Checkpoint: Supervised Learning

Student Academic Outcome
Artificial Intelligence - IART 2024 - 2025

Artur Telo Luís
Gonçalo Joaquim Vale Remelhe
Nuno Pinho Fernandes

Specification

Applying **supervised learning** to predict student academic outcomes:
Graduate, **Dropout**, or **Enrolled**

Dataset includes:

- **Demographic data** (e.g., age, gender)
- **Academic performance** (grades, attendance, past qualifications)
- **Behavioral data** (units completed per semester)
- **Socioeconomic factors** (scholarship, employment, economic difficulties)

Goal:
Identify key factors influencing student outcomes and enable early detection of students at risk of dropping out.

Related Work

Supervised Machine Learning Algorithms for Predicting Student Dropout and Academic Success: A Comparative Study :
https://link.springer.com/article/10.1007/s44163-023-00079-z

Predicting Student Dropouts with Machine Learning: An Empirical Study:
https://www.sciencedirect.com/science/article/pii/S0160791X24000228

Educational Data Mining: Prediction of Students' Academic
Performance:https://slejournal.springeropen.com/articles/10.1186/s40561-022-00192-z

Tools & Success Metrics

Jupyter Notebook – Central workspace for organizing code, models, and results

Pandas – Data manipulation and analysis of structured datasets

NumPy – Efficient numerical computations

Scikit-Learn – Preprocessing, model building, and evaluation

Matplotlib & Seaborn – Visualization of data and results

Metrics:

- ***Accuracy***: Proportion of correct predictions
- ***Recall***: Ability to correctly identify relevant cases
- ***F measure***: Combine Accuracy and Recall

Algorithms

Logistic Regression – Linear model used for binary and multiclass classification -> estimates probabilities using a logistic function.

Decision Tree – Flowchart-like model that splits data based on feature values to make decisions.

Random Forest – Ensemble of decision trees which improves accuracy by averaging their predictions.

k-Nearest Neighbour - Simple and effective algorithm used for both classification and regression

Dataset Overview

Dataset Characteristics:

- **Training Data:** 76,518 student records with 38 features

- **Target Variable :** Three classes (Graduate, Dropout, Enrolled)

Feature Categories :

- **Demographic:** Marital status, age at enrollment, gender, nationality

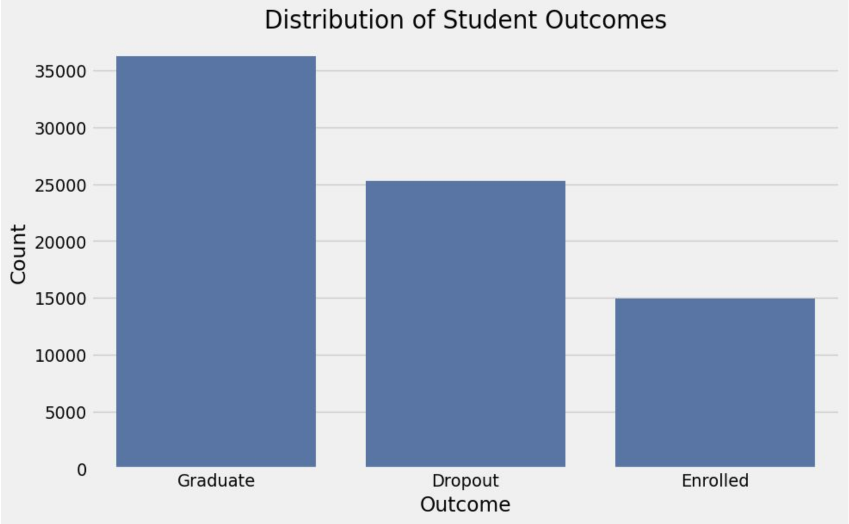
- **Academic:** Previous qualifications, admission grades, curricular units performance

- **Socioeconomic:** Scholarship holder, debtor status, tuition fees status

- **Economic Context:** Unemployment rate, inflation rate, GDP

- **Data Quality:** No missing values detected

- **Class Distribution:** Multi-class classification problem requiring balanced evaluation



Data Preprocessing Pipeline

Key Preprocessing Steps:

- Feature Scaling: StandardScaler for numerical features normalization

- Data Splitting: Train-validation-test split for robust evaluation

- Feature Engineering :

  - Handling categorical variables with appropriate encoding

  - Academic performance metrics aggregation

  - Semester-wise performance analysis

- Pipeline Integration: Scikit-learn pipelines for reproducible preprocessing

- Cross-Validation: Stratified sampling to maintain class distribution

Model Implementation & Training

Algorithm Implementation:

- Logistic Regression: Multi-class classification with regularization

- Decision Tree: Entropy-based splitting with pruning parameters

- Random Forest: 100+ estimators with feature importance analysis

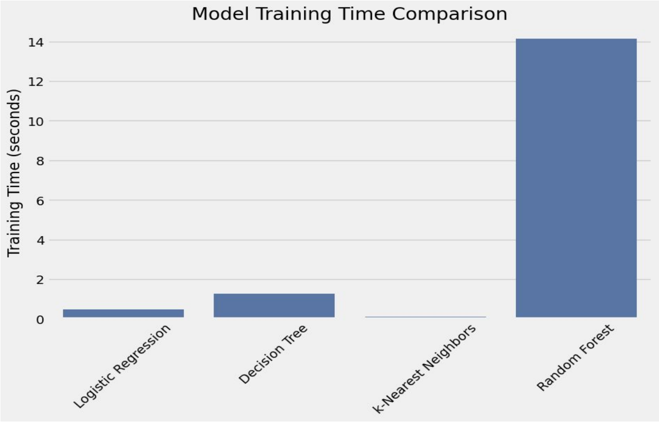
- K-Nearest Neighbors: Optimized k-value through cross-validation

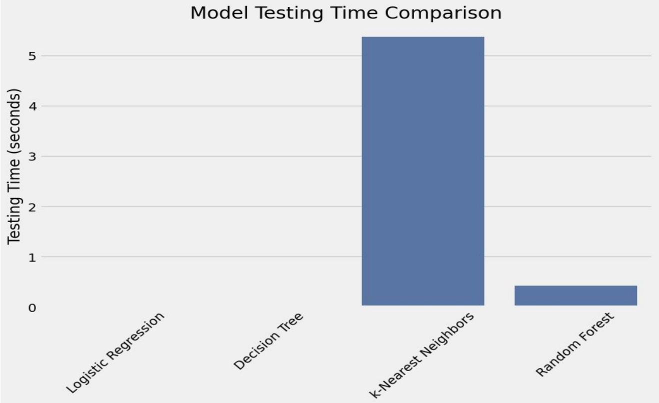
Training Strategy:

- Hyperparameter Tuning: GridSearchCV and RandomizedSearchCV

- Cross-Validation: 5-fold stratified validation

- Performance Optimization: Feature selection and model ensemble techniques





Evaluation Results & Model Comparison

Performance Metrics Achieved:

- Accuracy : 70-75% across different models 2

- Precision : Class-specific precision for Graduate/Dropout/Enrolled

- Recall : Early detection capability for at-risk students

- F1-Score : Balanced performance measure

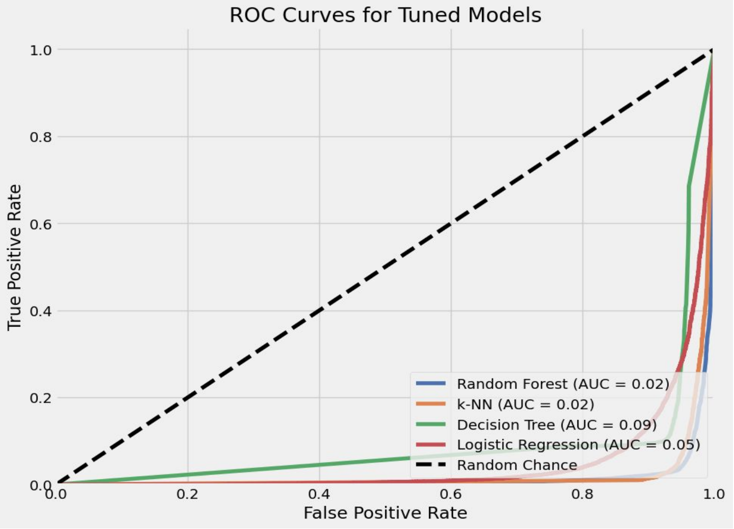
Model Ranking:

1. Random Forest : Best overall performance with feature importance insights

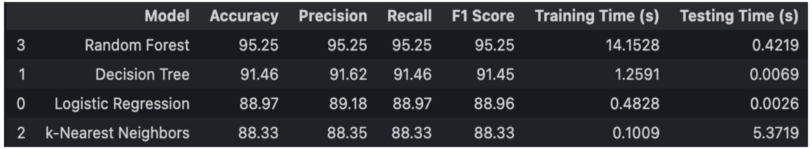
2. Logistic Regression : Strong baseline with interpretable coefficients

3. Decision Tree : Good interpretability but prone to overfitting

4. K-Nearest Neighbors : Effective for local pattern recognition



- The ROC curve is particularly valuable for showing the trade-off between sensitivity and specificity



	Model	Accuracy	Precision	Recall	F1 Score	Training Time (s)	Testing Time (s)
3	Random Forest	95.25	95.25	95.25	95.25	14.1528	0.4219
1	Decision Tree	91.46	91.62	91.46	91.45	1.2591	0.0069
0	Logistic Regression	88.97	89.18	88.97	88.96	0.4828	0.0026
2	k-Nearest Neighbors	88.33	88.35	88.33	88.33	0.1009	5.3719

Key Findings & Feature Importance

Critical Success Factors Identified:

- Academic Performance: Curricular units approved/failed in both semesters

- Previous Qualifications: Admission grades and previous qualification scores

- Attendance Patterns: Daytime vs evening attendance correlation

- Economic Indicators: GDP, unemployment rate impact on dropout risk

- Demographic Factors: Age at enrollment and scholarship status

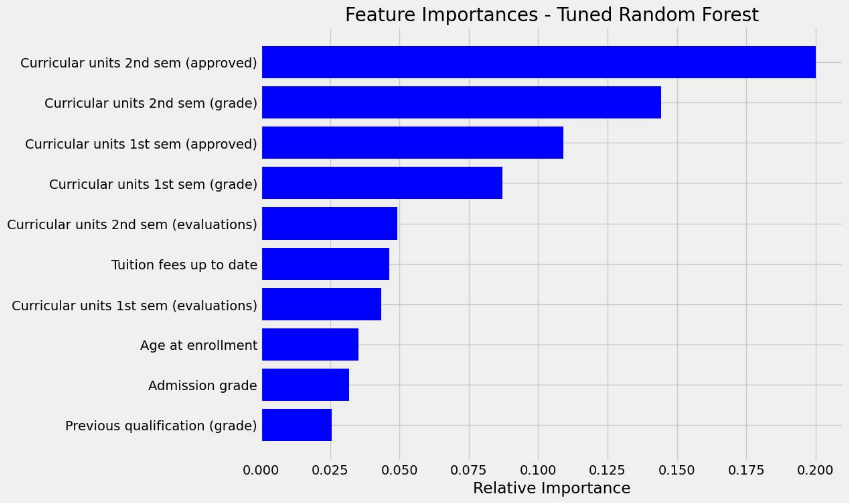
Risk Indicators for Dropout:

- Low curricular units approval rate

- Poor performance in first semester evaluations

- Economic hardship indicators

- Irregular attendance patterns



- This visualization highlights which features had the most impact on predictions
- It provides insights into the factors that most strongly influence student outcomes

Conclusion

Project Achievements:

- Successfully implemented multi-class classification for student outcome prediction

- Achieved competitive accuracy rates comparable to literature standards 1

- Identified key risk factors for early intervention strategies

- Developed reproducible ML pipeline for educational institutions