Student Dropout Prediction Using Machine Learning

AI Model for Predicting Student Dropout Risk using Academic and Behavioral Insights.

Presented by:

Sakshi Gaikwad – B72

Sakshi Baviskar – B84

Krushna Taur – B86

Business Understanding

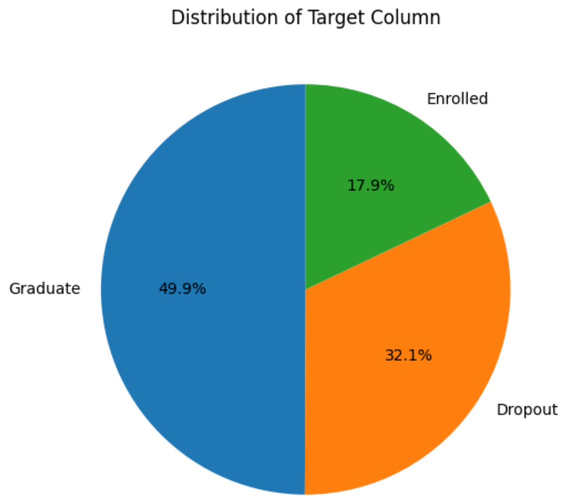
- Many colleges struggle to identify students at risk of dropout.
- Causes: poor academic performance, low engagement, financial issues.
- Early prediction helps institutes: Provide timely support, improve retention and student success.
- Objective: Predict dropout risk using ML models.

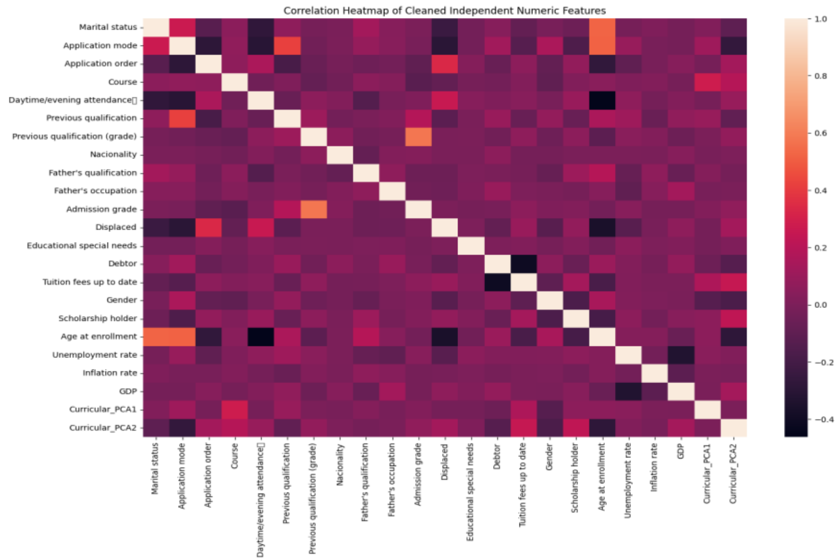
Dataset & Features

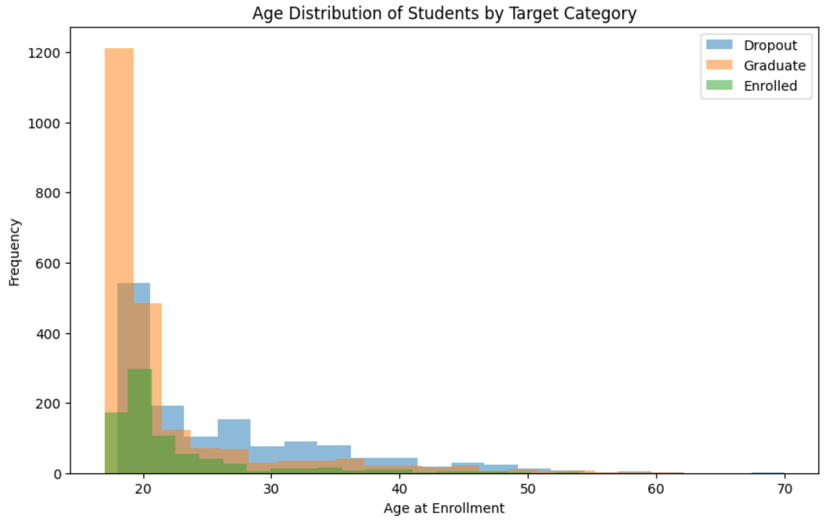
Feature Category	Examples
Demographics	Age, Gender, Marital Status
Academic	Grades, credits, evaluations, approvals
Financial	Debtor, Scholarship, Fee up-to-date
Socio-economic	Unemployment, Inflation, GDP
Target	Transformed to binary: **Dropout = 1**, Enrolled/Graduate = 0

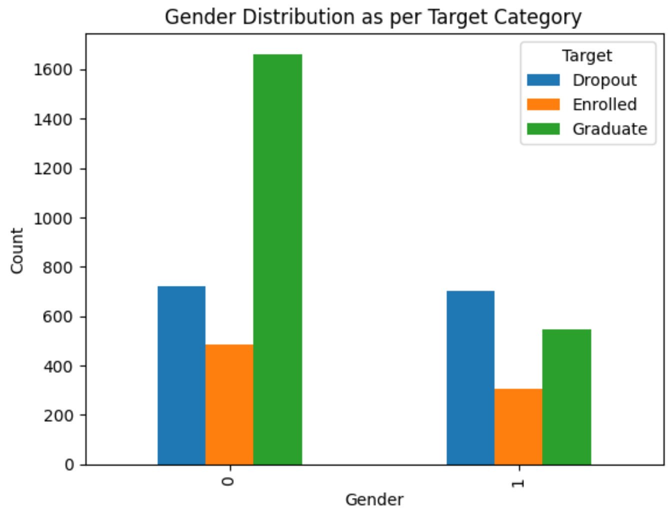
Exploratory Data Analysis (EDA)

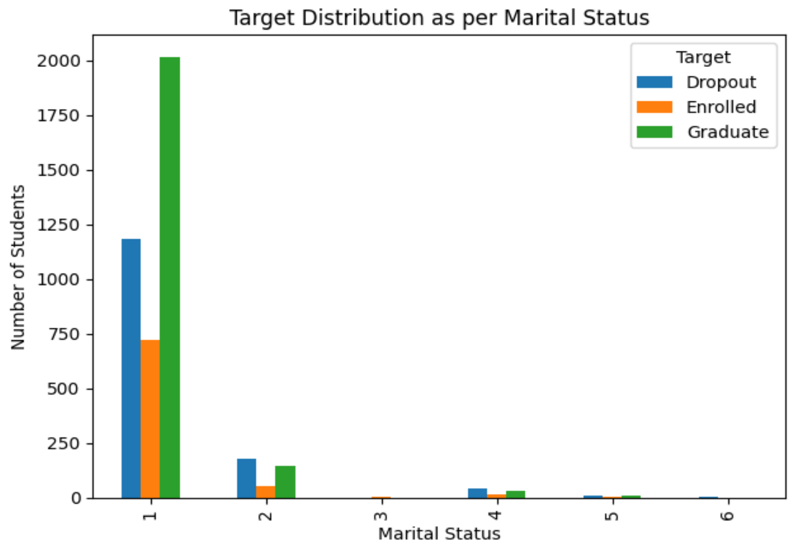
- Distribution of Target Column
- Age distribution across target categories
- Gender vs Target
- Course vs Target
- Marital status vs Target
- Correlation heatmap



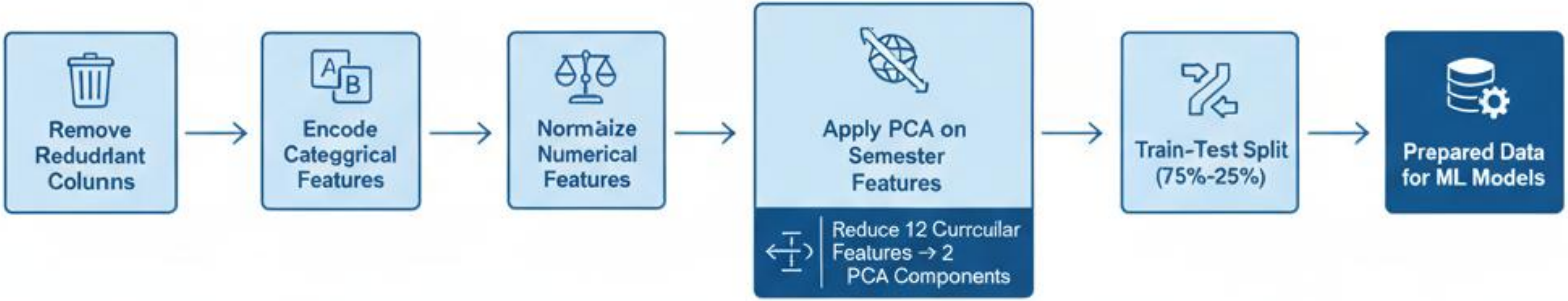






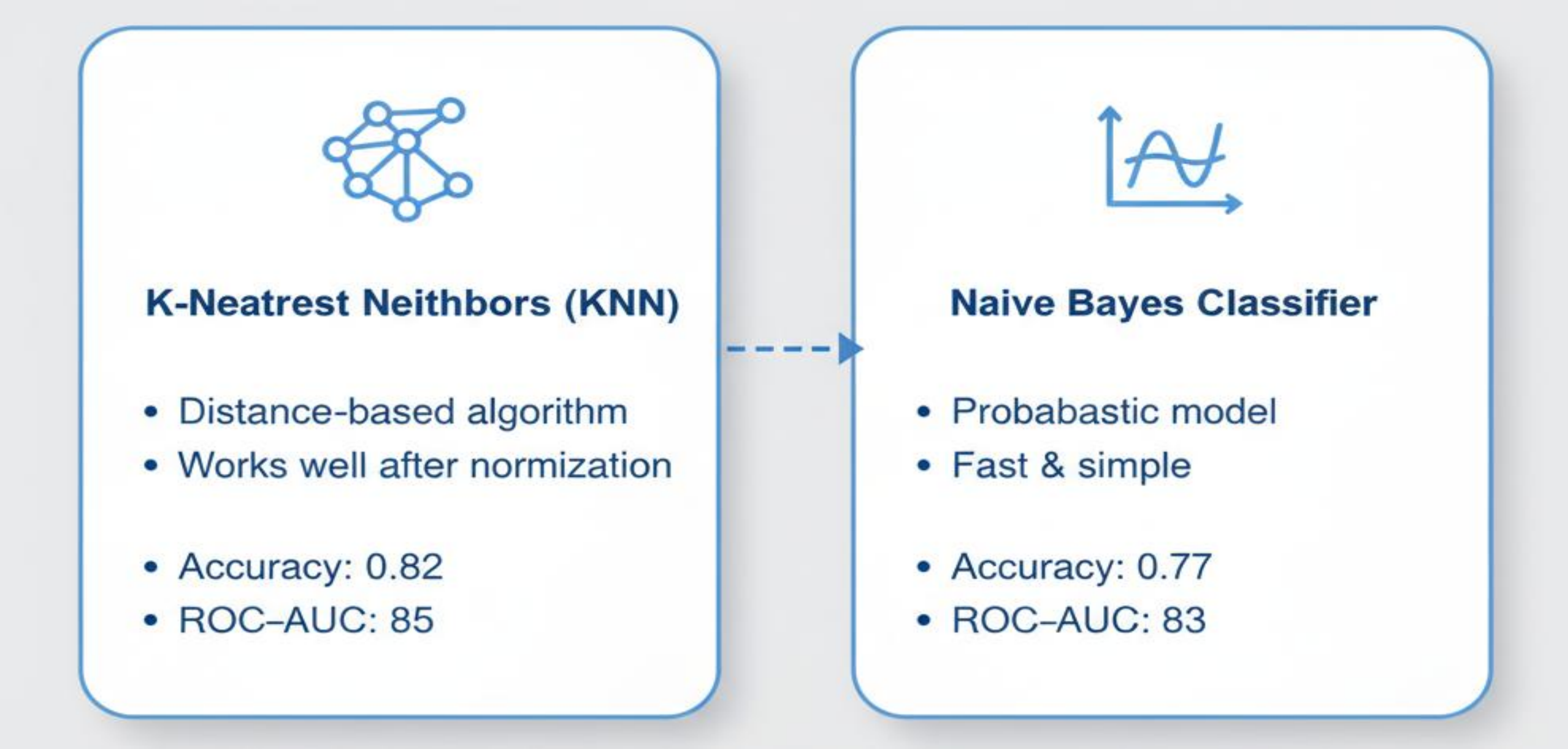


Data Preparation



PCA helped reduce dimoninsinity while keeping important information.

KNN & Naive Bayes Models



Decision Tree & Random Forest

Decision Tree

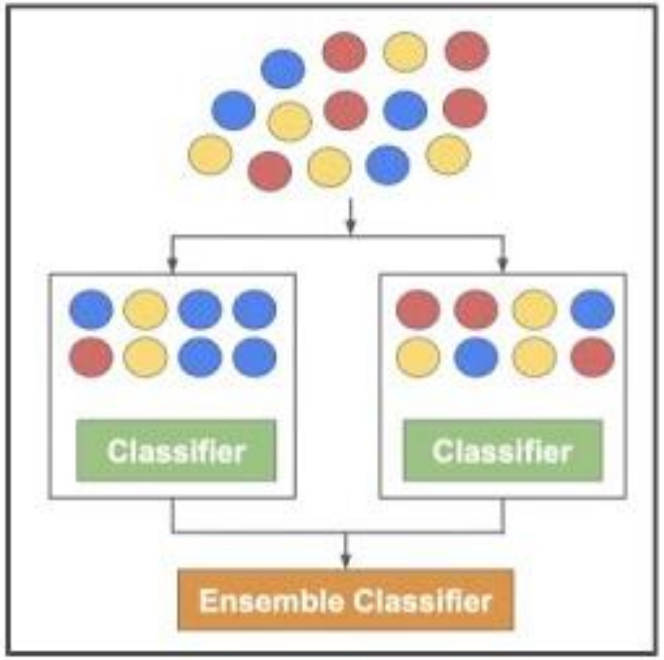
- Easy to intepret
- Moderate performance
- Accuracy: 0.81
- Prone to ovverfting

Random Forest

- Ensemble of multiple decision trees
- Reduces overfting
- Best standolane model
- Accuracy: 87, ROC-AUC: 0.92

Random Forest performed significanly better than a single Decision Tree.

Bagging Ensemble



Key Metrics

- ✓ Reduces variance using multiple Decision Trees
- ✓ **Accuracy:** 0.8670
- ✓ **AUC:** 0.921
- ✓ Produces stable and robust predictions

Boosting Techniques

AdaBoost

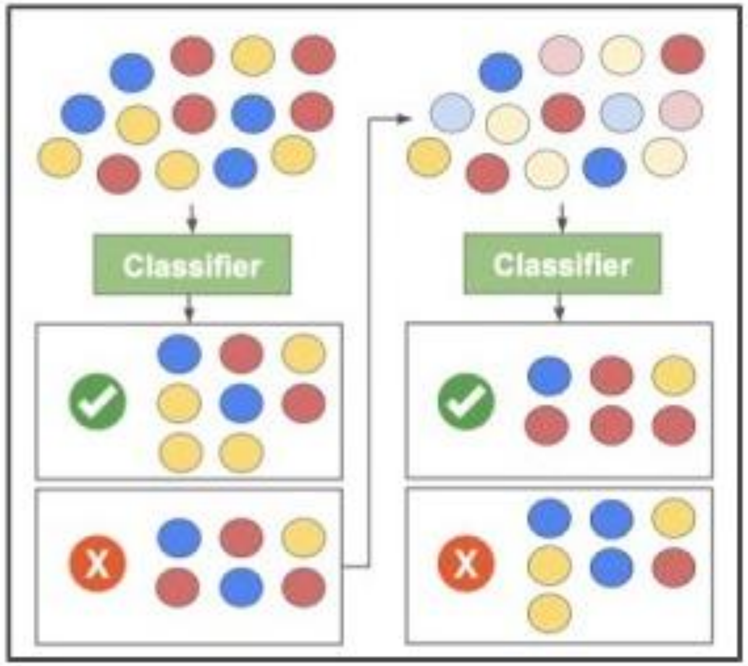
Accuracy: 0.872

AUC: 0.912

Gradient Boosting

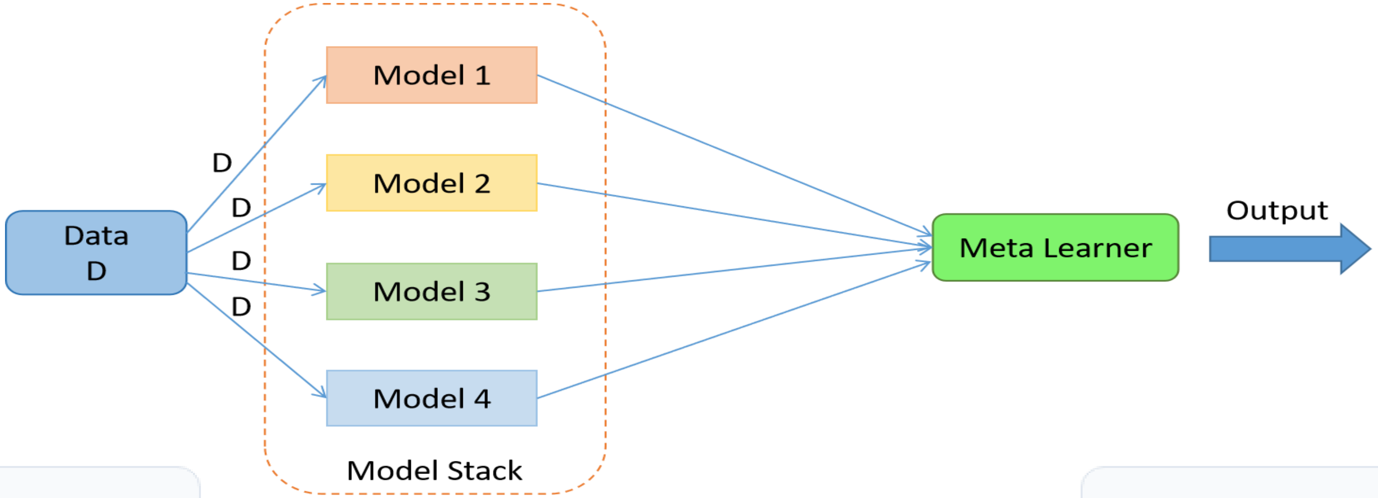
Accuracy: 0.874

AUC: 0.922



Note: Boosting improves classification by focusing on difficult samples.

Stacking Ensemble (Final Model)



Base Models

KNN, Naive Bayes, Decision Tree, and Random Forest.

Meta-Model

Logistic Regression combines the predictions from base models.

Final Result

Best ROC-AUC: 0.928
Conclusion: Stacking model is the most reliable and recommended for deployment.

Practical Outcome

- The model can help institutions identify at-risk students for early academic and counseling interventions.
- **Stacking Ensemble** achieved the highest ROC–AUC (0.928) and excellent accuracy.

Thank You !