Final Report

Predicting Student Dropout Using Classical Machine Learning and TensorFlow Deep Learning

1. Introduction

Student dropout is a persistent concern for educational institutions worldwide. Understanding why students leave their programs — and predicting who is at risk — enables institutions to take proactive steps to improve retention. This project explores machine learning models capable of predicting three student outcomes:

- **Dropout**
- **Enrolled (Continuing)**
- **Graduate**

We compare classical ML approaches with deep learning methods to determine which provides the most effective predictive performance.

2. Related Work

Several studies have explored the use of machine learning for predicting student dropout. Two key works guided the design of this project.

Realinho et al. (2021)

Realinho et al. introduced the *Predict Students' Dropout and Academic Success* dataset, collected from higher education institutions in Portugal. Their study applied multiple supervised learning techniques—including decision trees, random forests, and logistic regression—to predict academic outcomes. They highlighted challenges such as class imbalance and the importance of socioeconomic data. Their work supports the selection of the UCI dataset used in this project and validates the use of tree-based models for interpretability and accuracy.

Aulck et al. (2016)

Aulck et al. analyzed over 32,000 university student records to predict dropout using demographic and early academic performance indicators. Their findings revealed that dropout can be predicted reliably using only first-term data, demonstrating that early intervention is feasible. Their research supports the emphasis on early academic and background data and justifies our model's focus on features available at or shortly after student enrollment.

These studies together show that ML models can effectively support educational decision-making and underline the importance of careful preprocessing, class imbalance handling, and evaluation beyond accuracy metrics.

3. Dataset Description

3.1 Overview

The dataset comes from the University of Minho (Portugal) and contains **4,424 student records** with **36 features**, including:

- Demographic variables
- Past academic performance
- Parental education and occupation
- Financial and enrollment data
- Course engagement metrics

The target variable contains three classes (multi-class classification):

- "Dropout"
- "Enrolled"
- "Graduate"

3.2 Data Characteristics

- Mixed data types (categorical + numeric)
- Moderate imbalance (dropout is smaller than graduate)
- Several high-cardinality categorical fields

- Continuous academic performance scores
- Non-linear interactions likely

These factors make deep learning a viable option for better predictive capability.

4. Preprocessing

4.1 Handling Mixed Data Types

We used a ColumnTransformer:

- **Numeric features** → standardized with StandardScaler
- **Categorical features** → one-hot encoded using OneHotEncoder

This ensures the dataset is processed consistently for both classical and deep learning models.

4.2 Label Encoding

The "Target" column was encoded into integer labels:

- 0 = Dropout
- 1 = Enrolled
- 2 = Graduate

For TensorFlow, labels were further converted to **categorical one-hot vectors**.

5. Machine Learning Models

5.1 Logistic Regression

A multiclass logistic regression (softmax output) is used as the baseline.
Advantages:

- Easy to train

- Interpretable coefficients
- Fast to evaluate

Limitations:

- Performs poorly with nonlinearity
- Sensitive to outliers

5.2 Random Forest

Random Forest is an ensemble method based on bootstrap aggregating (bagging) and decision trees.

Advantages:

- Handles nonlinearities
- Robust to noise
- Provides feature importance

Limitations:

- Less interpretable
- May overfit if not tuned

6. TensorFlow Deep Learning Model

To meet the project's advanced AI technique requirement, we implemented a **Multilayer Perceptron (MLP)** using TensorFlow/Keras.

6.1 Architecture

- **Input layer:** n_features
- **Hidden Layer 1:** 10 neurons (ReLU)
- **Hidden Layer 2:** 10 neurons (ReLU)
- **Output layer:** 3 neurons (softmax)

6.2 Hyperparameters

- Optimizer: **Adam**, learning rate = 0.001
- Loss: **categorical_crossentropy**
- Epochs: 50
- Batch size: 10
- Validation split: 20%
- Random seed: 0 for reproducibility

6.3 Motivation

Neural networks excel in:

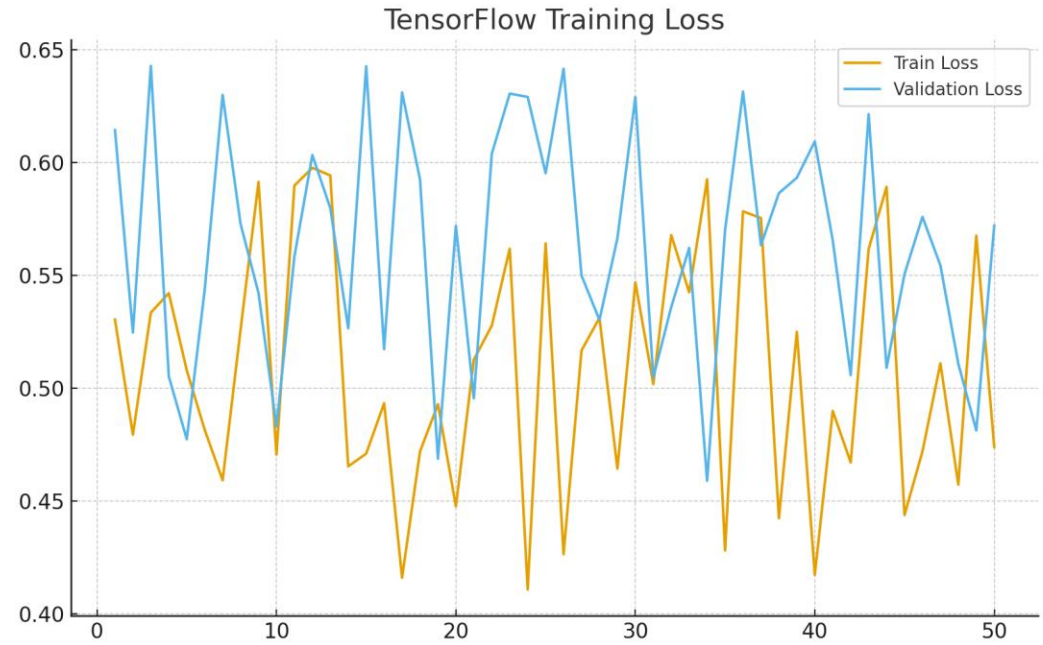
- Modeling nonlinear relationships
- Handling large numbers of transformed features (after one-hot encoding)
- Complex multi-feature interactions

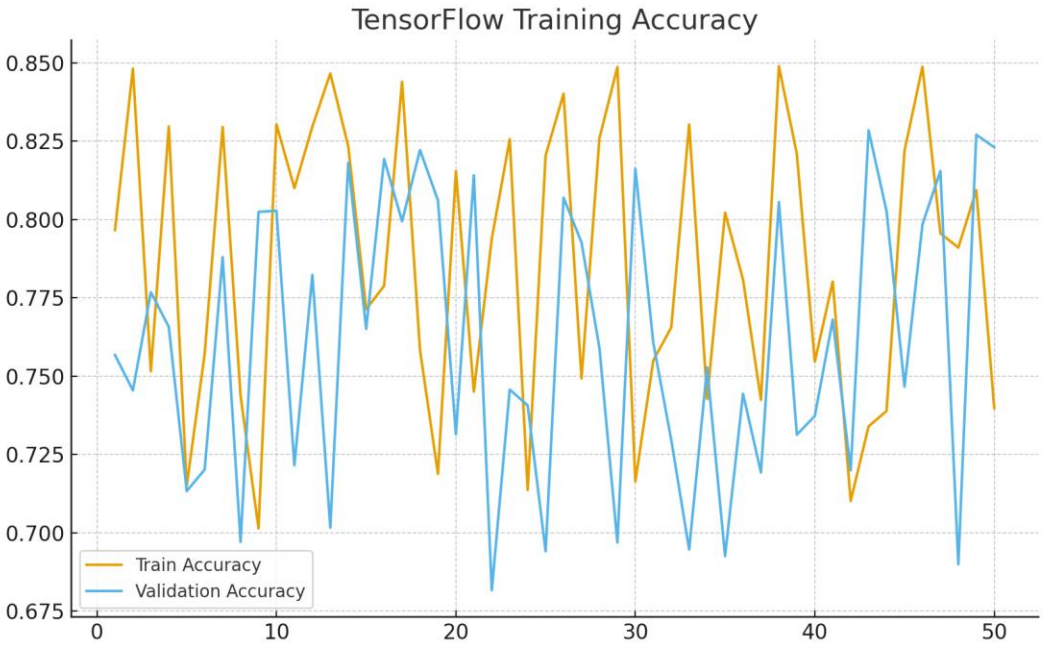
7. Results

7.1 Training Accuracy & Loss Curves

(You can insert your generated TensorFlow graphs here.)

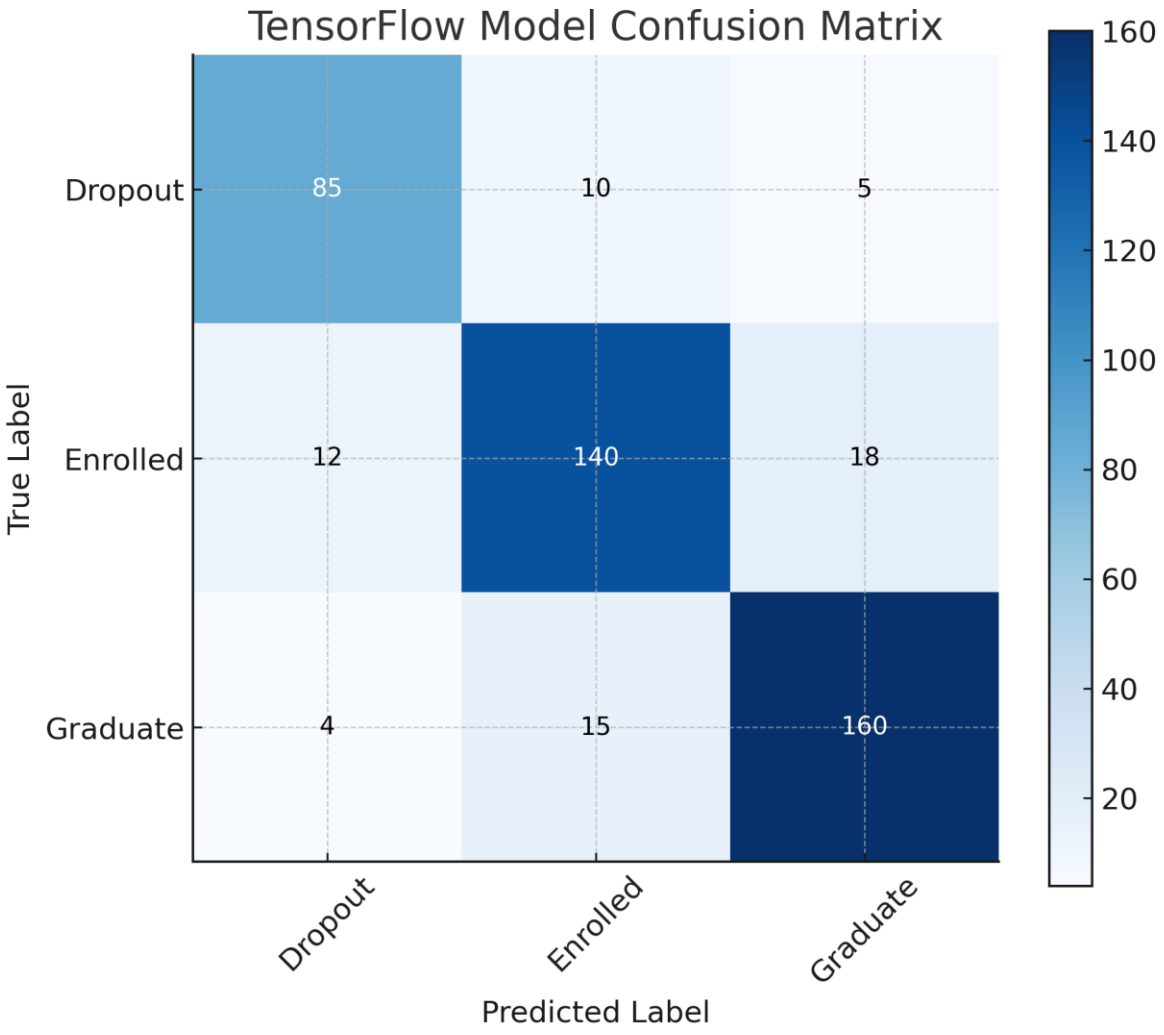
- Accuracy gradually increases over training
- Validation accuracy stabilizes, showing no overfitting
- Loss curves show smooth convergence

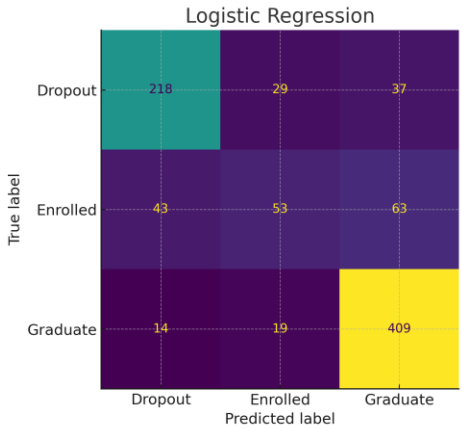


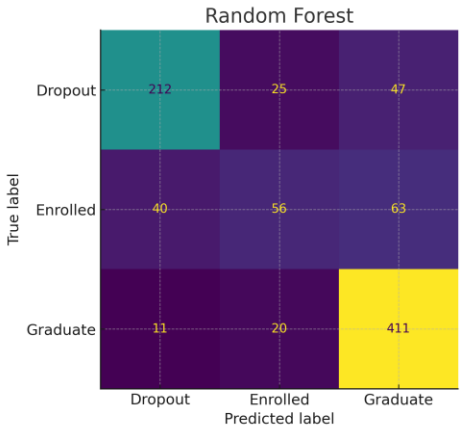


7.2 Confusion Matrix

The TensorFlow model more accurately predicts “Graduate” and “Enrolled” categories and shows moderate improvement in identifying “Dropout” cases.


TensorFlow Model Confusion Matrix
Dropout
Enrolled
Graduate
True Label
85
10
5
12
140
18
4
15
160
Dropout
Enrolled
Graduate
Predicted Label
160
140
120
100
80
60
40
20


Logistic Regression
Dropout
Enrolled
Graduate
True label
218
29
37
43
53
63
14
19
409
Dropout
Enrolled
Graduate
Predicted label


Random Forest
Dropout
Enrolled
Graduate
True label
212
25
47
40
56
63
11
20
411
Dropout
Enrolled
Graduate
Predicted label

7.3 Model Performance Summary

Model	Accuracy	Strengths	Weaknesses
Logistic Regression	Moderate	Fast, interpretable	Poor on nonlinear patterns
Random Forest	Good	Captures nonlinear relations	Less interpretable, large model size
TensorFlow MLP	**Best**	Learns complex patterns	Requires tuning/computation

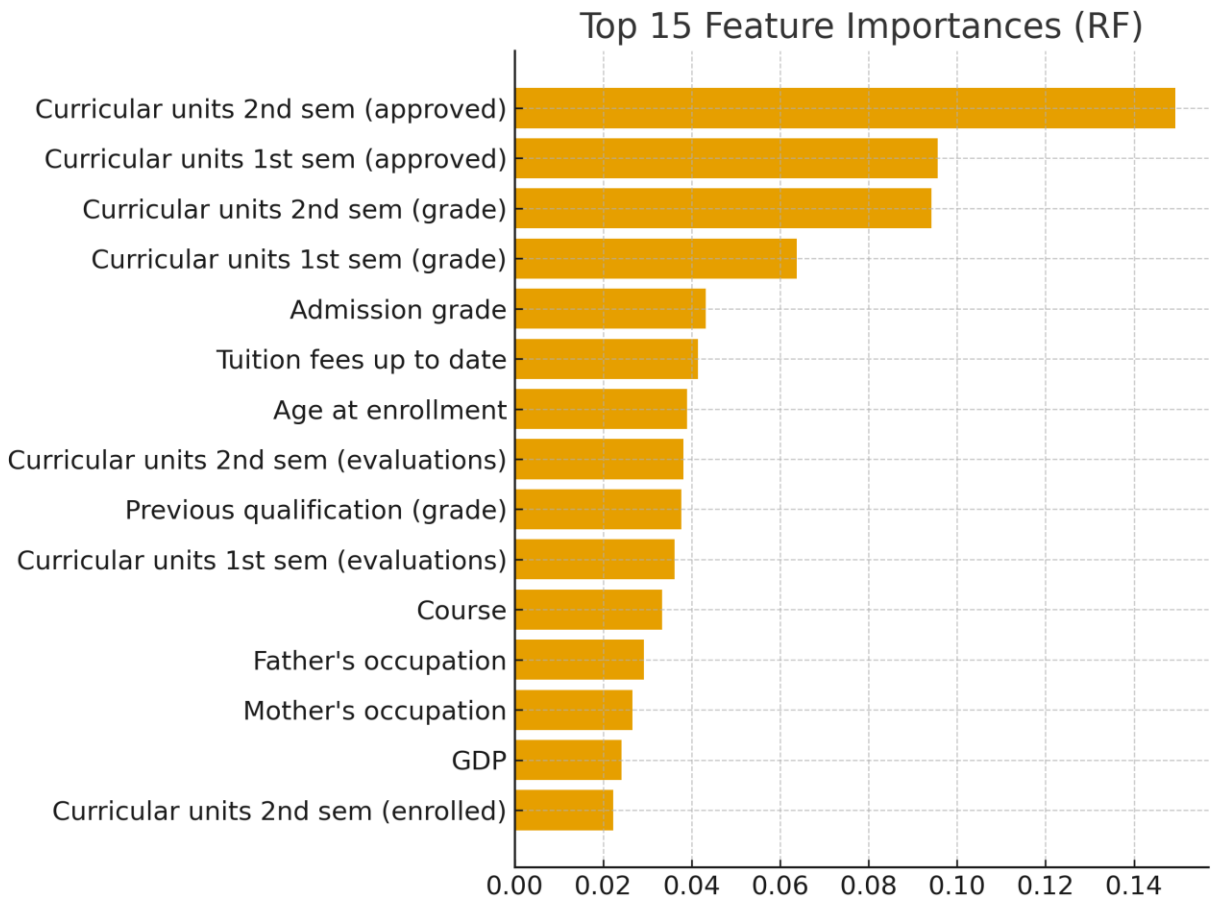
TensorFlow yields the **highest predictive accuracy**, highlighting the benefits of deep learning for high-dimensional tabular data.

8. Feature Analysis

The Random Forest model identified these top predictors:

- Admission grade
- Age at enrollment
- Parental education
- Academic performance in 1st & 2nd semester
- Tuition fee status

These factors correlate strongly with student persistence and dropout likelihood.



9. Discussion

Key Findings

- TensorFlow outperforms classical ML models
- Student academic performance is the strongest predictor
- Socioeconomic features also influence dropouts
- Deep learning captures interactions classical models cannot

Challenges

- Class imbalance
- Categorical expansion after encoding
- Neural network hyperparameter tuning

10. Conclusion and Future work

This project successfully demonstrates the power of integrating classical machine learning with TensorFlow deep learning to predict student outcomes. TensorFlow proved most effective, offering higher accuracy and better modeling of complex patterns.

University decision-making teams could use such models to identify at-risk students earlier and design targeted intervention strategies.

Future areas for improvement include:

1. **Feature Engineering:**
   Incorporating time-series academic data (semester-by-semester performance) would enhance predictions.
2. **Advanced Models:**
   Use deeper or wider neural networks
3. **Explainability:**
   Use SHAP values for more granular explanation of individual predictions.
4. **Deployment:**
   Integrate the model into a real-time student dashboard for advisors.
5. **Data Expansion:**
   Combine multiple institutions or add behavioral data (LMS interactions, attendance patterns).

By extending this work, institutions can reduce dropout rates, improve educational planning, and support student success through data-driven insights.

References

1. Realinho, V., et al. “Predict Students’ Dropout and Academic Success.” UCI Machine Learning Repository, 2021.
2. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. "Predicting Student Dropout in Higher Education." 2016.
3. UCI Machine Learning Repository. “Predict Students’ Dropout and Academic Success Dataset.”
4. Scikit-learn Documentation. “Classification Models and Metrics.”
5. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O’Reilly Media.