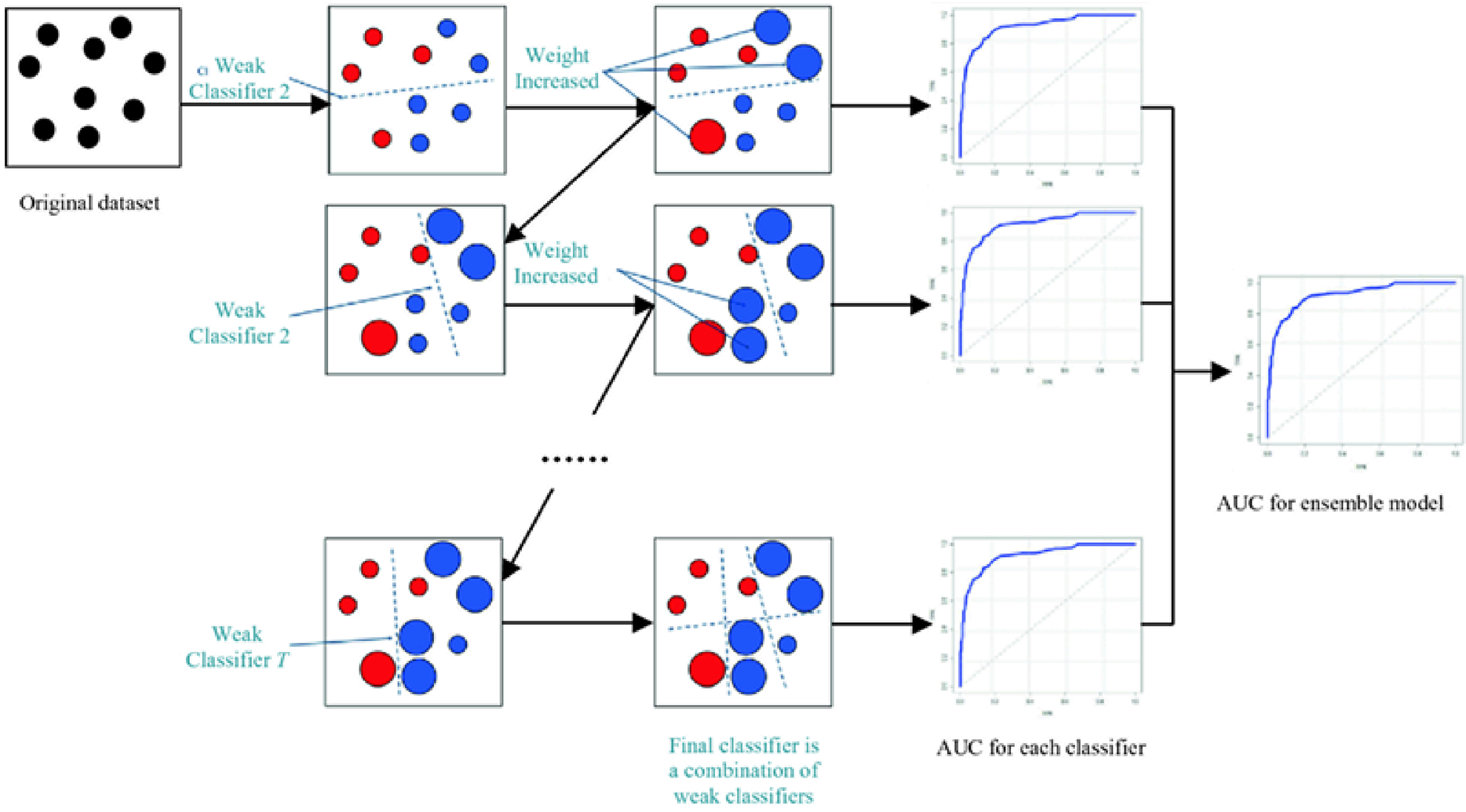
O

Pipeline Chuẩn

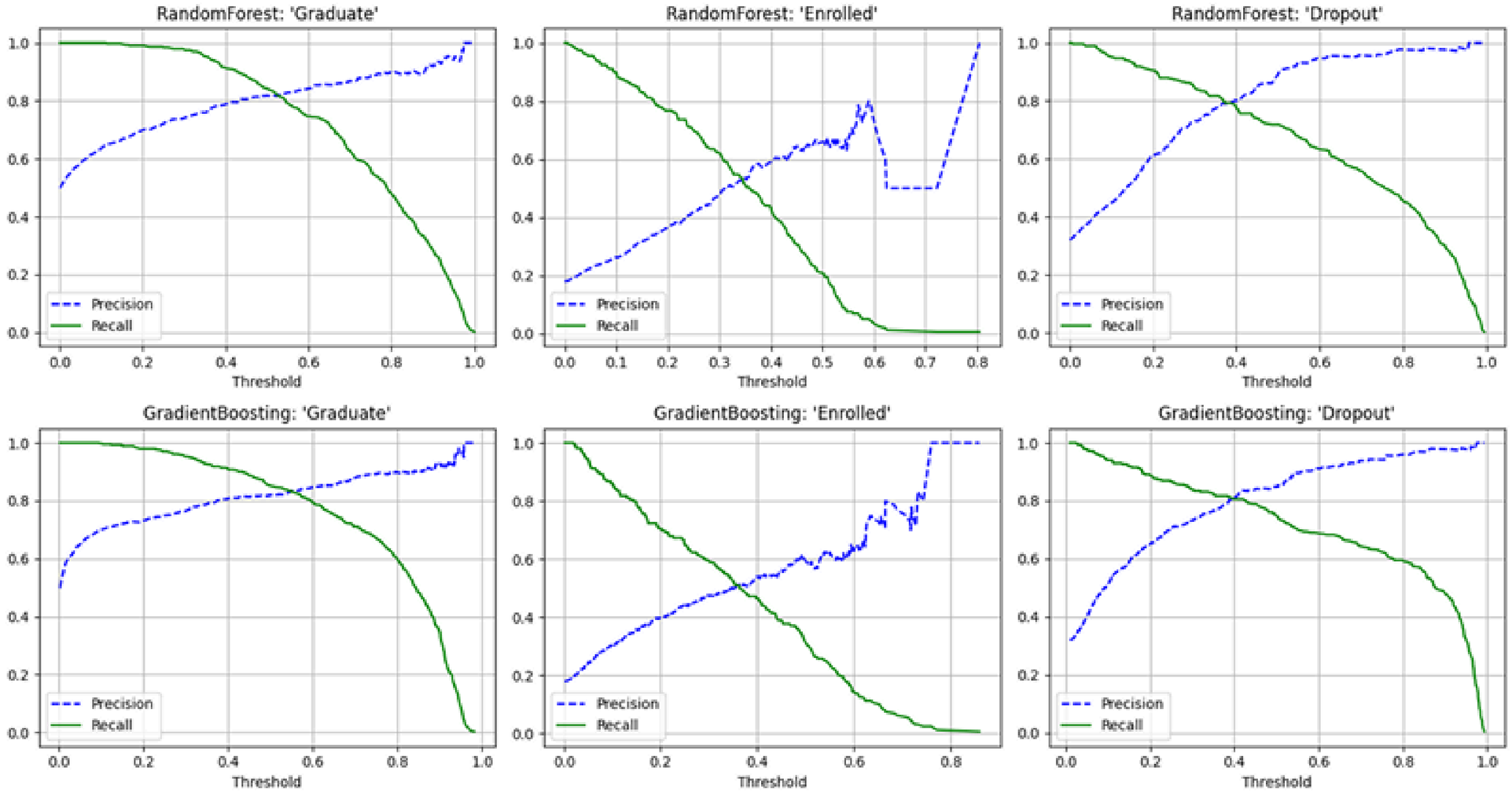
RobustScaler cho số, OneHotEncoder cho phân loại, chia dữ liệu phân tầng không rò rỉ



Random Forest Classifier

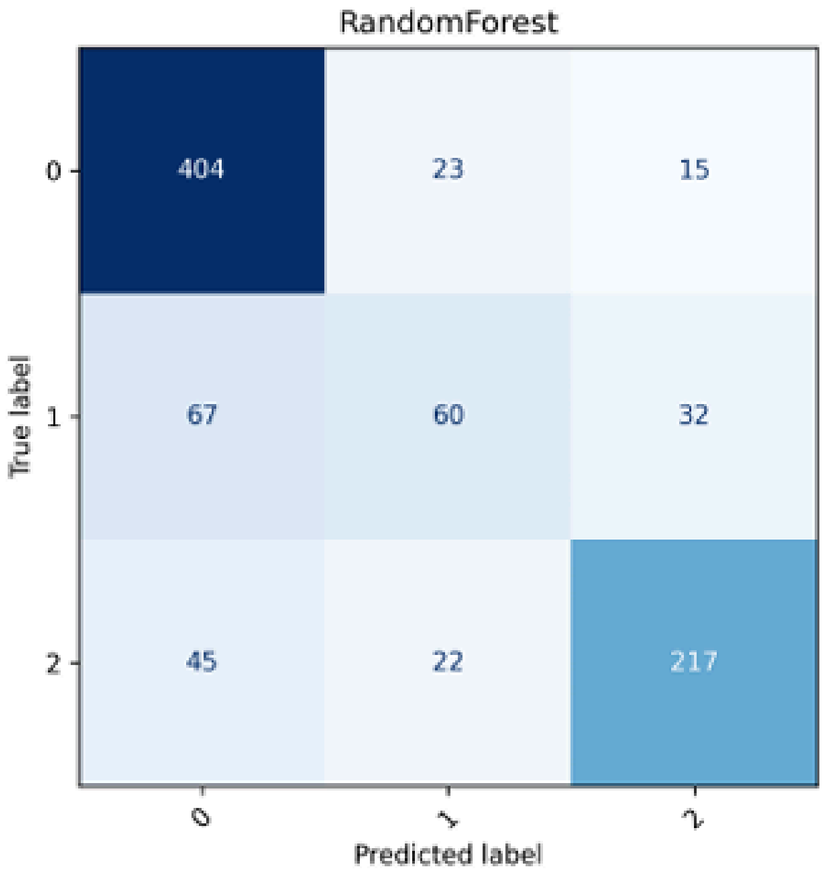


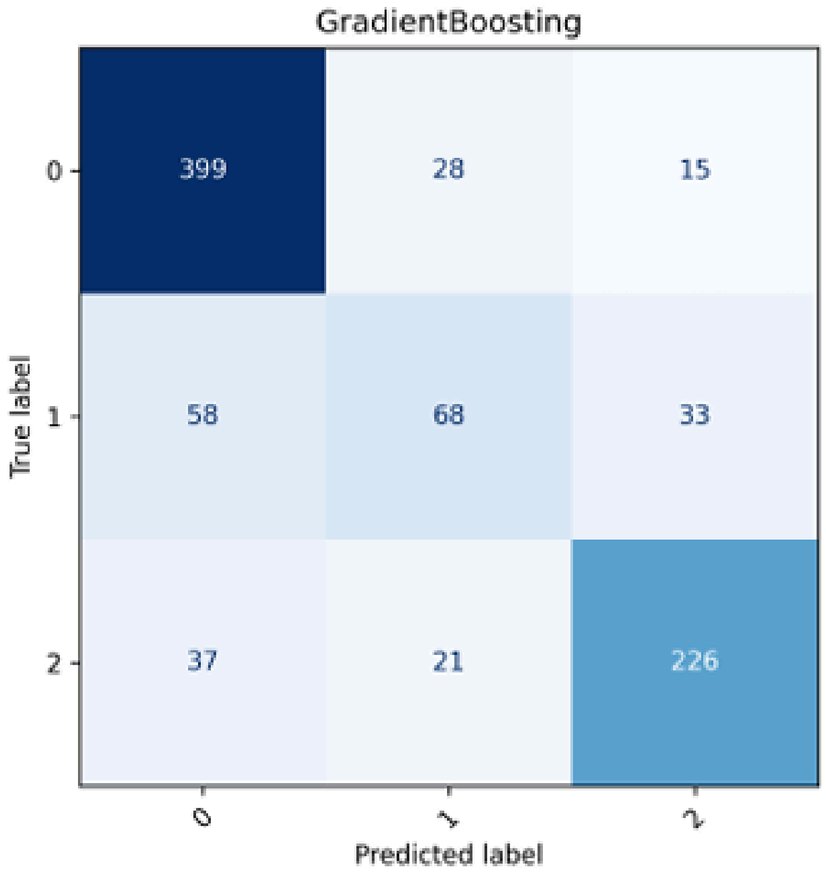
Gradient Boosting Classifier



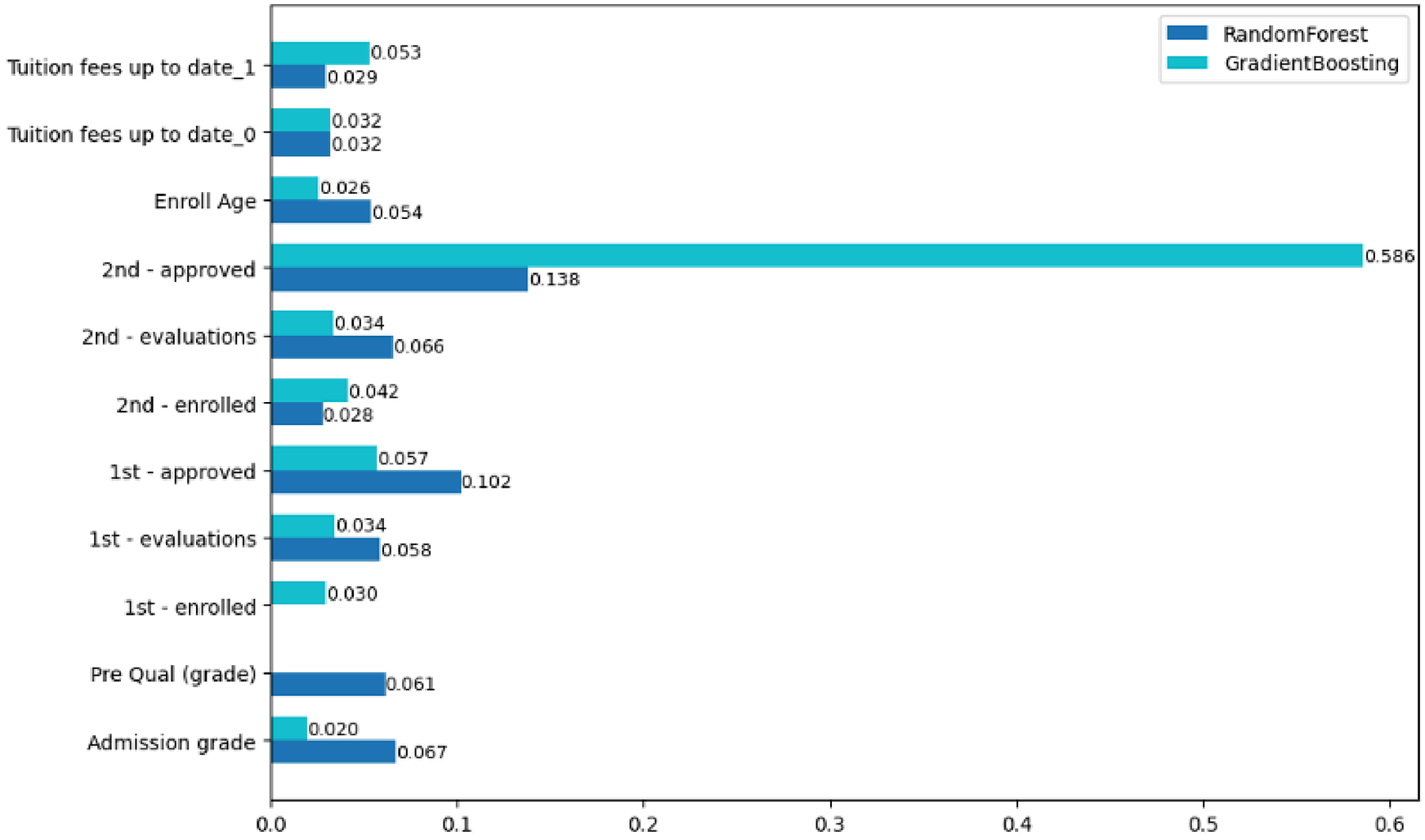
Precision- Recall Curve

	accuracy	f1_macro	f1_weighted	precision_macro	precision_weighted	recall_macro	recall_weighted	roc_auc_ovr
RandomForest	0.769492	0.696647	0.757046	0.725448	0.757467	0.685157	0.769492	0.878347
GradientBoosting	0.783051	0.718451	0.774273	0.737902	0.772495	0.708721	0.783051	0.882816



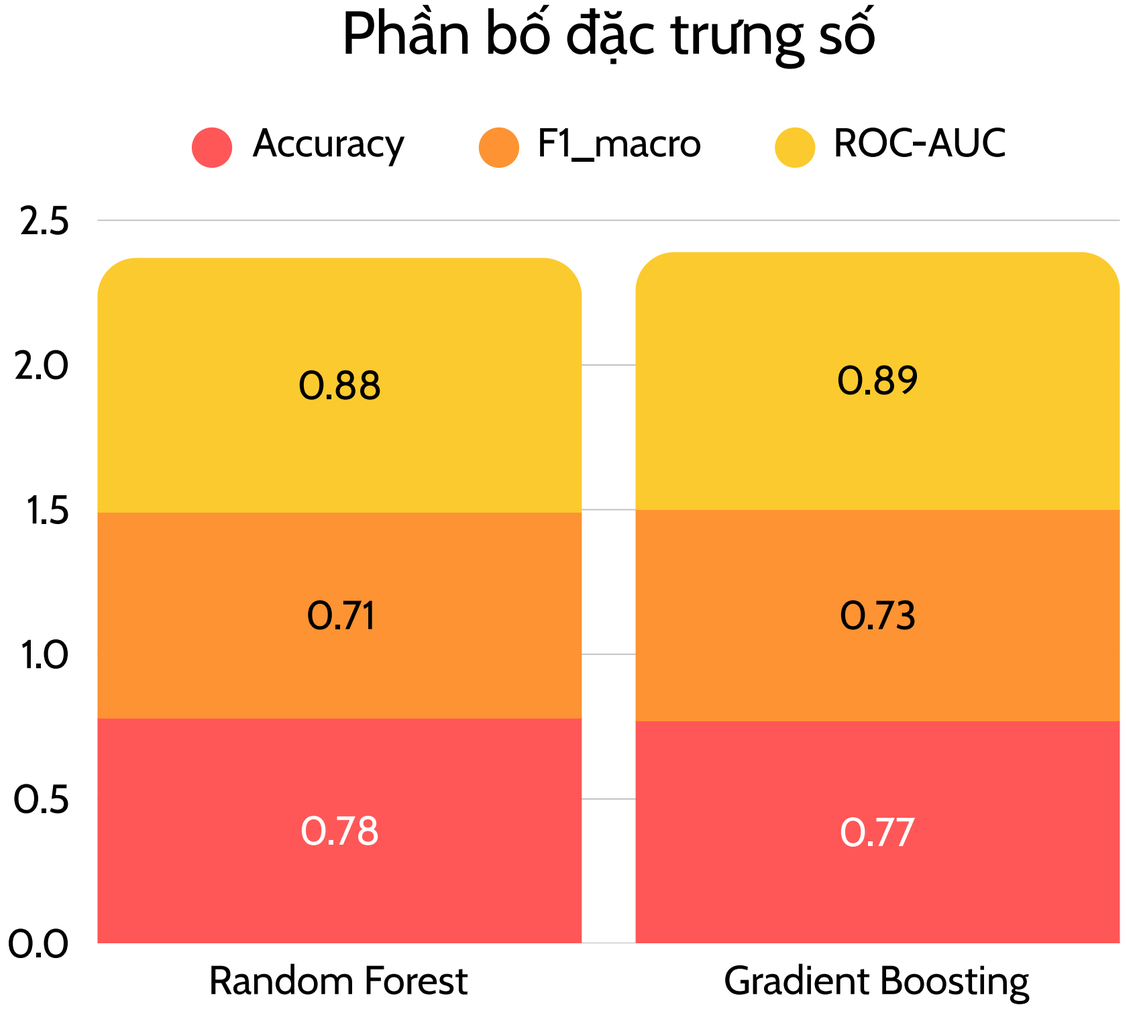


Some Evaluation Metrics



Top 10 Feature Importances

Kết Quả Mô Hình



Insights Từ Ma Trận Nhầm Lẫn

Gradient Boosting tốt hơn ở lớp Graduate nhưng cả hai đều nhầm Enrolled với Dropout do mất cân bằng

- Cần tuning để tăng recall lớp thiểu số
- 2nd_approved là đặc trưng quan trọng nhất
- Phân bố tầm quan trọng không đều

Cải Tiến Mô Hình - Engineering & Tuning

Tạo Đặc Trưng Nâng Cao

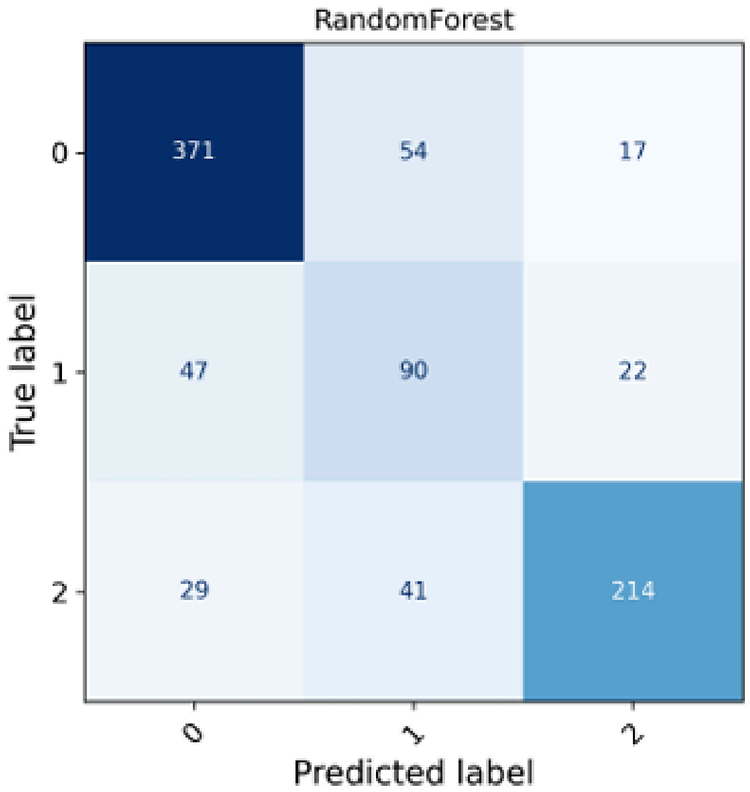
Thêm tỷ lệ và delta giúp mô hình tập trung vào hiệu suất tương đối, cải thiện F1_macro lên ~0.72

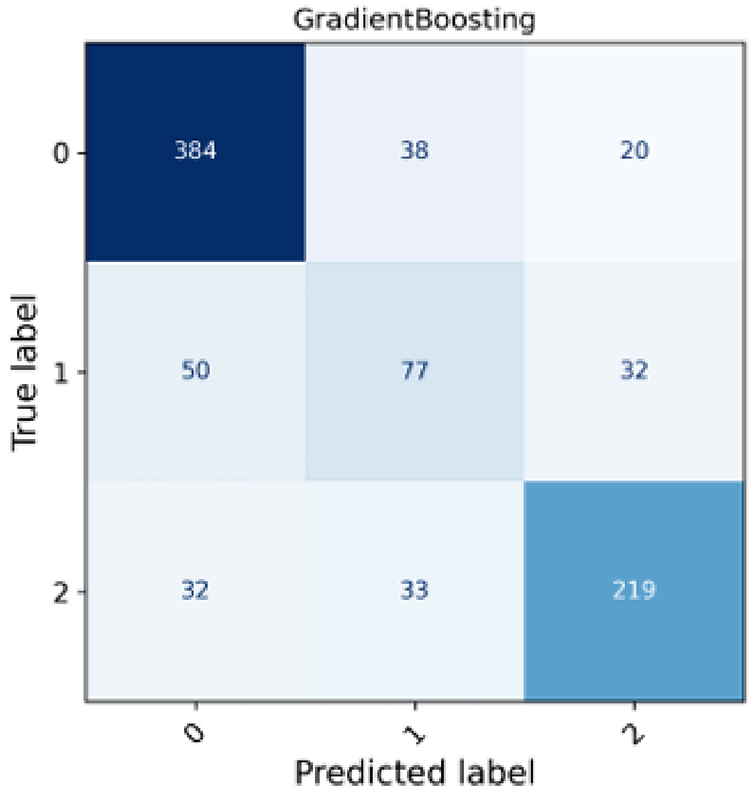
Tối Ưu Siêu Tham Số

RandomizedSearchCV với StratifiedKFold và scorer F1_macro, hiệu quả hơn grid search truyền thống

Kết Quả Cải Tiến

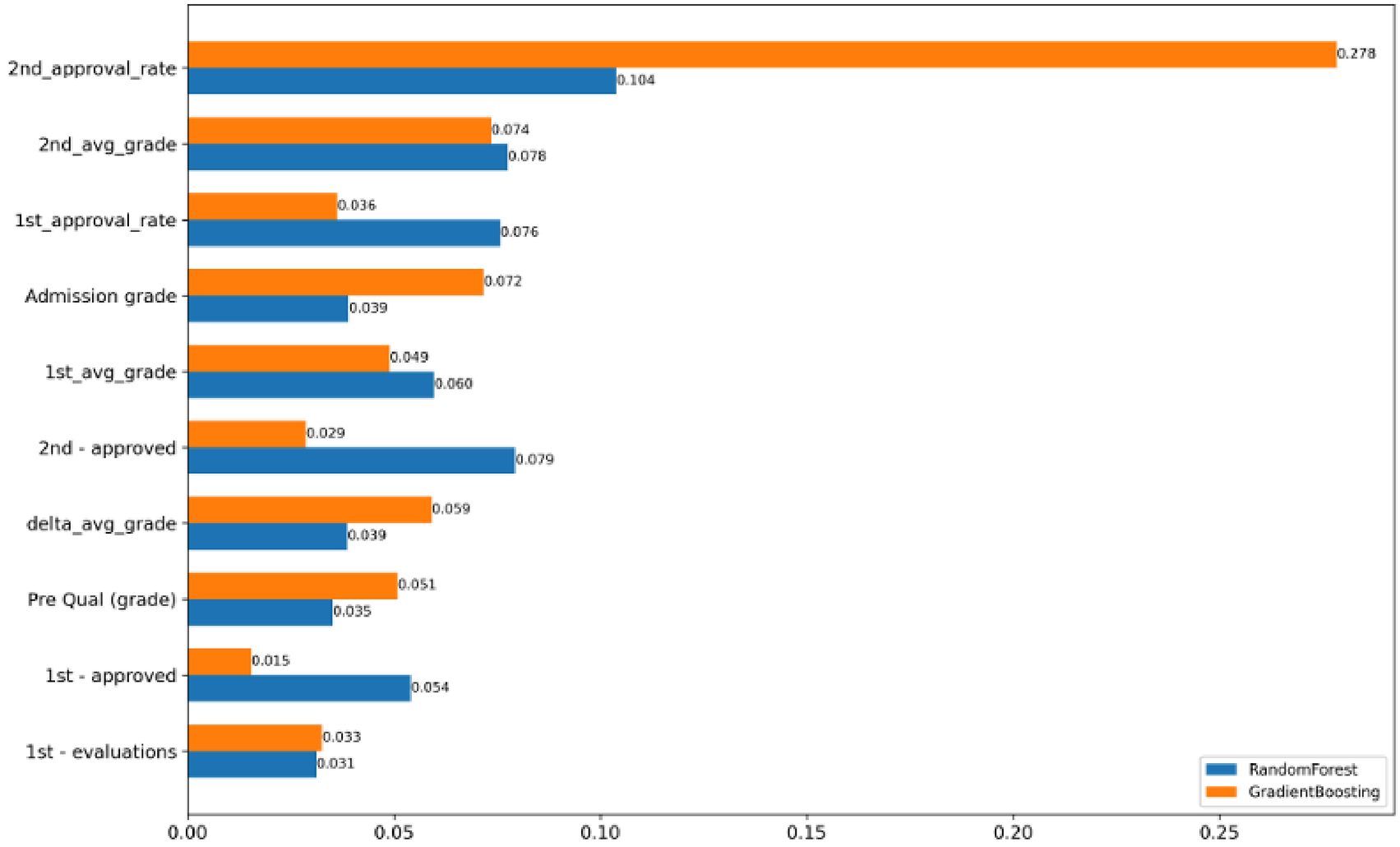
Accuracy ~0.76-0.77, F1_macro tăng 3-5%, ROC-AUC ổn định - hy sinh accuracy nhẹ để cân bằng lớp





Model	Accuracy	F1_Macro	F1_Weighted	Precision_Macro	Recall_Macro	ROC_AUC
RandomForest	0.762712	0.718307	0.766627	0.720771	0.719642	0.887097
GradientBoosting	0.768362	0.712211	0.765807	0.717474	0.708061	0.881817

Hypertuned Models' Metrics



Enhanced Models' Top 10 Feature Importances

Phương Pháp Ensemble

1

Soft Voting

Kết hợp Random Forest và Gradient Boosting qua trung bình xác suất, tận dụng độ tin cậy dự đoán

2

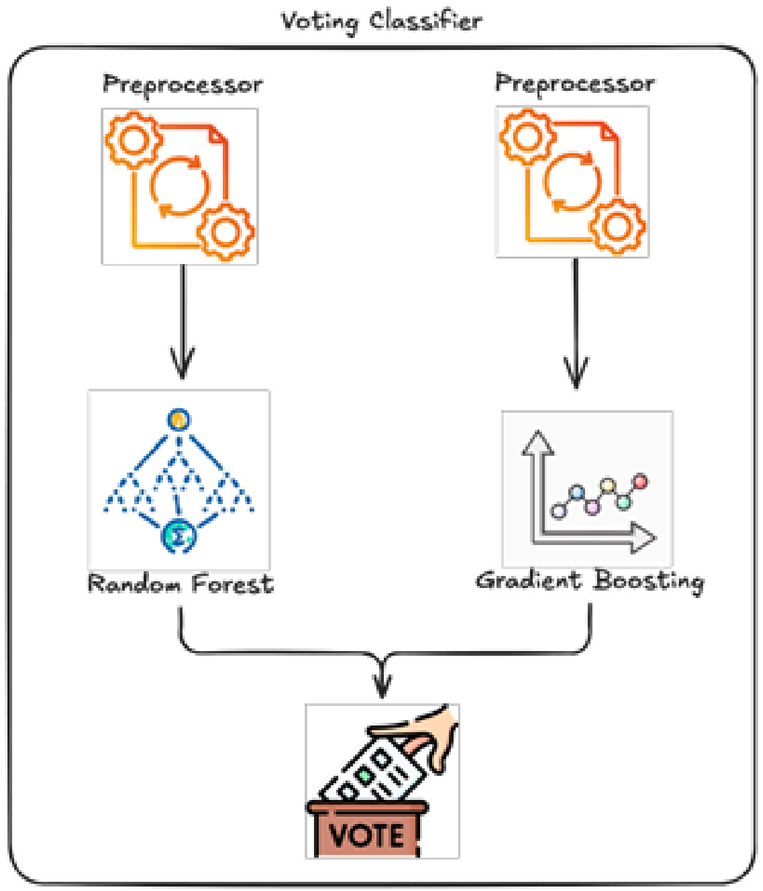
Lý Do Chọn

Kết hợp sức mạnh hai mô hình để giảm yếu điểm cá nhân và cải thiện hiệu suất tổng thể

3

Kết Quả Ensemble

Đạt accuracy cao nhất 0.776, cân bằng tốt nhất cho đa lớp mất cân bằng




```python
ensemble = VotingClassifier(
  estimators=[
      ('rf', rf_search.best_estimator_),
      ('gb', gb_search.best_estimator_)
  ],
  voting='soft') # To be discussed
ensemble.fit(X_train, y_train)

```

Voting Classifier Pipeline

So Sánh Kết Quả Cuối Cùng

Mô hình	Accuracy	F1_marco	ROC-AUC
Random Foreast	0.78	0.71	0.88
Gradient Boosting	0.77	0..73	0.89
Ensemble	0.776	0.718	0.887

0.776

Accuracy Cao Nhất

0.718

F1_marco Cân Bằng

Kết Luận Và Hướng Phát Triển

Tóm tắt thành công

Dự án xây dựng mô hình hiệu quả chứng minh giá trị machine learning trong giáo dục, từ EDA đến ensemble hoàn chỉnh

Hạn chế:

- Dữ liệu cụ thể, thiếu thời gian thực
- Có thể thiên kiến
- Cần kiểm tra tổng quát hóa

Tương lai:

- Thêm neural networks
- Tích hợp web app
- Mở rộng đặc trưng ngoài

Tác Động Giáo Dục

Giúp trường xác định sinh viên rủi ro sớm để can thiệp kịp thời

Cảm Ơn Sự Lắng Nghe Của Mọi Người

Câu hỏi và thảo luận được hoan nghênh!