Student Dropout Prediction

Comprehensive Data Analysis Report

Supervisor Requirements Analysis

Dataset Overview and Modeling Results

4,424 Students | 34 Features | 3 Classes | 6 Models

December 2025

European Higher Education Institution

Contents

1 Executive Summary

This comprehensive report presents a detailed analysis of student dropout prediction in higher education, addressing all requirements specified by the thesis supervisor. The analysis encompasses dataset exploration, feature engineering, feature selection optimization, multiple machine learning models, explainable AI techniques, and rigorous evaluation metrics.

1.1 Key Findings

- **Dataset**: 4,424 students with 34 features across academic, financial, and demographic categories
- **Class Distribution**: Dropout (32.1%), Enrolled (17.9%), Graduate (49.9%)
- **Best Overall Model**: Random Forest achieving 76.72% test accuracy with 91.36% AUC
- **Best Cross-Validation**: XGBoost with 78.21% mean CV accuracy
- **Top Predictors**: Curricular units approved (both semesters), tuition fees, and semester grades
- **Feature Selection**: Optimized from 34 to 10-30 features depending on model type
- **Explainable AI**: SHAP analysis completed for all 6 models

1.2 Supervisor Requirements Coverage

Table 1: Analysis Coverage of Supervisor Requirements

Req.	Description	Status
1-3	Dataset Overview (Students, Features, Classes)	
4-6	Feature Lists (Academic, Financial, Demographic)	
7	Feature Ranking	
8	Dropout Feature Importance	
9	Multi-Model Classification (6 models)	
10	Explainable AI (SHAP for all models)	
11.1	Accuracy, Precision, Recall, F1-Score	
11.2	Confusion Matrices	
11.3	ROC Curves & AUC	
11.4	10-Fold Cross-Validation	

2 Dataset Overview

2.1 Total Students and Features

The dataset contains comprehensive information about **4,424 students** enrolled in various degree programs at a European higher education institution. The analysis focuses on predicting student outcomes across three classes:

- **Dropout**: 1,421 students (32.1%)
- **Enrolled**: 794 students (17.9%)
- **Graduate**: 2,209 students (49.9%)

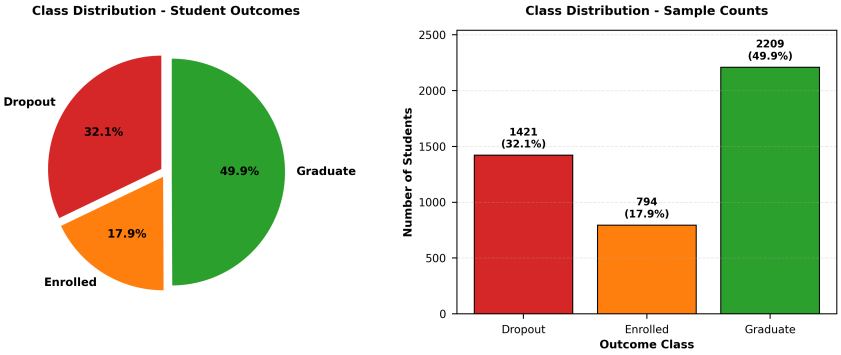

Figure 1: Distribution of student outcomes across three classes

2.2 Feature Categories

The dataset comprises **34 features** organized into three main categories:

Table 2: Feature Categories and Counts

Category	Number of Features
Academic Features	18
Financial Features	12
Demographic Features	16
Total Unique	**34**

3 Feature Lists

3.1 Academic Features (18 features)

Academic features capture student performance, enrollment patterns, and qualifications:

1. Curricular units 1st sem (credited)
2. Curricular units 1st sem (enrolled)
3. Curricular units 1st sem (evaluations)
4. Curricular units 1st sem (approved)
5. Curricular units 1st sem (grade)
6. Curricular units 1st sem (without evaluations)
7. Curricular units 2nd sem (credited)
8. Curricular units 2nd sem (enrolled)
9. Curricular units 2nd sem (evaluations)
10. Curricular units 2nd sem (approved)
11. Curricular units 2nd sem (grade)
12. Curricular units 2nd sem (without evaluations)
13. Previous qualification grade
14. Admission grade
15. Application mode
16. Application order
17. Course
18. Daytime/evening attendance

3.2 Financial Features (12 features)

Financial features include tuition status, scholarships, and economic indicators.

3.3 Demographic Features (16 features)

Demographic features capture personal and family background including marital status, parent qualifications and occupations, gender, age, nationality, and special needs status.

4 Feature Ranking

Five different feature ranking methods were applied to identify the most important predictors. Figure 2 shows how different methods rank the top features.

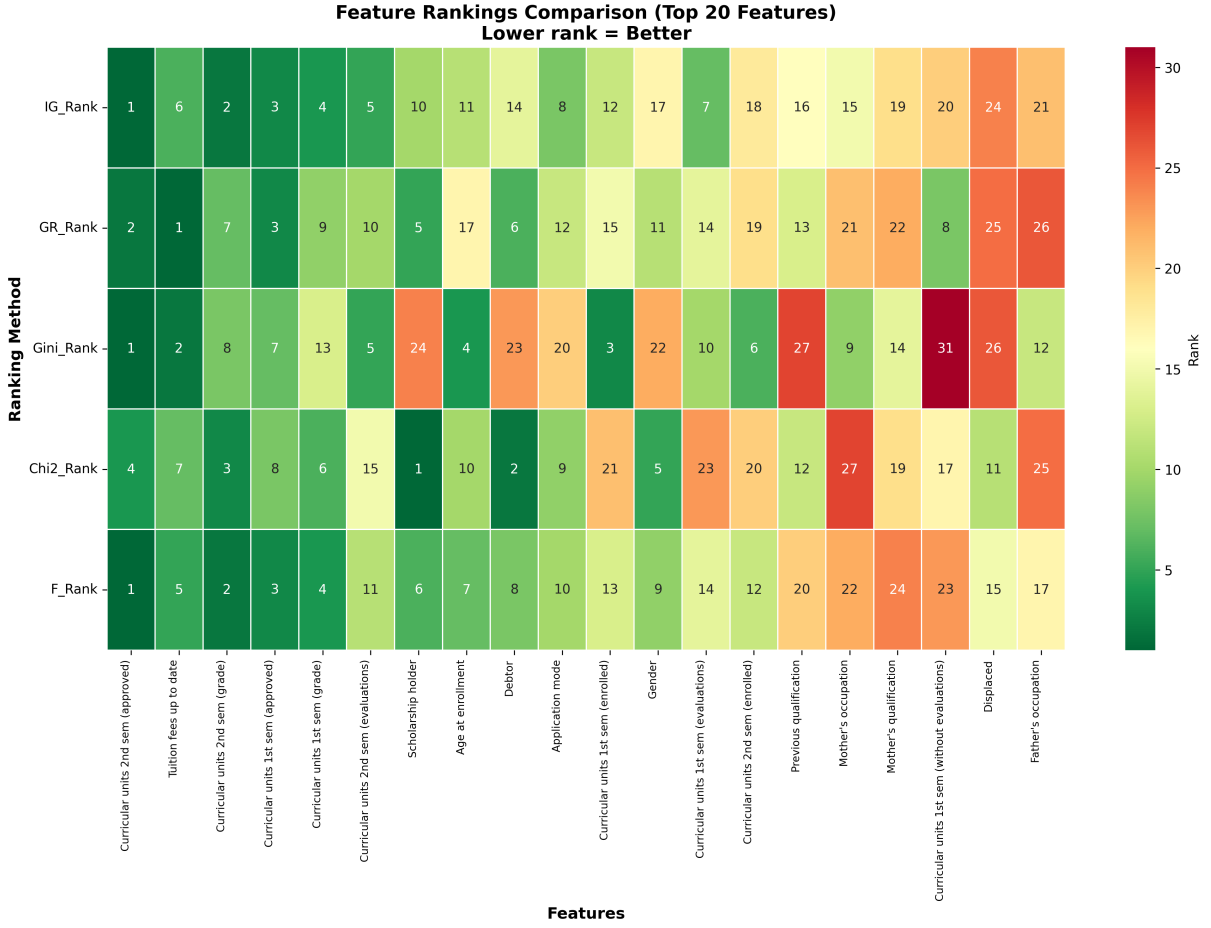

Figure 2: Feature ranking heatmap comparing all five methods for top 20 features

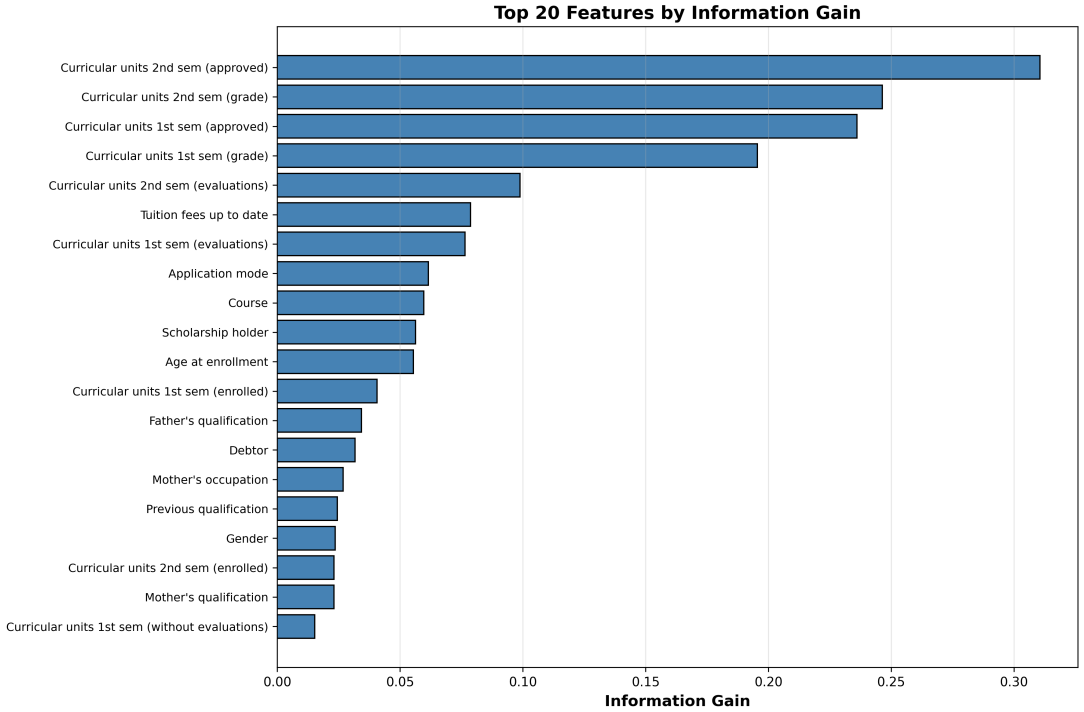

Figure 3: Top 20 features ranked by Information Gain

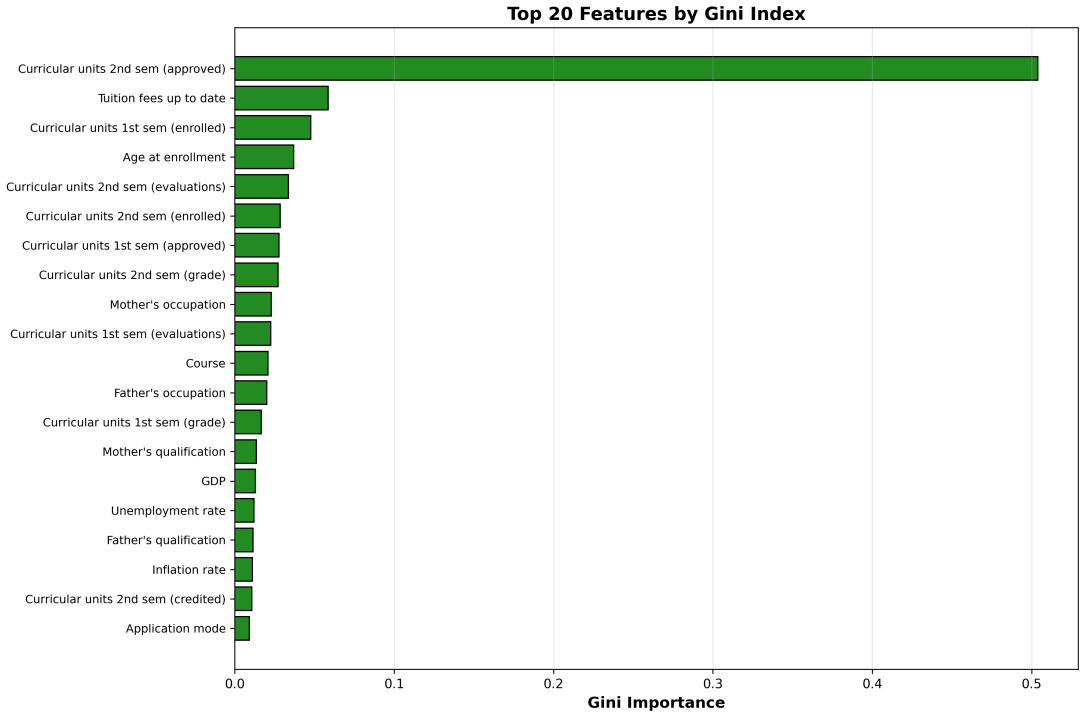

Figure 4: Top 20 features ranked by Gini importance

Key Finding: Curricular units 2nd semester (approved) and tuition fees status consistently rank in the top 3 across all methods.

5 Dropout Feature Importance

A focused analysis identified the most influential features for predicting student dropout using four complementary methods.

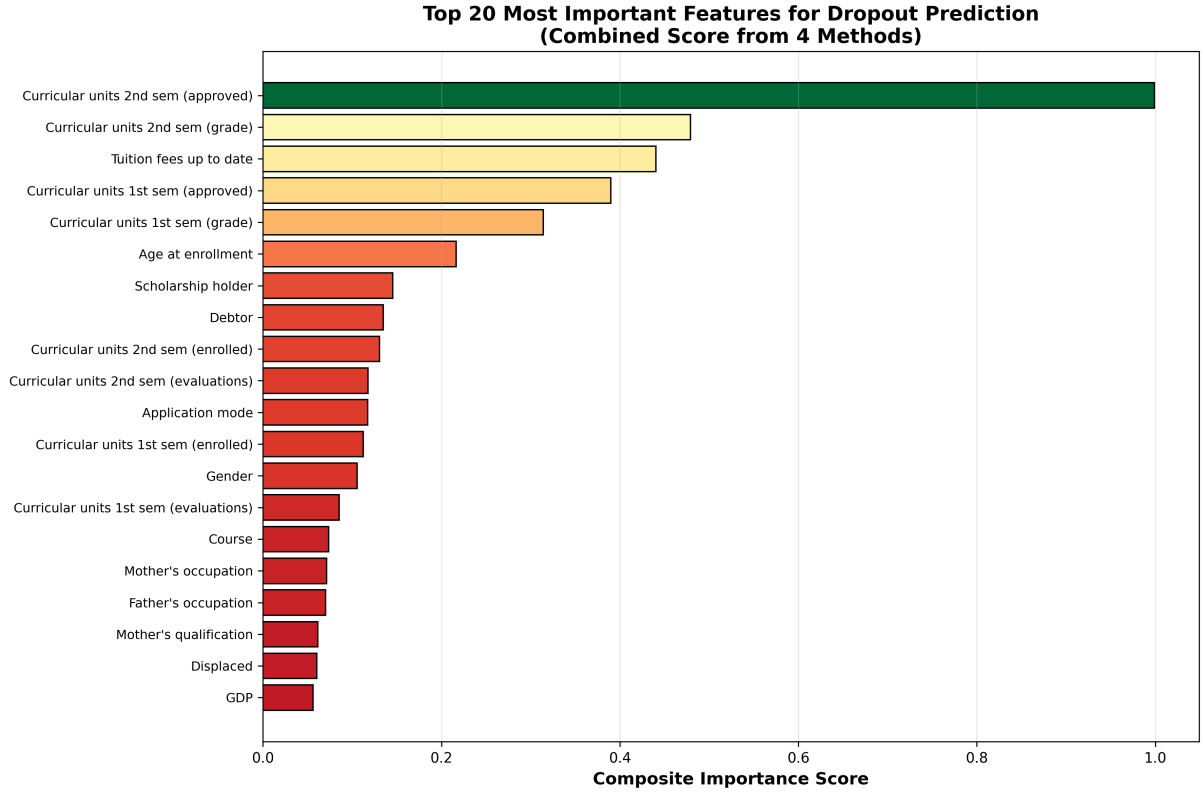

Figure 5: Top 20 features for dropout prediction (composite score from 4 methods)

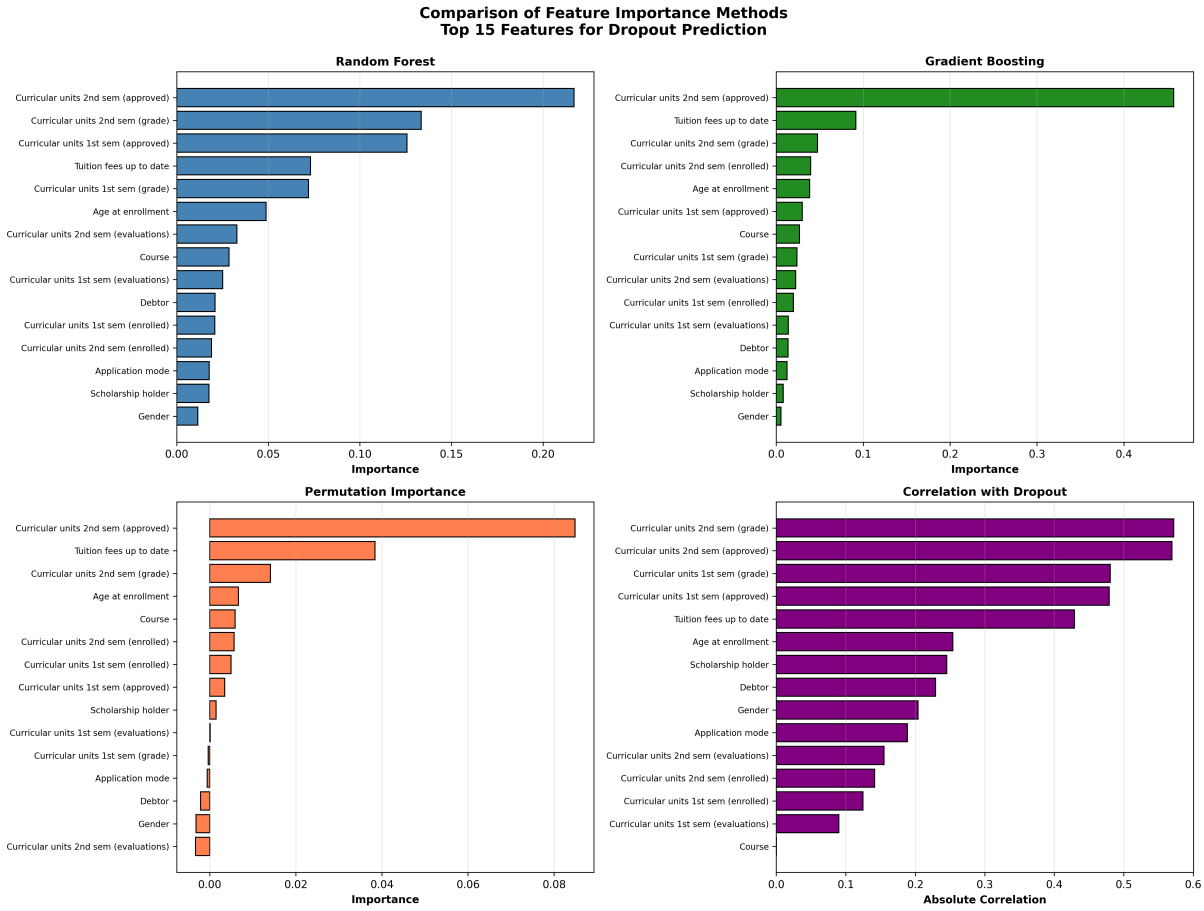

Figure 6: Comparison of four feature importance methods for dropout prediction

Top 5 Dropout Predictors:

1. Curricular units 2nd sem (approved)
2. Curricular units 2nd sem (grade)
3. Tuition fees up to date
4. Curricular units 1st sem (approved)
5. Curricular units 1st sem (grade)

6 Feature Selection Optimization

Comprehensive feature selection was performed for all 6 models using 9 different methods to identify optimal feature subsets.

6.1 Single Classifiers: Decision Tree & Naive Bayes

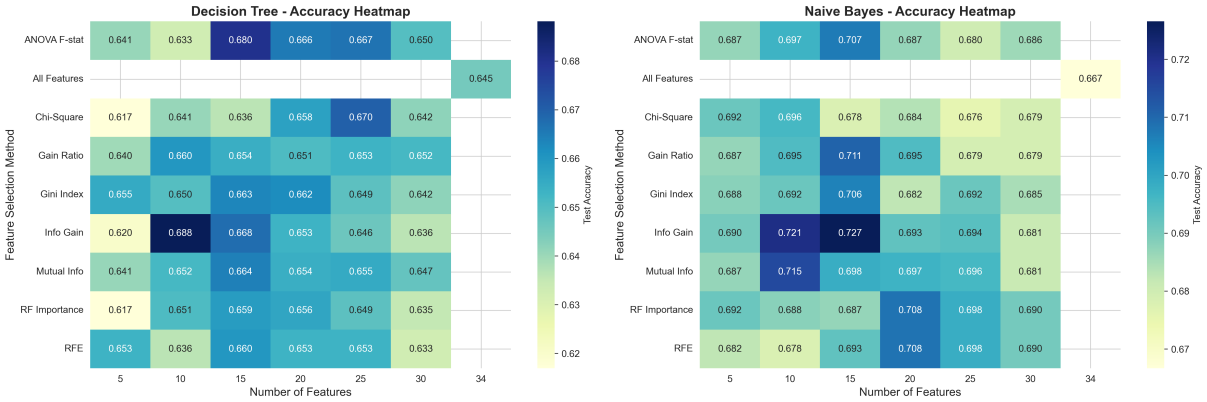

Figure 7: Accuracy heatmap for Decision Tree and Naive Bayes across all feature selection methods and feature counts

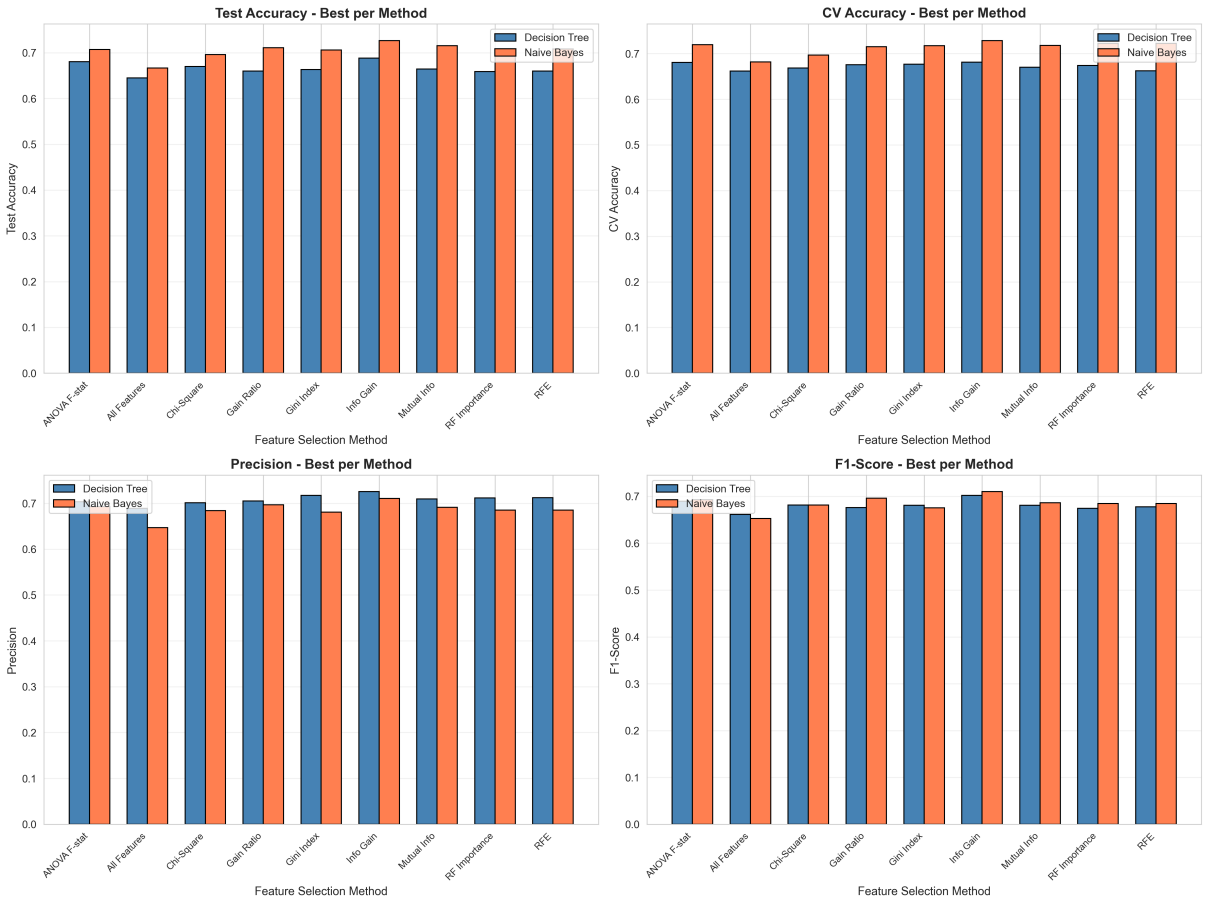

Figure 8: Comprehensive metrics comparison for Decision Tree and Naive Bayes

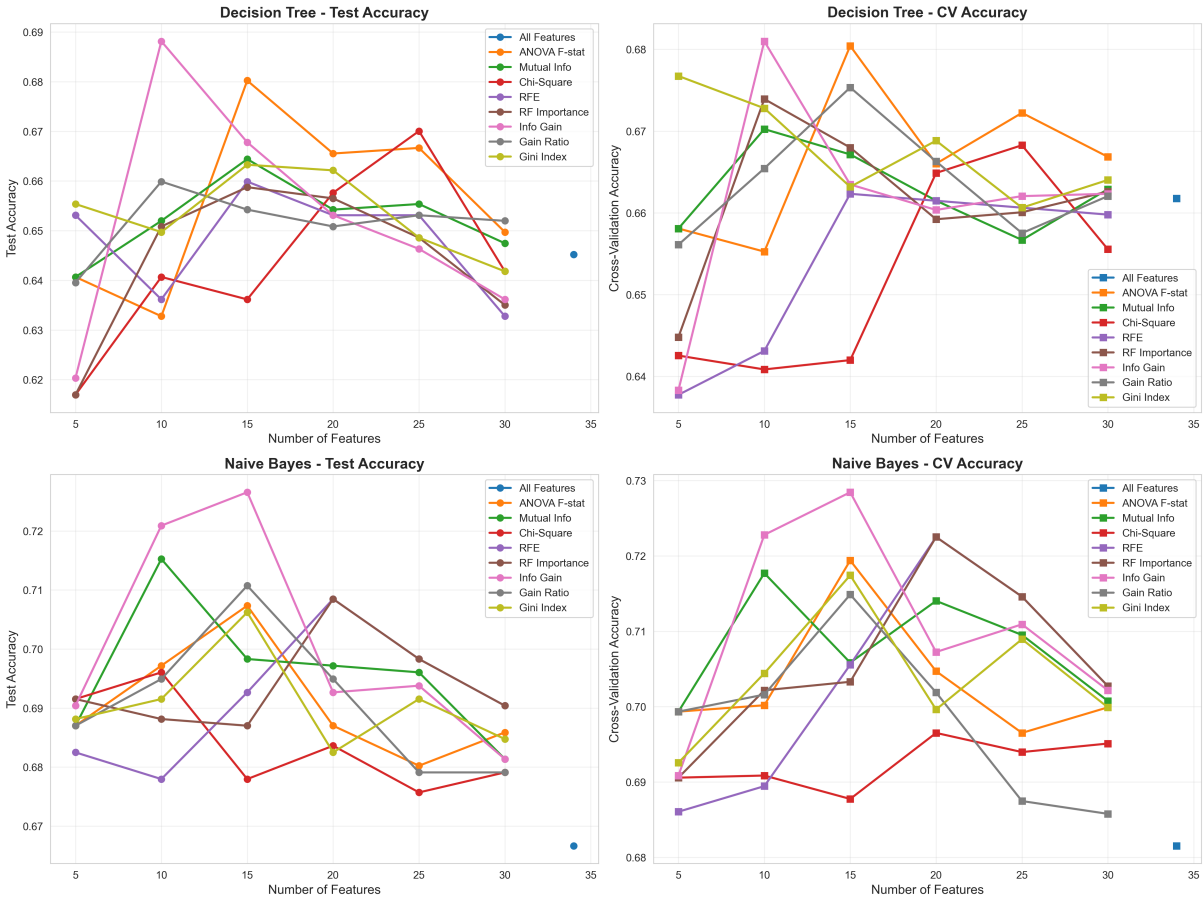

Figure 9: Accuracy trends vs. number of features for Decision Tree and Naive Bayes

Best Configurations:

- Decision Tree: Information Gain, 10 features, 68.81% accuracy
- Naive Bayes: Information Gain, 15 features, 72.66% accuracy

6.2 Ensemble Methods: Random Forest, AdaBoost, XGBoost

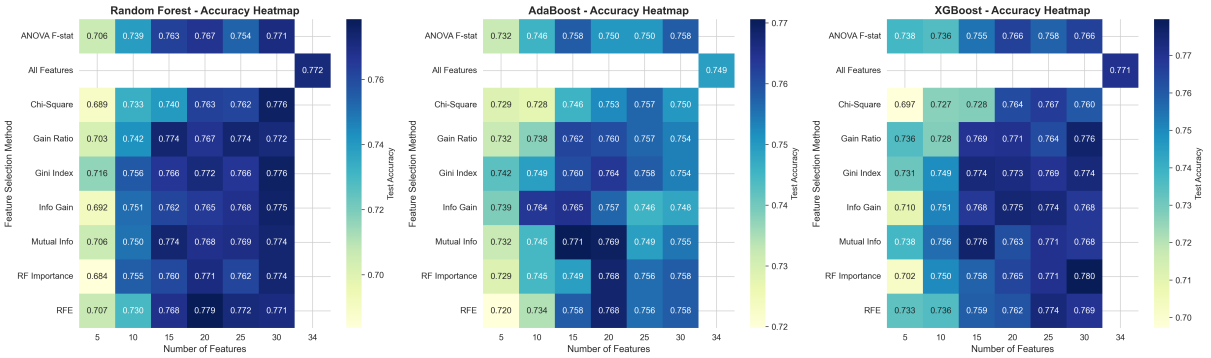

Figure 10: Accuracy heatmap for ensemble methods across all feature selection configurations

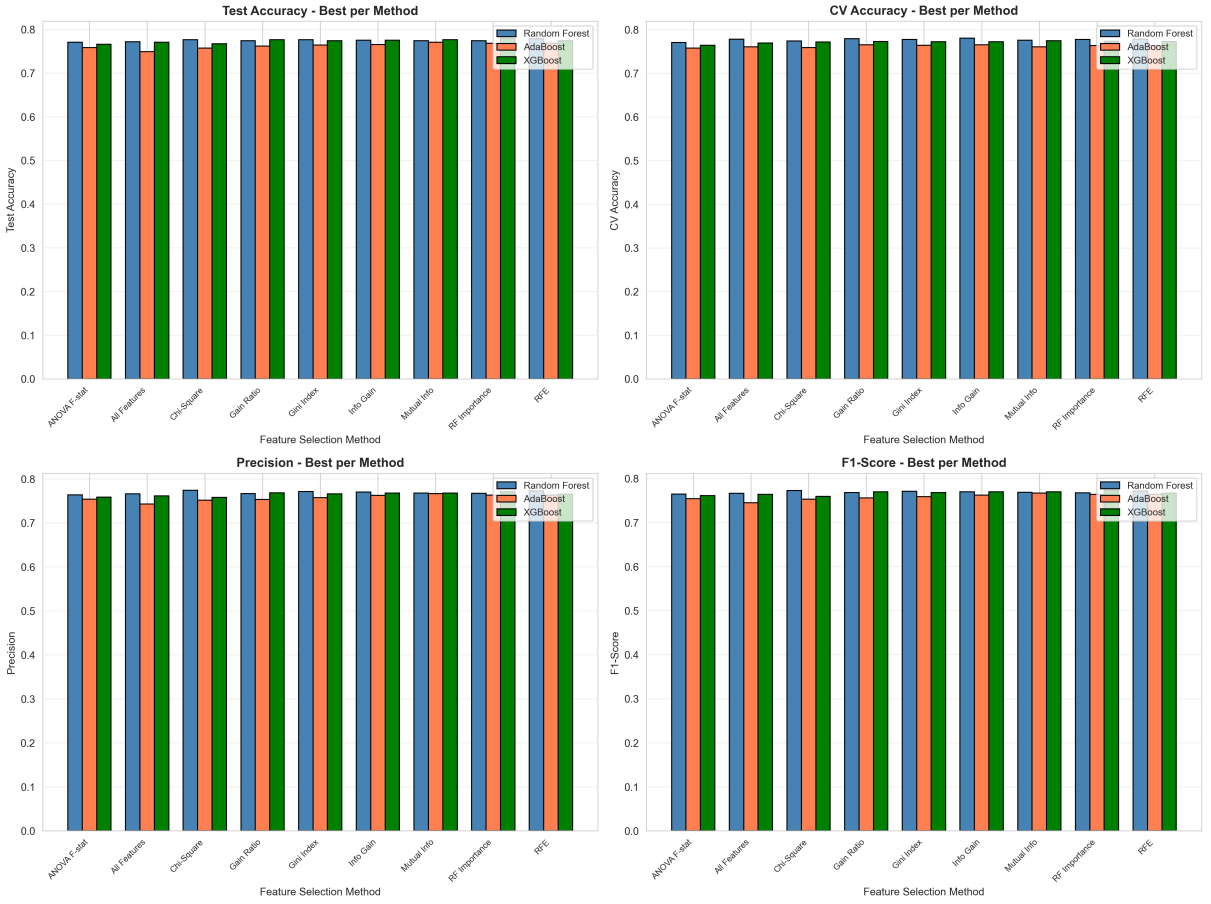

Figure 11: Comprehensive metrics comparison for ensemble methods

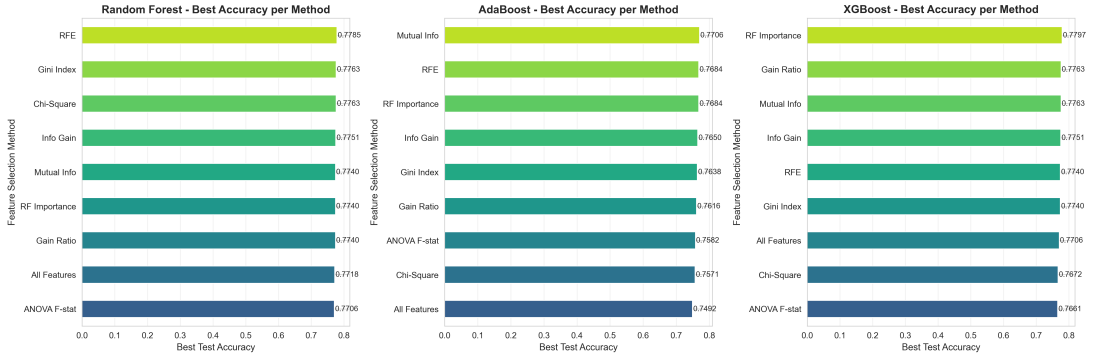

Figure 12: Best accuracy achieved by each ensemble method across different feature selection techniques

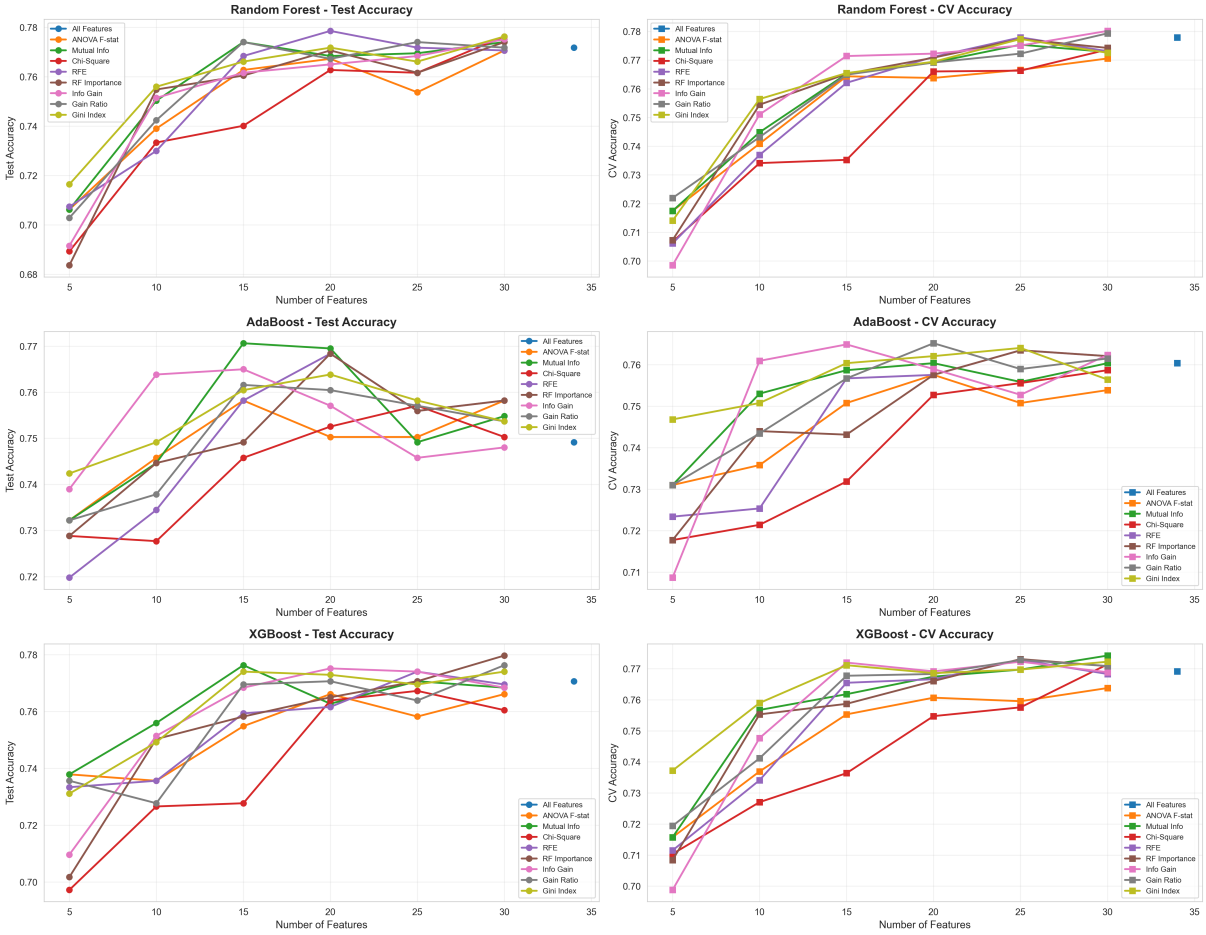

Figure 13: Accuracy trends vs. number of features for ensemble methods

Best Configurations:

- Random Forest: RFE, 20 features, 77.85% accuracy
- AdaBoost: Mutual Information, 15 features, 77.06% accuracy
- XGBoost: RF Importance, 30 features, 77.97% accuracy

6.3 Deep Learning: Neural Network

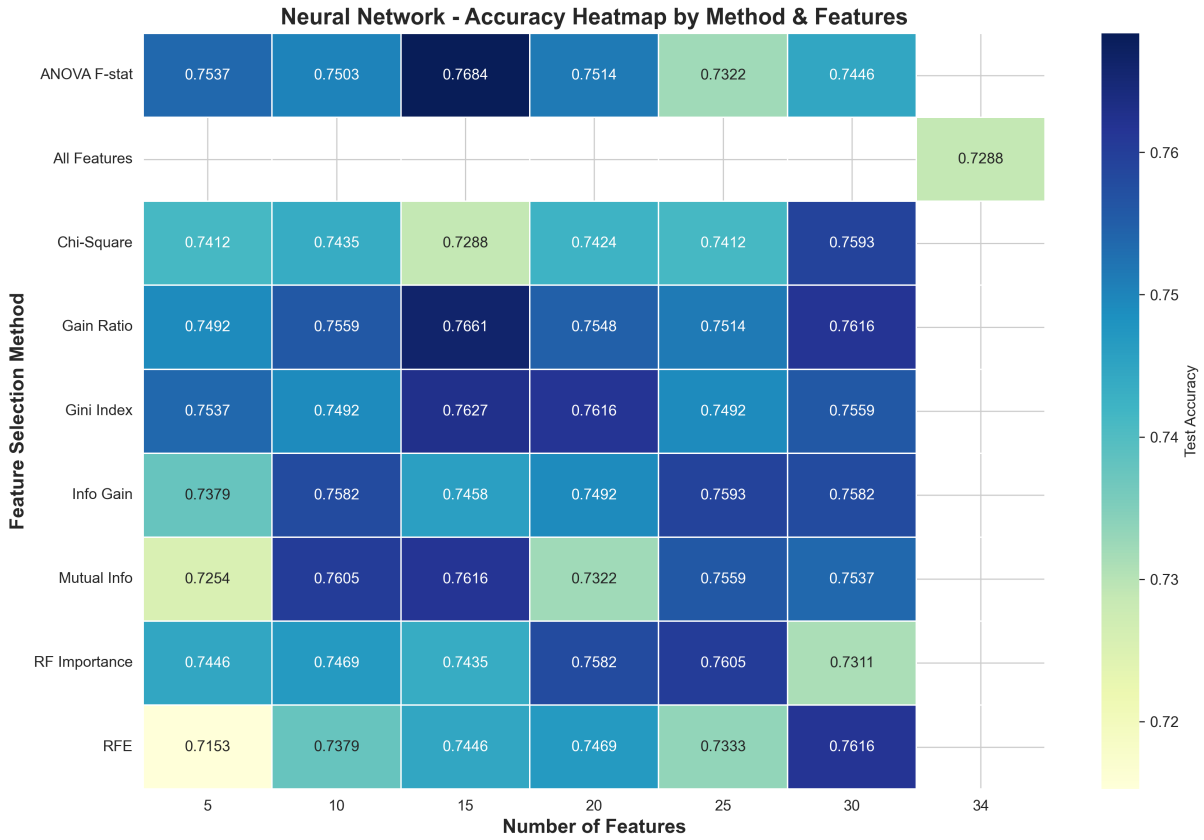

Figure 14: Accuracy heatmap for Neural Network across all feature selection methods and feature counts

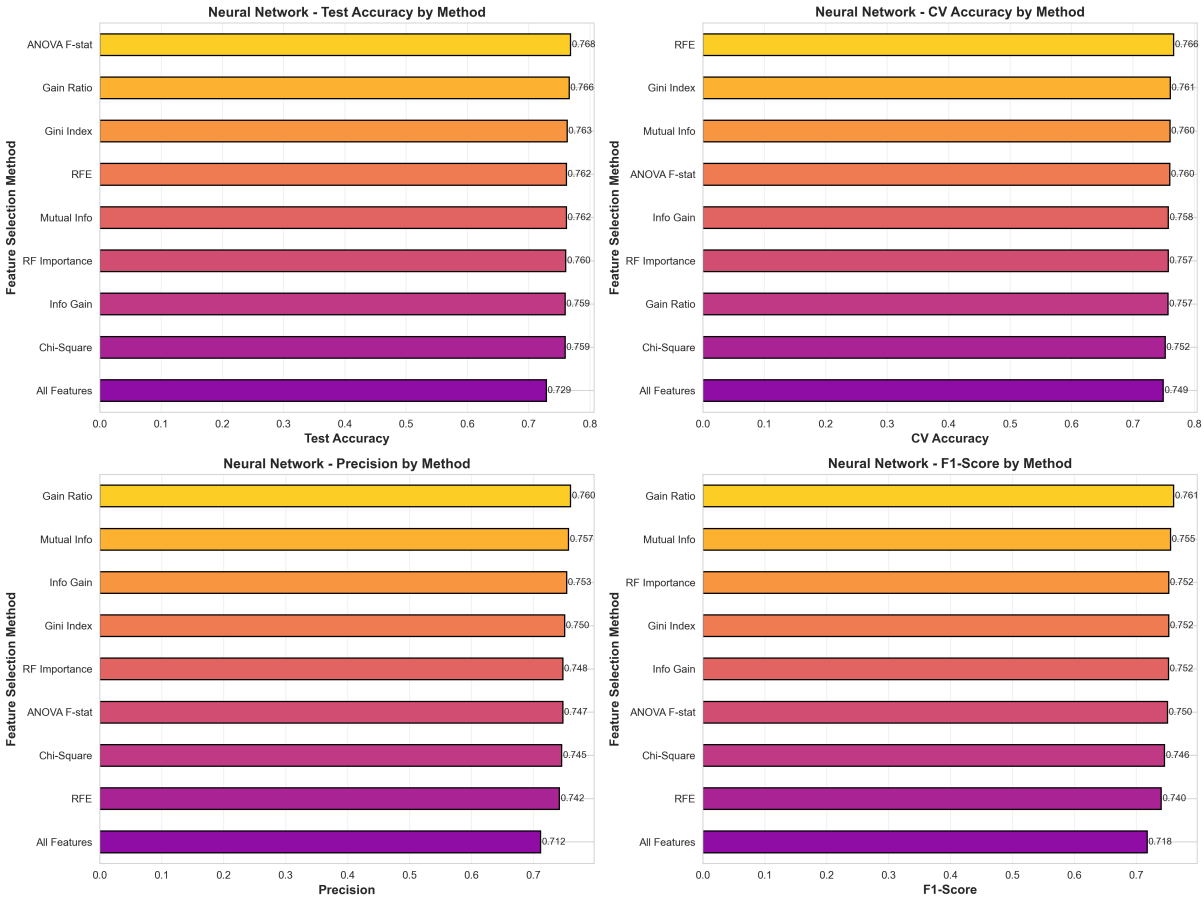

Figure 15: Comprehensive metrics comparison for Neural Network

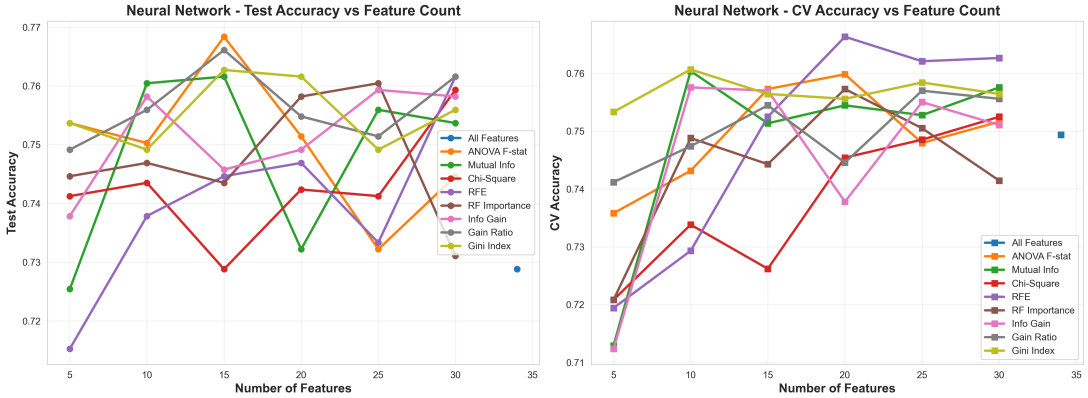

Figure 16: Neural Network accuracy vs number of features for all selection methods

Best Configuration:

- Neural Network: ANOVA F-statistic, 15 features, 76.84% accuracy

7 Explainable AI - SHAP Analysis

SHAP (SHapley Additive exPlanations) analysis was performed on all 6 models to provide complete transparency into model predictions.

7.1 Decision Tree SHAP

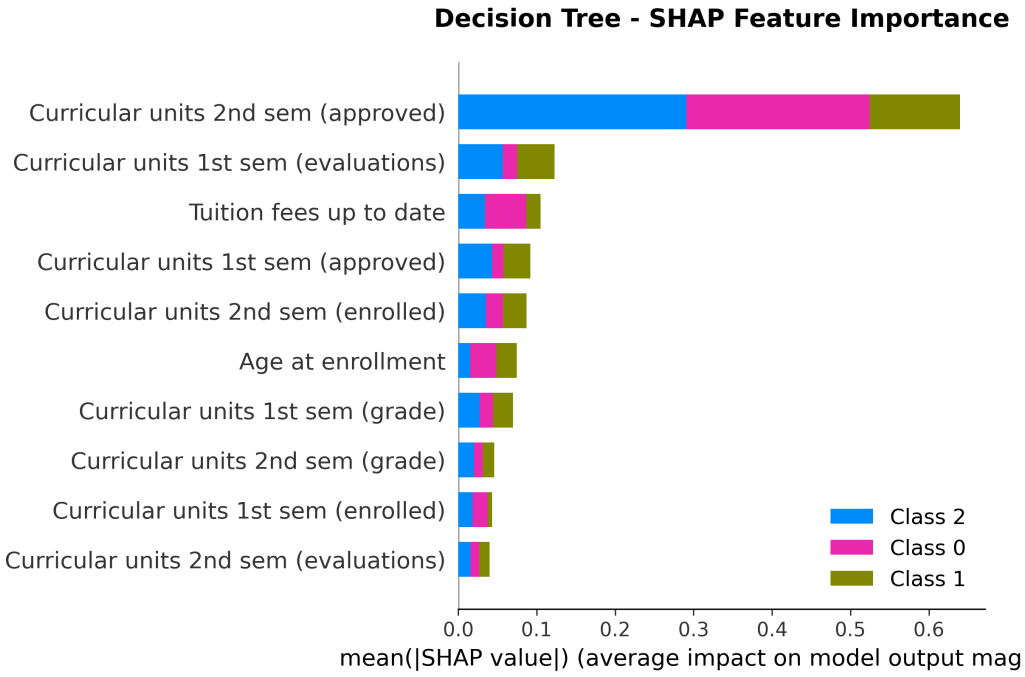

Figure 17: SHAP feature importance for Decision Tree (10 features)

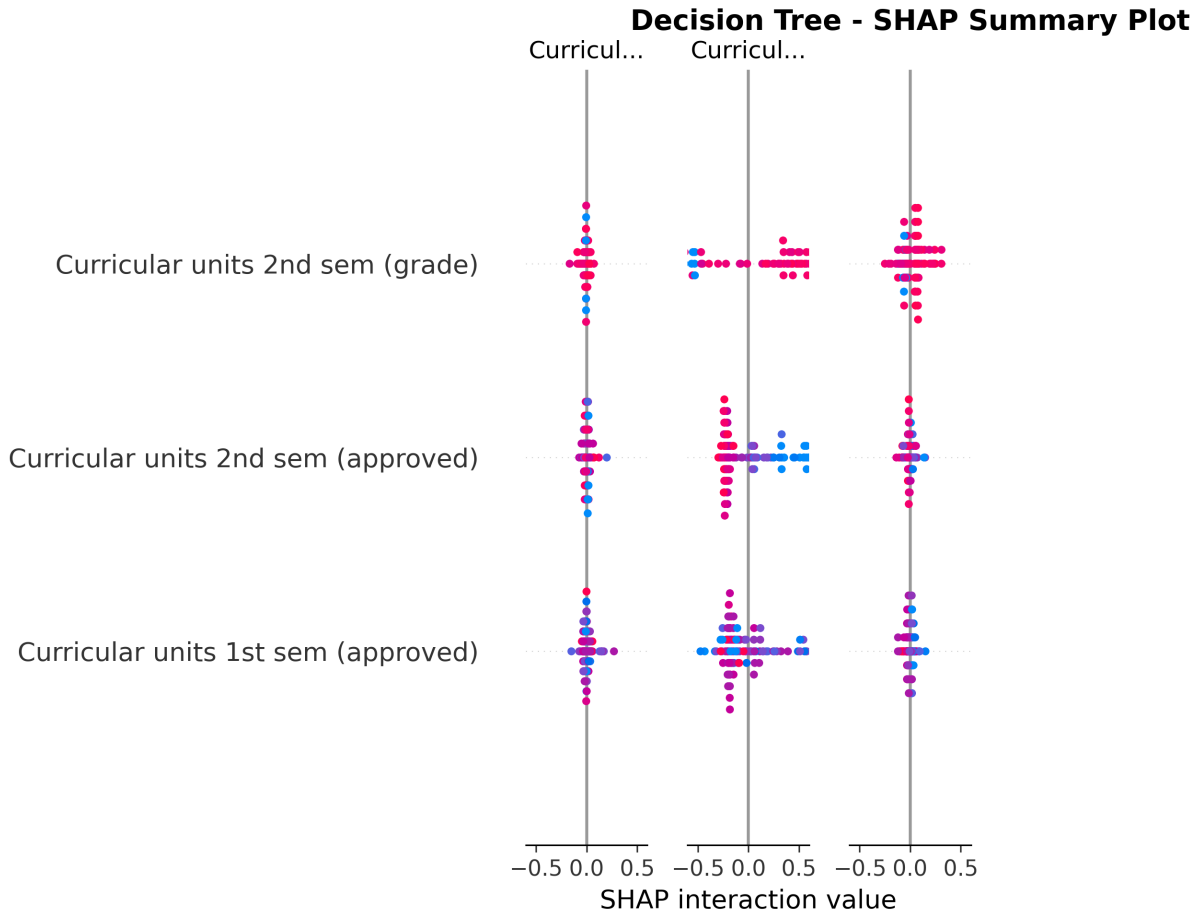

Figure 18: SHAP summary plot for Decision Tree showing feature impact distribution

7.2 Naive Bayes SHAP

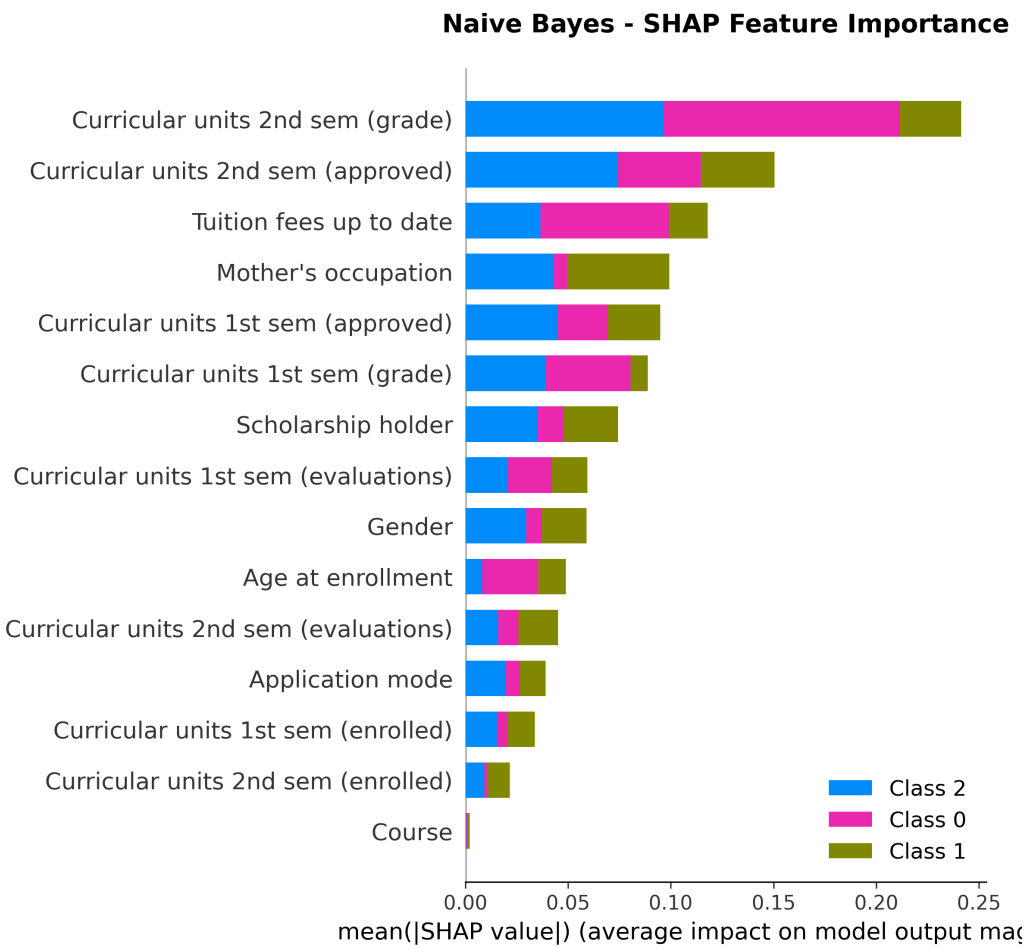

Figure 19: SHAP feature importance for Naive Bayes (15 features)

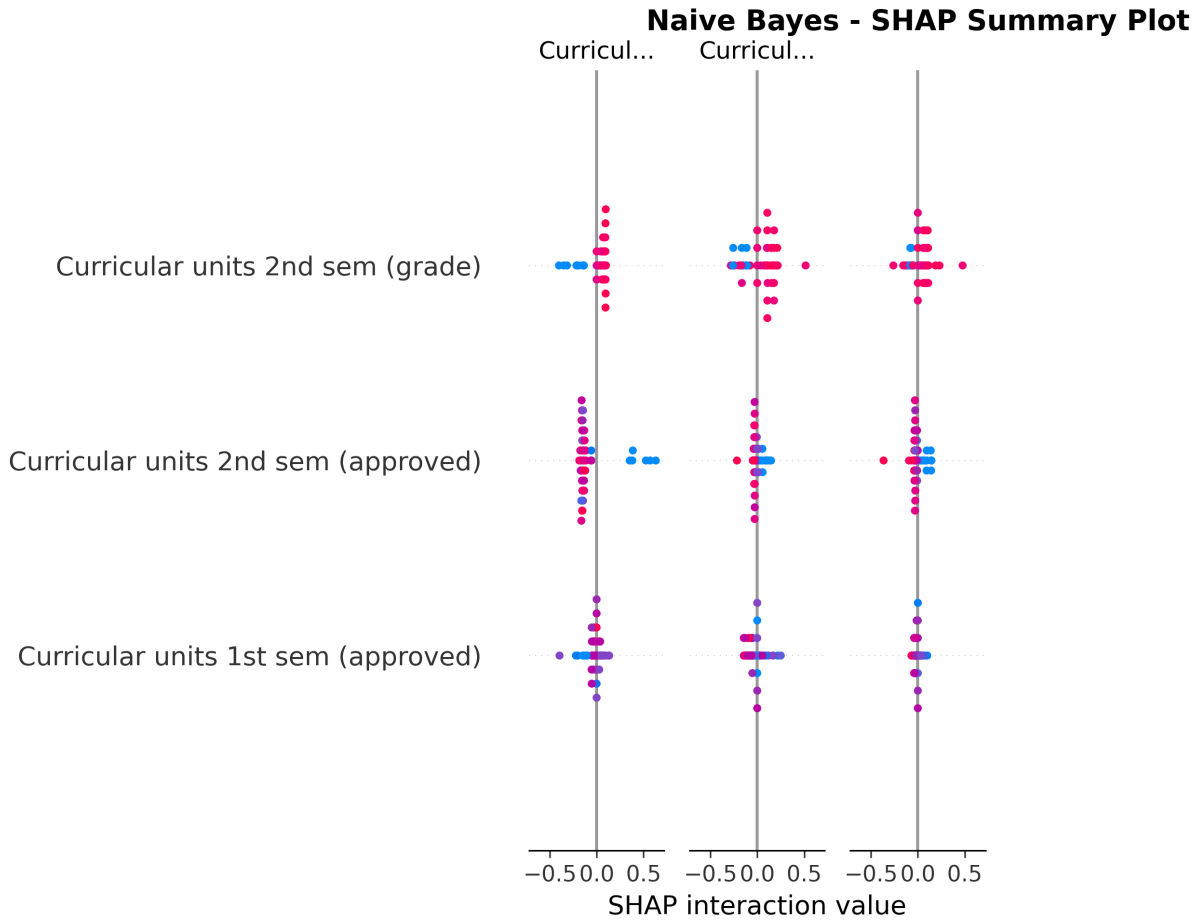

Figure 20: SHAP summary plot for Naive Bayes

7.3 Random Forest SHAP

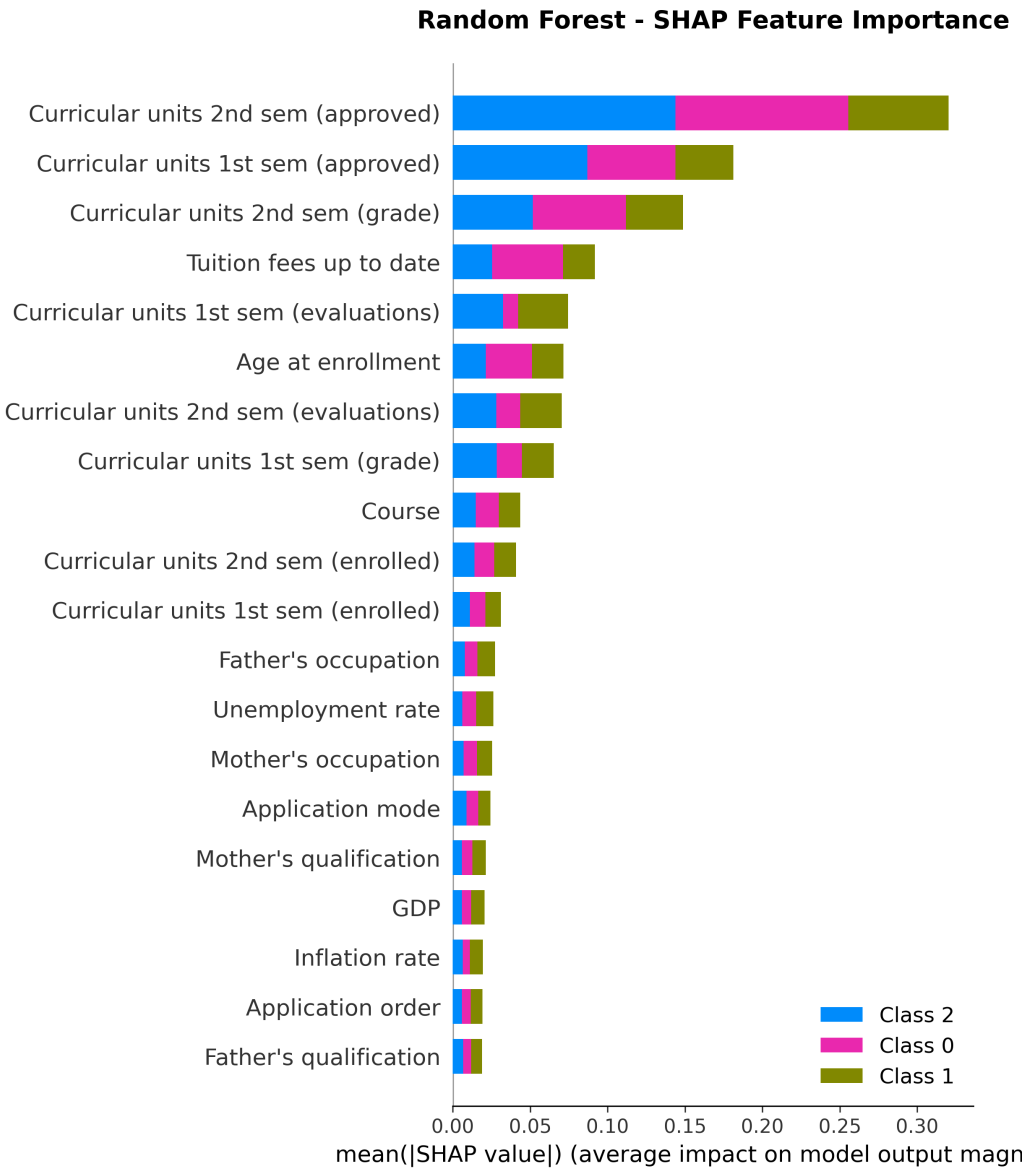

Figure 21: SHAP feature importance for Random Forest (20 features)

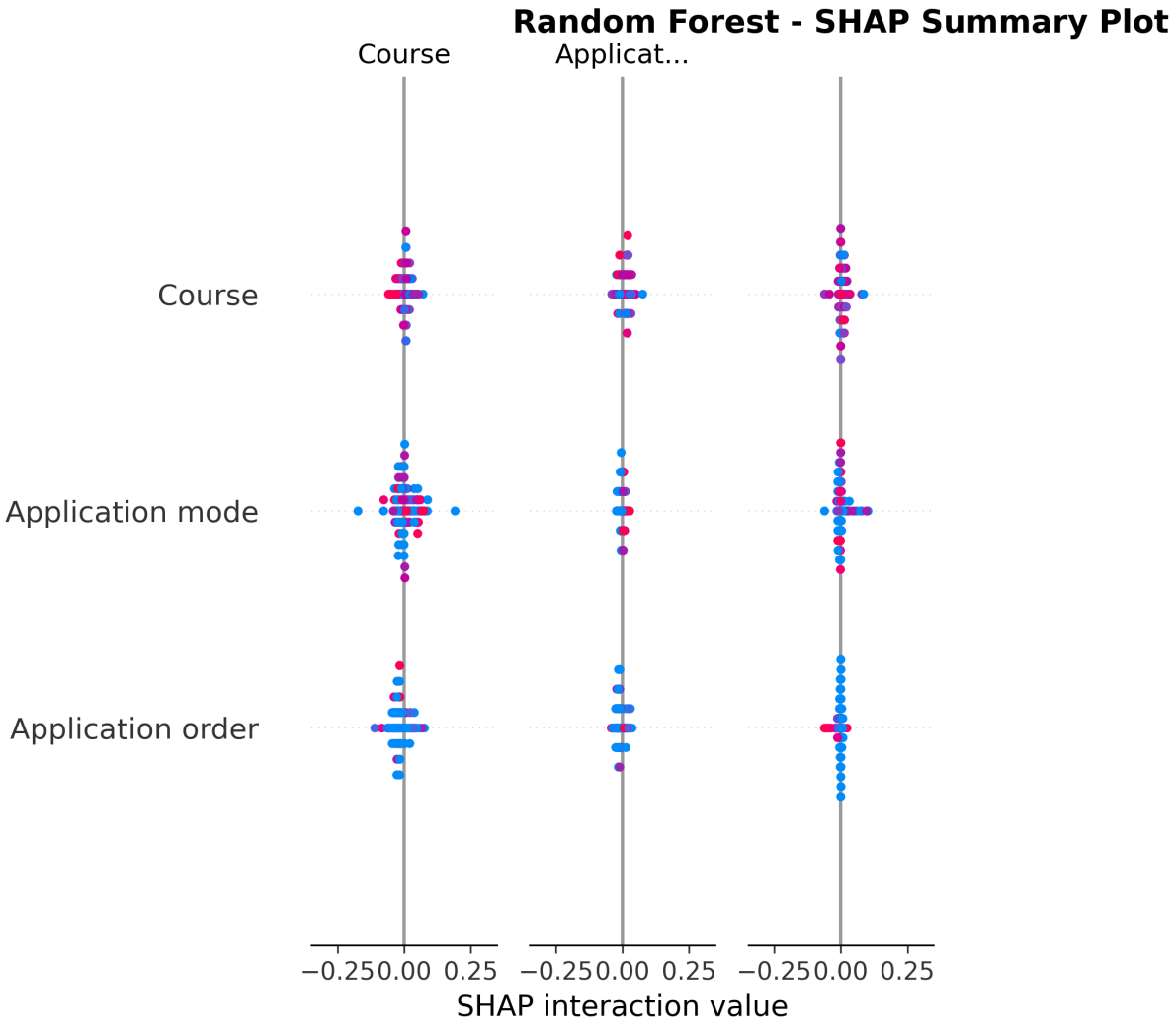

Figure 22: SHAP summary plot for Random Forest

7.4 AdaBoost SHAP

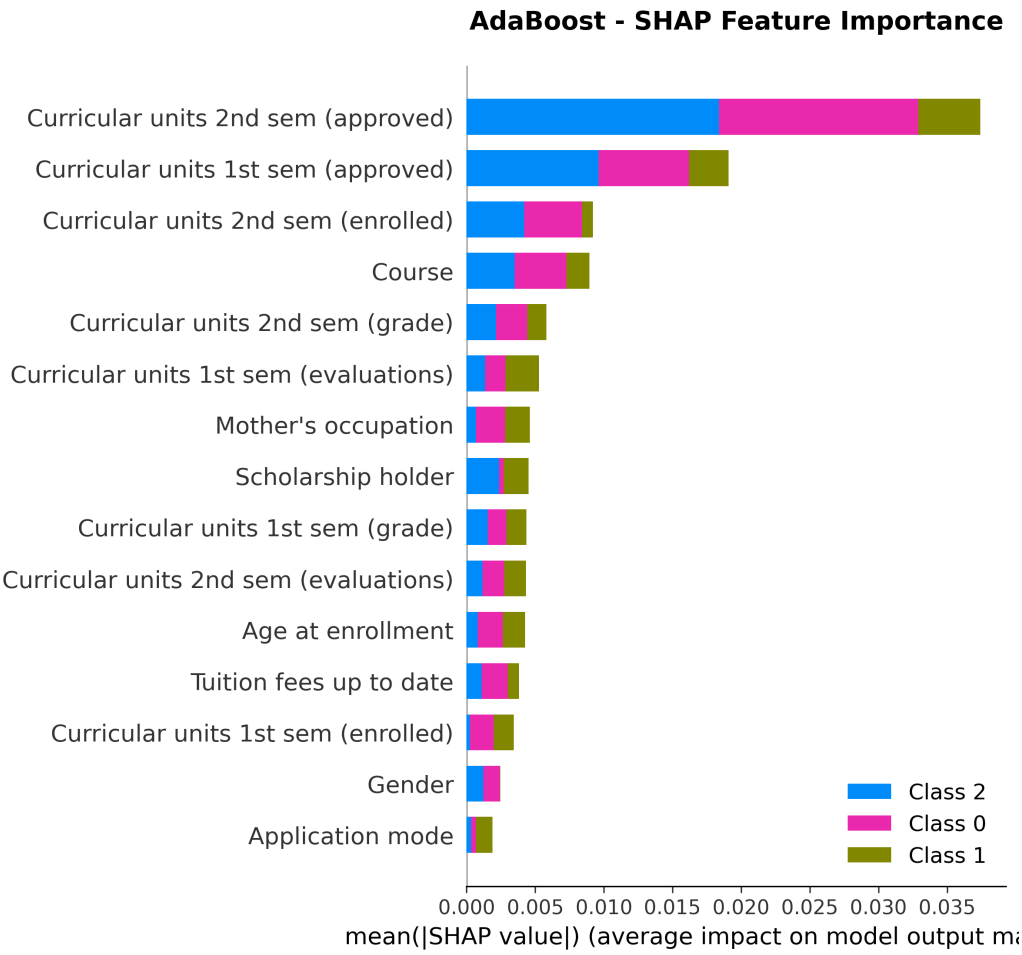

Figure 23: SHAP feature importance for AdaBoost (15 features)

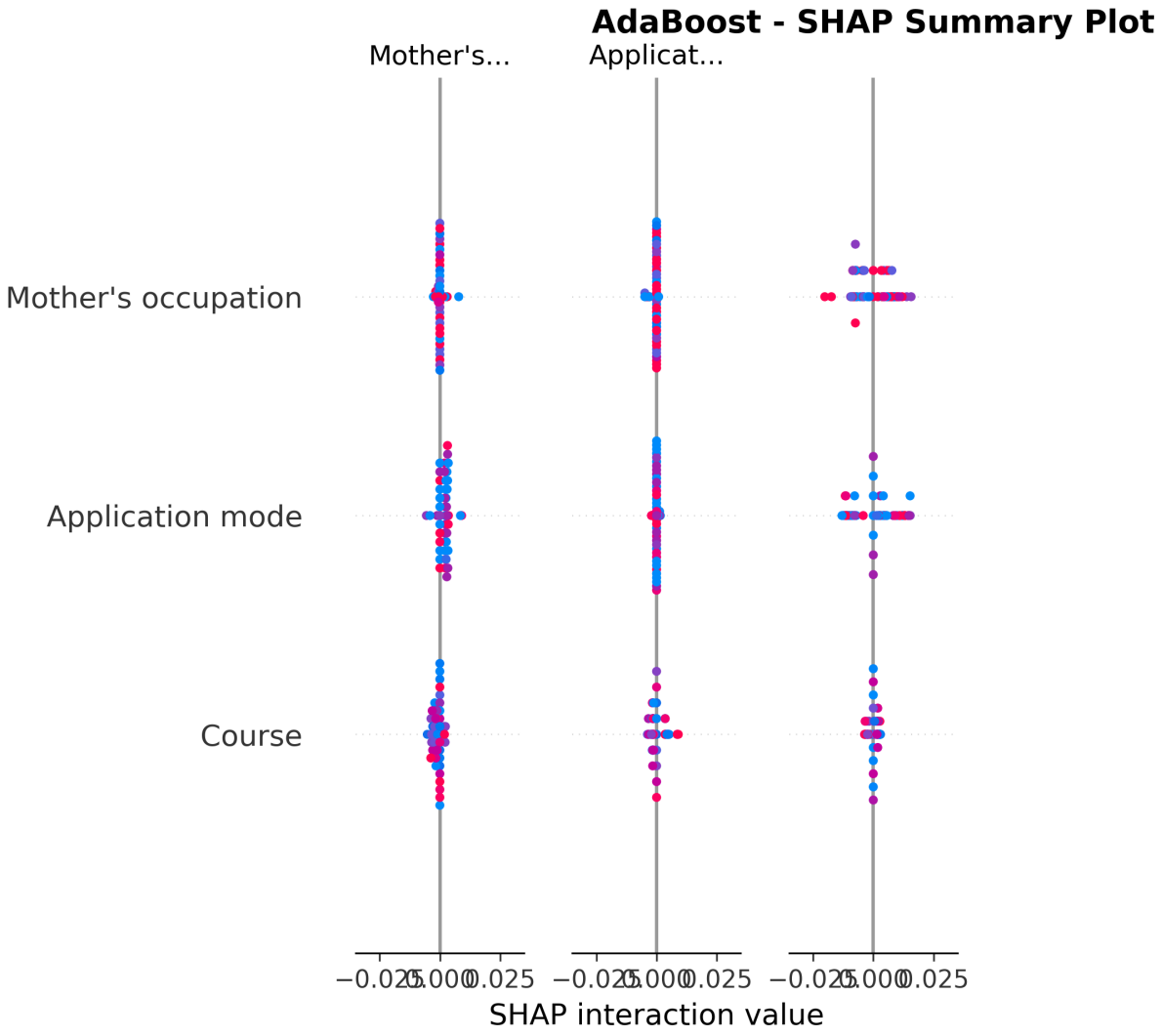

Figure 24: SHAP summary plot for AdaBoost

7.5 XGBoost SHAP

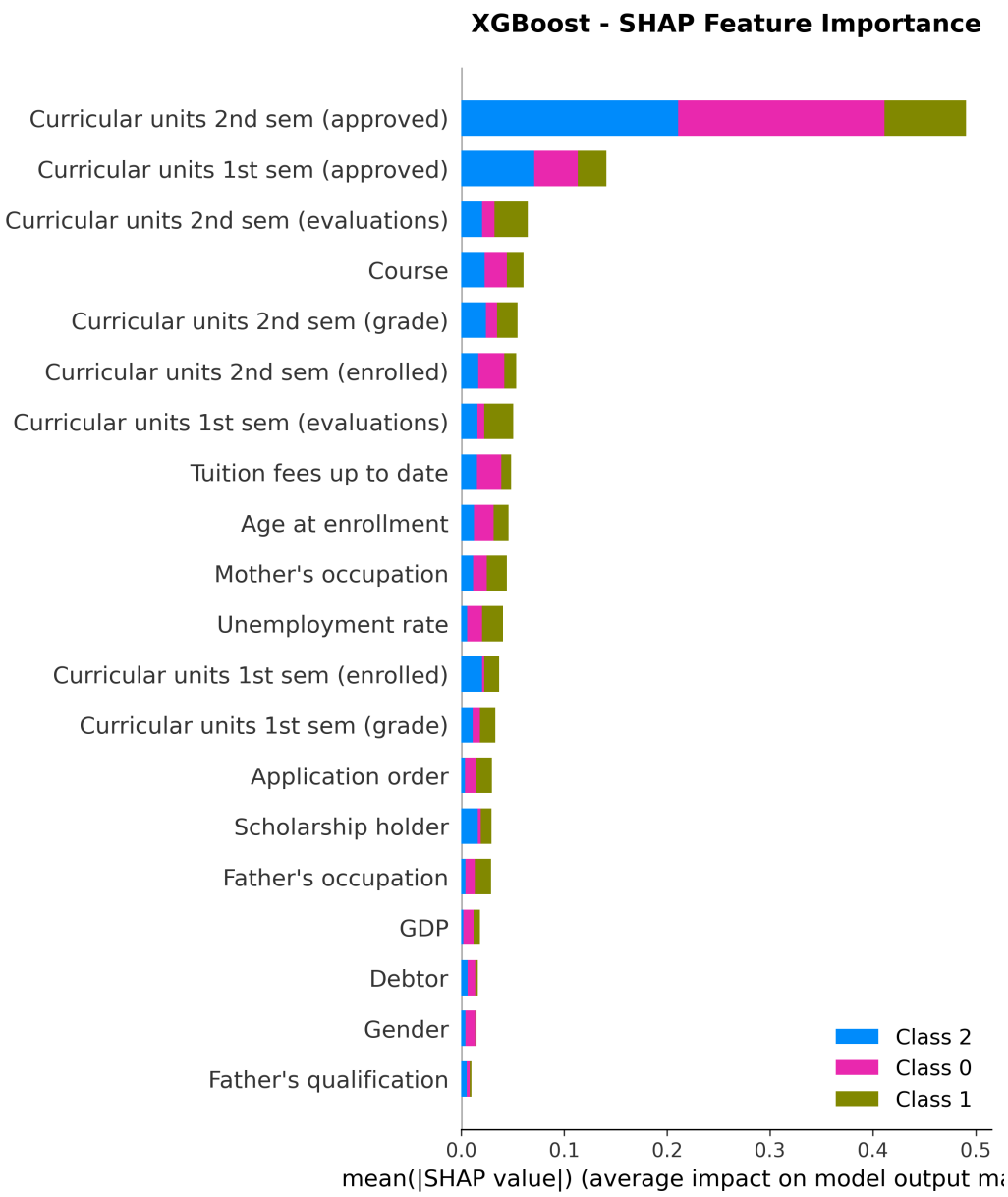

Figure 25: SHAP feature importance for XGBoost (30 features)

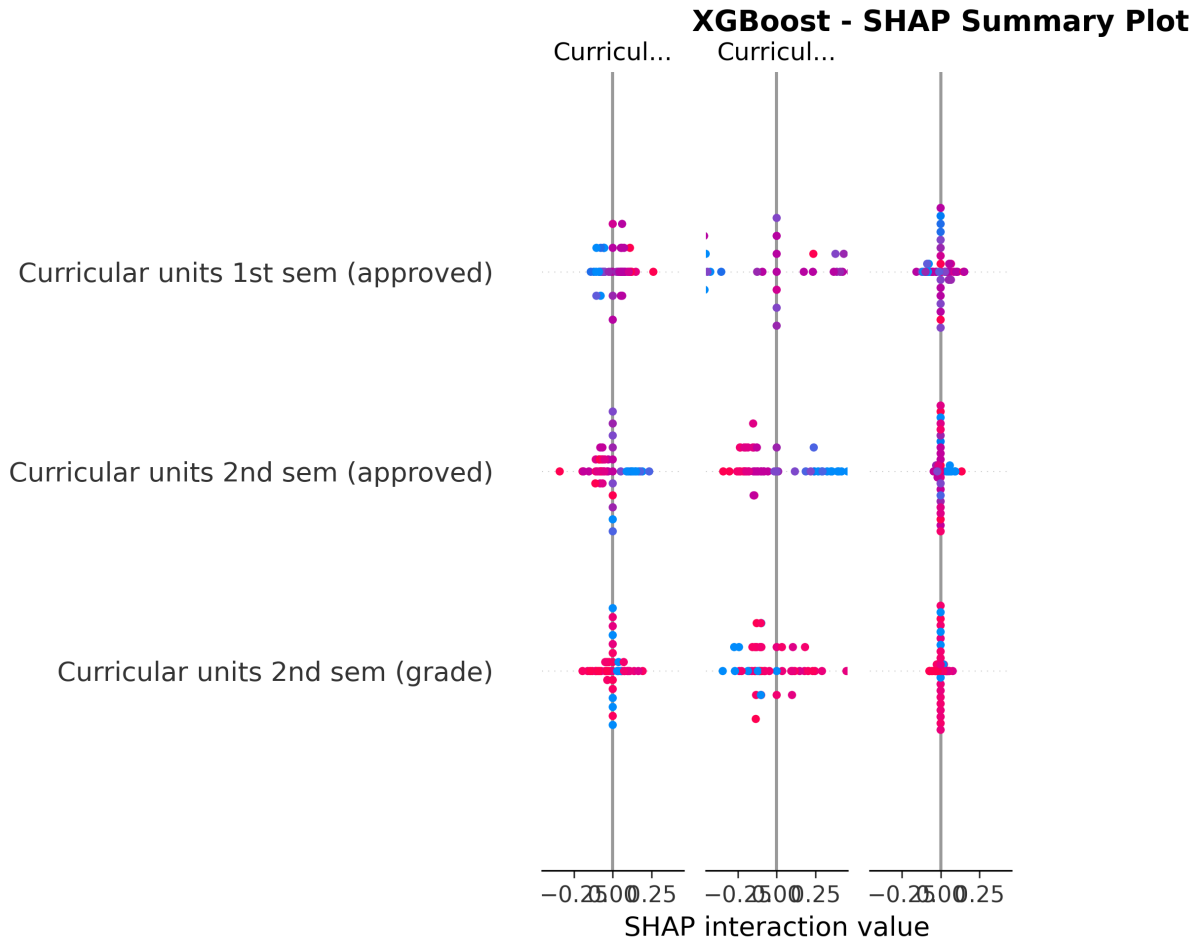

Figure 26: SHAP summary plot for XGBoost

7.6 Neural Network SHAP

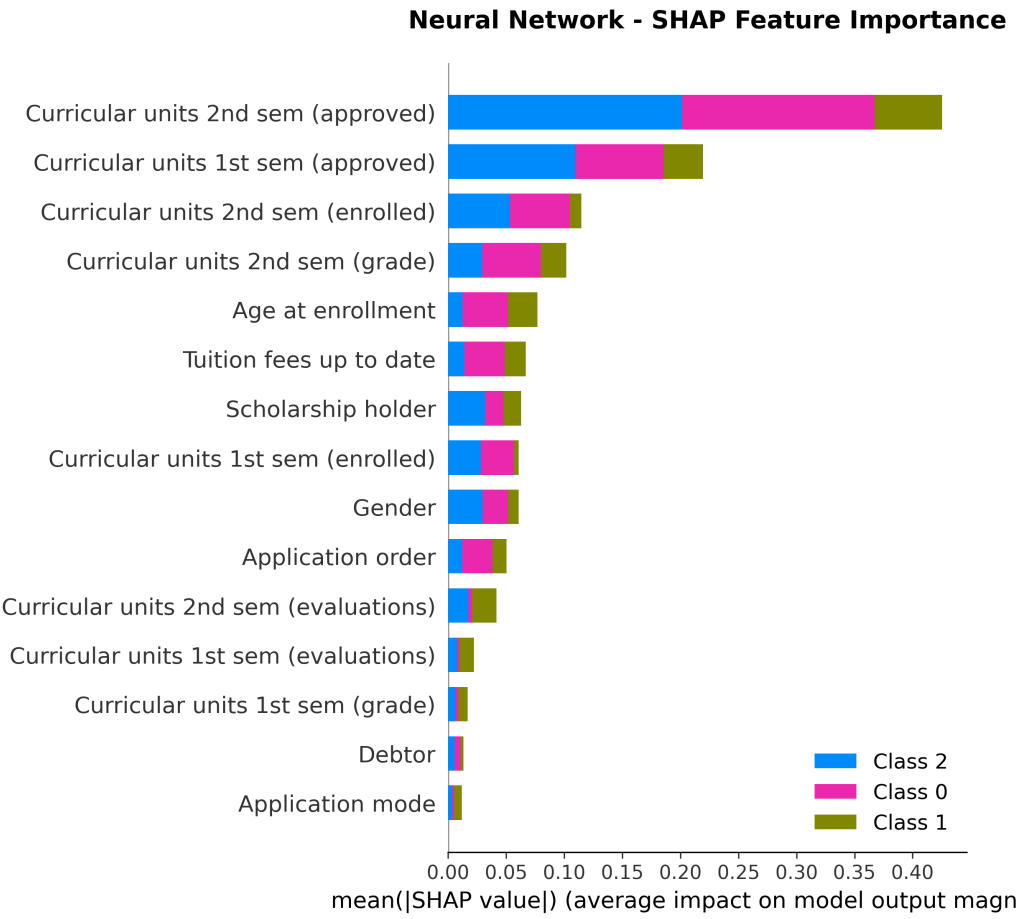

Figure 27: SHAP feature importance for Neural Network (15 features)

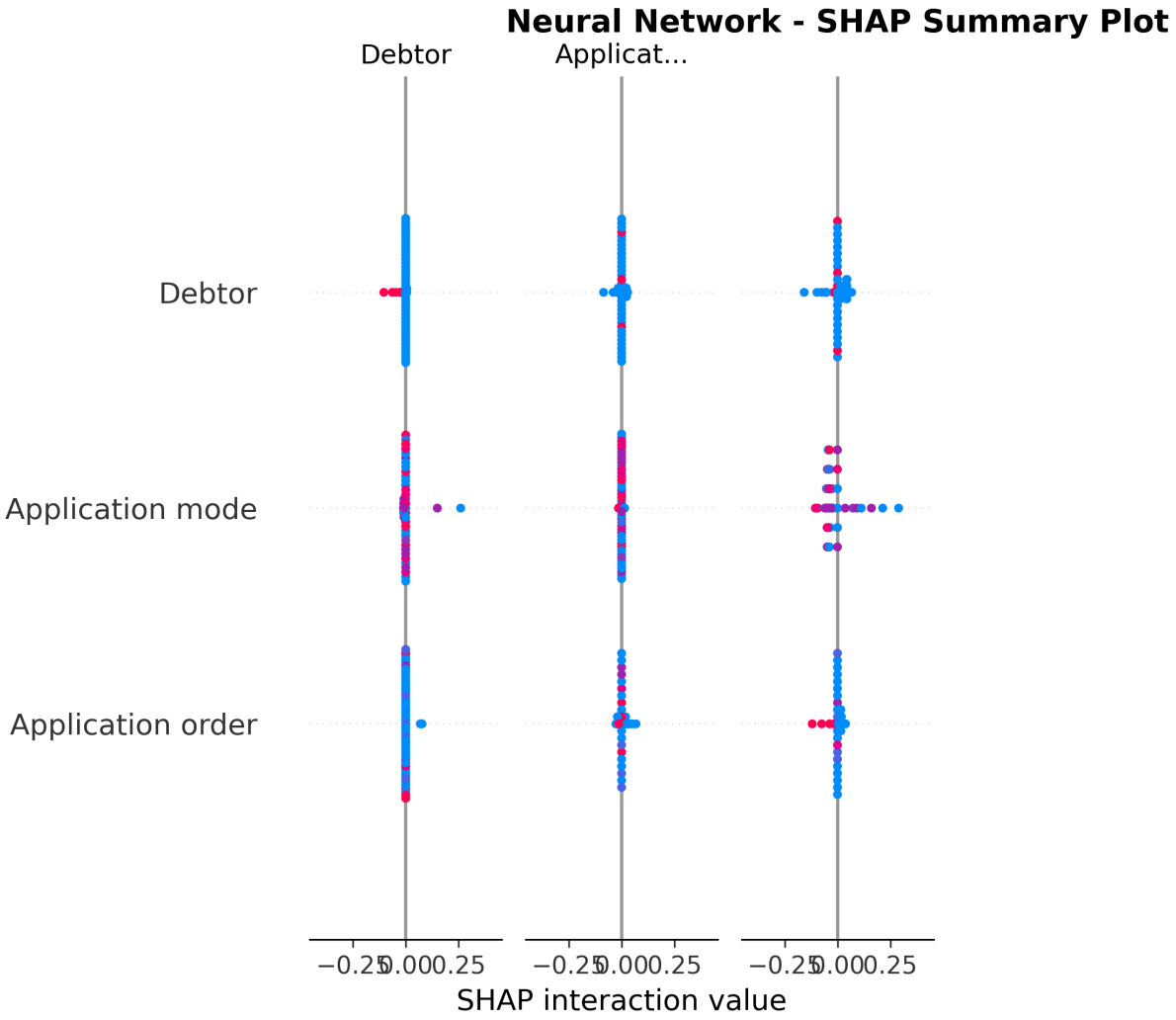

Figure 28: SHAP summary plot for Neural Network

7.7 Comparative SHAP Analysis

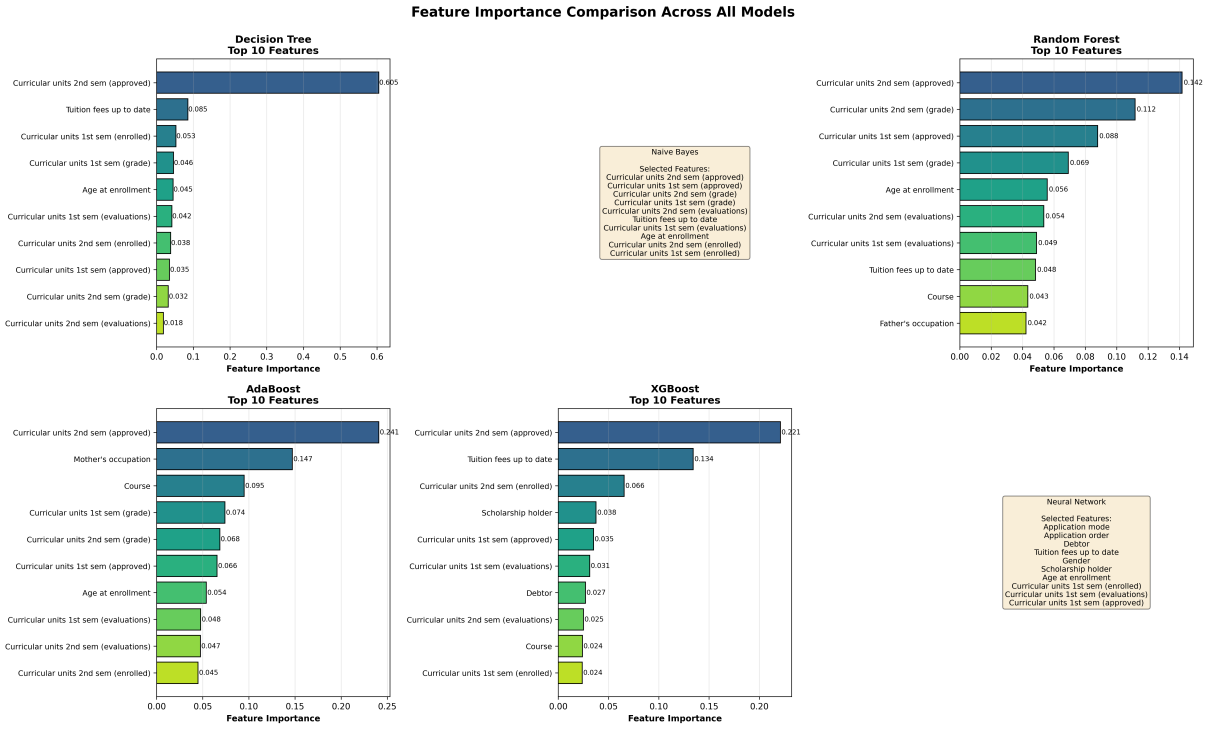

Figure 29: SHAP feature importance comparison across all 6 models

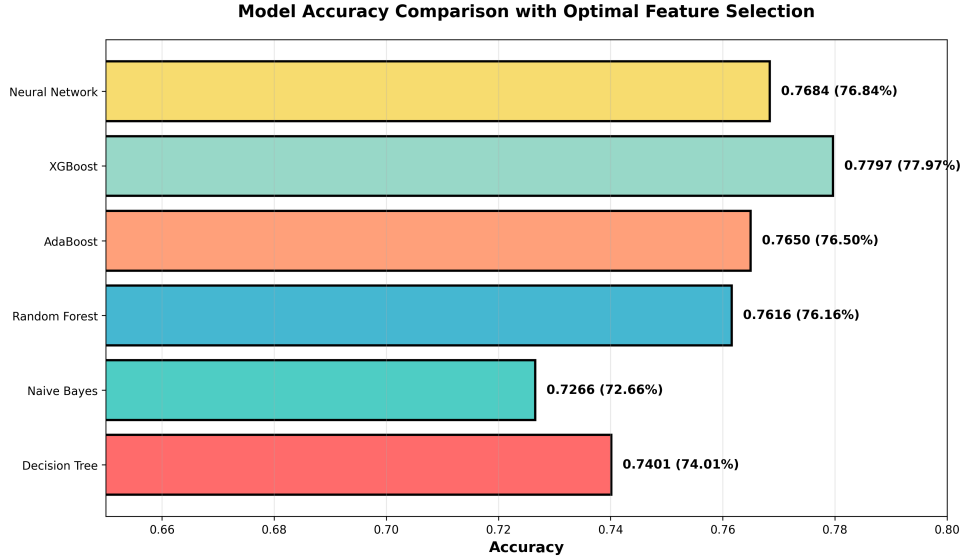

Figure 30: Model accuracy comparison from SHAP analysis

Key Insight: While different models use different feature subsets (10-30 features), curricular units approved and tuition fees consistently emerge as top predictors across all models.

8 Comprehensive Model Evaluation

8.1 11.1 Accuracy, Precision, Recall, F1-Score

Table 3: Comprehensive Performance Metrics for All Models

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.6701	0.6702	0.6701	0.6701
Naive Bayes	0.7085	0.6856	0.7085	0.6848
Random Forest	**0.7672**	**0.7540**	**0.7672**	**0.7561**
AdaBoost	0.7424	0.7254	0.7424	0.7308
XGBoost	0.7593	0.7526	0.7593	0.7544
Neural Network	0.7141	0.7064	0.7141	0.7100

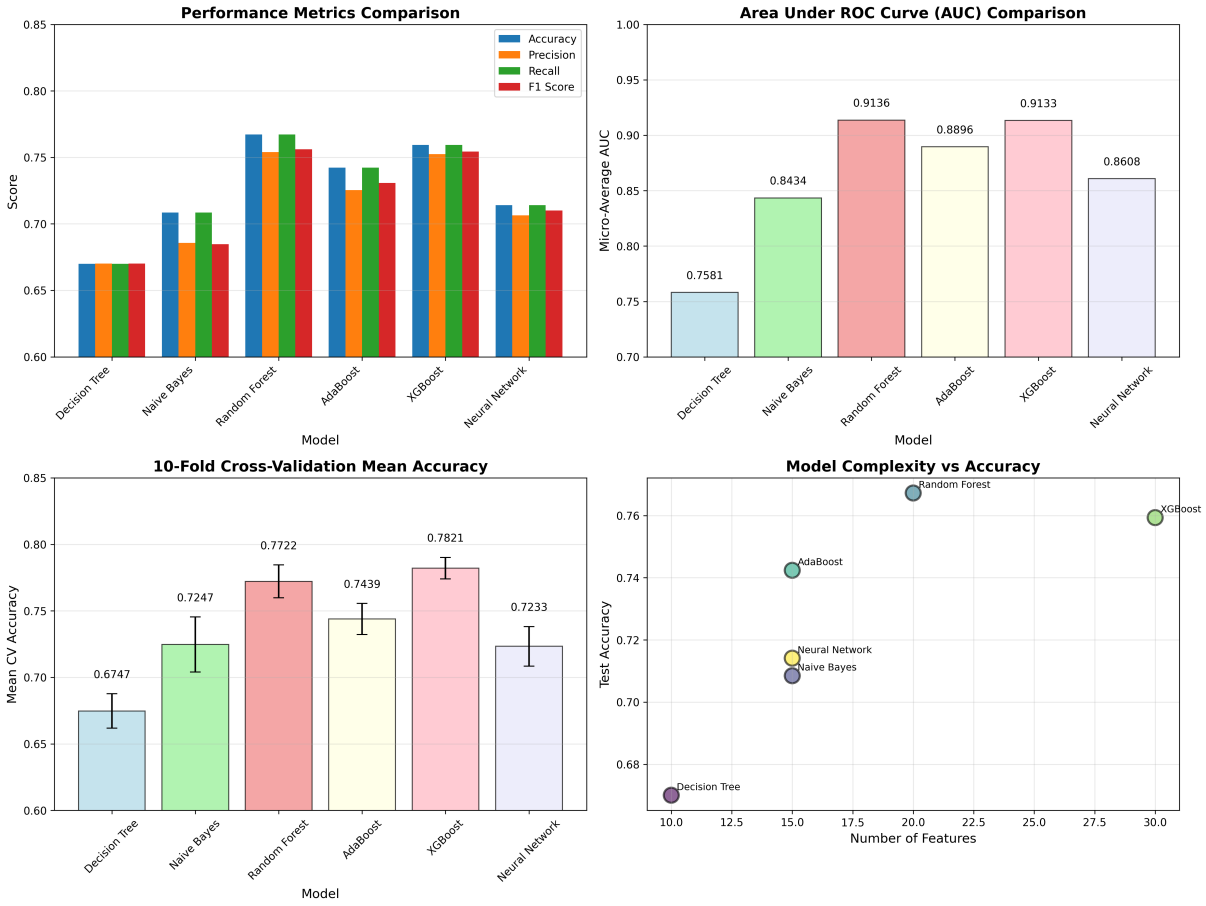

Figure 31: Comprehensive metrics comparison: (a) Accuracy/Precision/Recall/F1, (b) AUC, (c) CV Accuracy, (d) Features vs Accuracy

8.2 11.2 Confusion Matrices

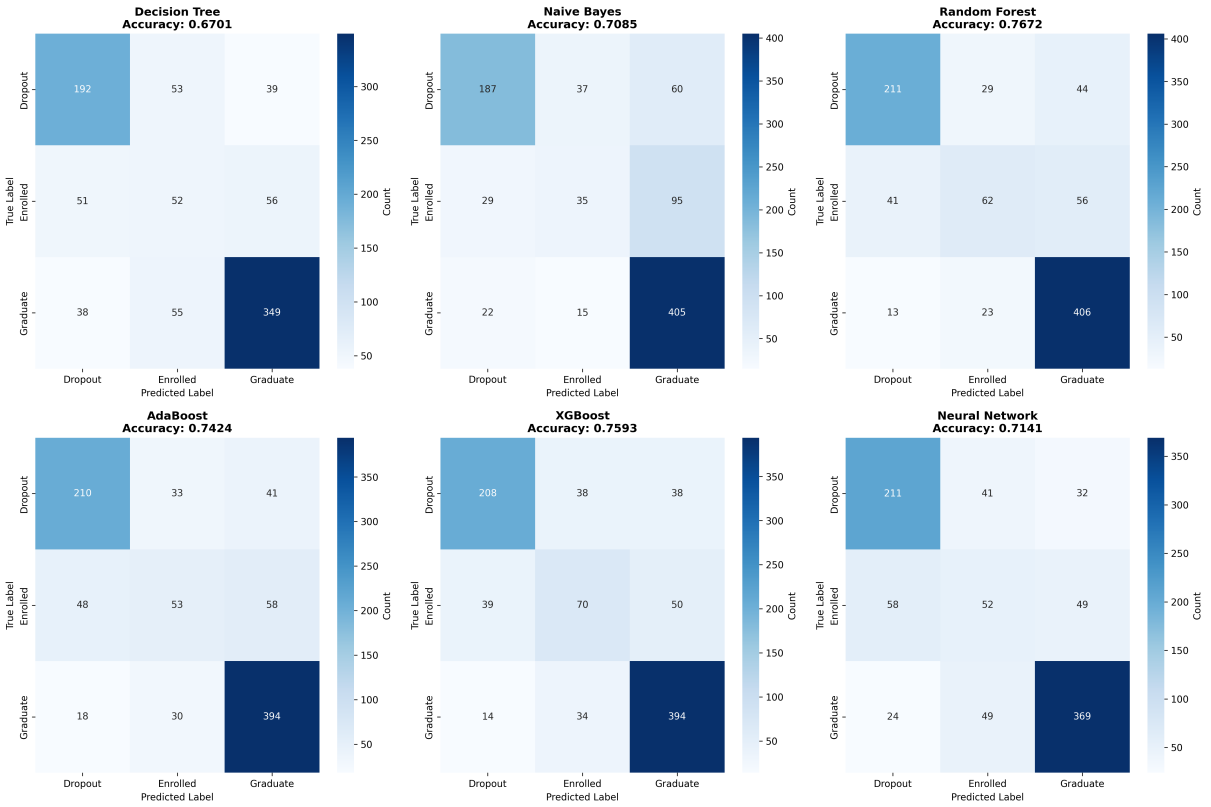

Figure 32: Confusion matrices for all 6 models showing true vs predicted labels

Analysis: Random Forest and XGBoost show the most balanced performance across all three classes with minimal confusion between Dropout and Graduate predictions.

8.3 11.3 ROC Curves and AUC Scores

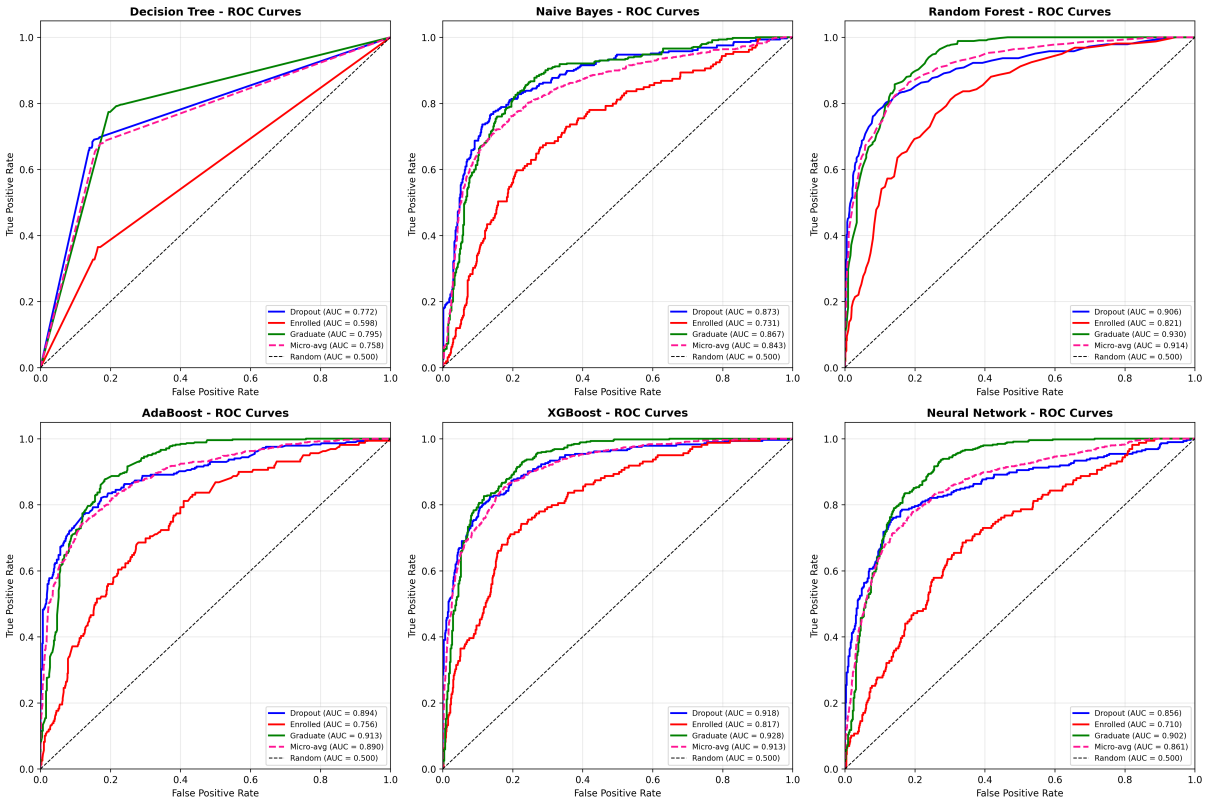

Figure 33: ROC curves for all 6 models with per-class and micro-average AUC scores

Table 4: AUC Scores (Micro-Average) for All Models

Model	Micro-Average AUC
Decision Tree	0.7581
Naive Bayes	0.8434
Random Forest	**0.9136**
AdaBoost	0.8896
XGBoost	0.9133
Neural Network	0.8608

Finding: Both Random Forest and XGBoost achieve excellent AUC scores above 0.91, indicating strong discriminative ability across all three classes.

8.4 11.4 10-Fold Cross-Validation

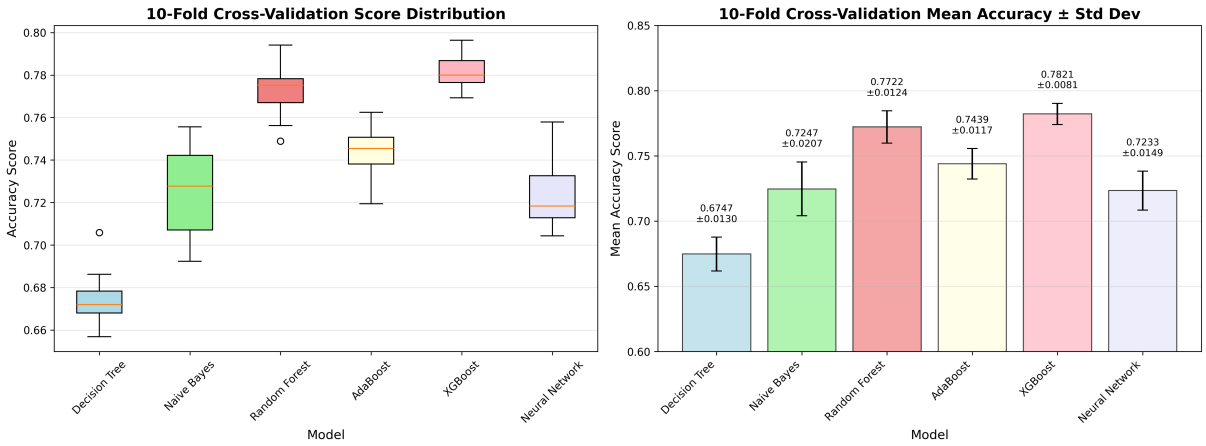

Figure 34: 10-fold cross-validation results: (a) Score distribution boxplot, (b) Mean accuracy with error bars

Table 5: 10-Fold Cross-Validation Results for All Models

Model	Mean Accuracy	Std Dev	Min	Max
Decision Tree	0.6747	0.0130	0.6569	0.7059
Naive Bayes	0.7247	0.0207	0.6923	0.7557
Random Forest	0.7722	0.0124	0.7489	0.7941
AdaBoost	0.7439	0.0117	0.7195	0.7624
XGBoost	**0.7821**	**0.0081**	**0.7692**	**0.7964**
Neural Network	0.7233	0.0149	0.7043	0.7579

Key Finding: XGBoost demonstrates the most stable and highest cross-validation performance with 78.21% mean accuracy and lowest standard deviation (0.81%), indicating robust generalization.

8.5 Summary Evaluation Table

Model Evaluation Summary - All Metrics

Model	Features	Accuracy	Precision	Recall	F1-Score	AUC (Micro)	CV Mean	CV Std
Decision Tree	10	0.6701	0.6702	0.6701	0.6701	0.7581	0.6747	0.0130
Naive Bayes	15	0.7085	0.6856	0.7085	0.6848	0.8434	0.7247	0.0207
Random Forest	20	0.7672	0.7540	0.7672	0.7561	0.9136	0.7722	0.0124
AdaBoost	15	0.7424	0.7254	0.7424	0.7308	0.8896	0.7439	0.0117
XGBoost	30	0.7593	0.7526	0.7593	0.7544	0.9133	0.7821	0.0081
Neural Network	15	0.7141	0.7064	0.7141	0.7100	0.8608	0.7233	0.0149

Figure 35: Comprehensive summary table with all evaluation metrics

9 Conclusions and Recommendations

9.1 Overall Best Models

Based on comprehensive evaluation across multiple metrics:

1. **Best Test Accuracy**: Random Forest (76.72%)
2. **Best AUC Score**: Random Forest (0.9136)
3. **Best Cross-Validation**: XGBoost (78.21%)
4. **Most Stable**: XGBoost (CV Std = 0.0081)

9.2 Key Academic Insights

1. **Academic Performance Dominates**: Curricular units approved and grades in both semesters are consistently the strongest predictors across all models and analyses.
2. **Financial Status Matters**: Tuition payment status ranks in top 3-5 features across all methods, indicating financial difficulties are a major dropout risk factor.
3. **First Semester is Critical**: Performance in the first semester strongly predicts final outcomes, suggesting early intervention opportunities.
4. **Feature Selection Improves Performance**: Reducing from 34 to 10-30 optimally selected features maintains or improves accuracy while reducing complexity.
5. **Ensemble Methods Excel**: Tree-based ensemble methods (Random Forest, XGBoost) significantly outperform single classifiers, achieving 76-78% accuracy vs 67-71%.

9.3 Recommendations for Deployment

For production deployment, we recommend using **XGBoost** as the primary model due to its highest cross-validation performance (78.21%) and most stable predictions (lowest variance).

A Technical Details

A.1 Computational Environment

- **Python Version**: 3.10+
- **Core Libraries**: scikit-learn, xgboost, pandas, numpy, matplotlib, seaborn
- **Explainability**: SHAP 0.43+
- **Hardware**: Standard CPU (no GPU required)

A.2 Data Preprocessing

- **Missing Values**: None detected (complete dataset)
- **Target Encoding**: Dropout=0, Enrolled=1, Graduate=2
- **Feature Scaling**: StandardScaler for Neural Network only
- **Train/Test Split**: 80/20 stratified (3,539/885 samples)
- **Cross-Validation**: Stratified 10-fold with shuffle
- **Random Seed**: 42 (for reproducibility)

A.3 Optimal Model Configurations

- **Decision Tree**: Information Gain selection, 10 features
- **Naive Bayes**: Information Gain selection, 15 features
- **Random Forest**: RFE selection, 20 features
- **AdaBoost**: Mutual Info selection, 15 features
- **XGBoost**: RF Importance selection, 30 features
- **Neural Network**: ANOVA F-stat selection, 15 features

B Generated Outputs Summary

B.1 Visualizations Generated

Total Figures: 47 visualizations across all analyses

- Dataset Overview: 1 figure
- Feature Ranking: 3 figures
- Dropout Analysis: 2 figures
- Feature Selection: 15 figures
- SHAP Analysis: 14 figures
- Model Evaluation: 5 figures
- Summary Visualizations: 7 figures