DATX05 PREDICTIVE ANALYTICS 2025

Final Assessment

Title

Predicting School Dropout Risk in South Africa Using Machine Learning and Deep Learning Approaches

Overview/Brief

This final assessment aims to develop predictive models that identify learners at risk of school dropout in South Africa by applying modern Machine Learning (ML) and Deep Learning (DL) techniques. The primary objective is not only to achieve strong predictive performance but also to produce novel interpretable, equitable, and actionable insights that can guide policymakers, educators, and school administrators in evidence-based decision-making.

In the model development phase, apply binary classification approaches using either machine learning or deep learning methods, such as logistic regression, random forests, CNN or neural networks, where the target variable is defined as *dropout (1) versus completion (0).*

Novelty Angles to Explore

Students are expected to incorporate novelty and analytical rigour into their research. You may select one of the proposed novelty angles below or suggest your own innovative direction, provided it aligns with the overall study objectives.

1. Beyond Standard Binary Classifiers

Instead of just using Logistic Regression, Random Forests (**RF**), or basic Neural Networks (**NN**), consider:

- *Ensemble Stacking/Super Learners***:** Combine the predictions of your initial models (Logistic, RF, NN) using a *second-layer meta-model* (like another Logistic Regression or a Ridge Classifier) to see if you can achieve superior predictive power. This is more advanced than simple averaging or voting ensembles. OR
- *Gradient Boosting Machines (GBMs): Techniques like XGBoost, LightGBM, or CatBoost* are often state-of-the-art for structured/tabular data classification and are typically more powerful than standard Random Forests. Their performance often sets a high bar. OR
- *Deep Learning (DL) Novelty***:** For your **NN**, go beyond a simple Multi-Layer Perceptron (MLP). Explore other deep learning models.

2. Explainability Layer

Most existing dropout prediction studies focus primarily on predictive accuracy, often reporting metrics such as the AUC or F1 score. This novelty angle encourages the development of interpretable models that explain why certain predictions are made.

Approach:

- Apply model-agnostic interpretability techniques such as SHAP (Shapley Additive Explanations) or LIME to identify which features contribute most to dropout risk predictions.
- Present and visualise feature importance in ways that policymakers and educators can easily interpret and apply.
- Example insight: *"Learners from households earning less than R3,000 per month who have repeated grades are 3.5× more likely to drop out."*

Novelty:

- Moves beyond predictive accuracy to introduce interpretable AI in education, thereby ensuring that model outcomes are transparent, explainable, and policy-relevant.

OR

3. Fairness / Equity Analysis

School dropout risk is not evenly distributed across the population. Significant disparities often exist along gender, geographic, and socio-economic dimensions. This novelty angle focuses on assessing and improving fairness within ML models.

Approach:

- Evaluate model performance across key subgroups such as Gender (Male vs Female), Location (Urban vs Rural), and Province (e.g., Gauteng, Western Cape, Eastern Cape).
- Utilise fairness metrics such as Equal Opportunity Difference (to determine if predictions are equally accurate across groups) and Demographic Parity.
- If disparities are detected, adjust the models using methods such as reweighting, group-specific thresholds, or fairness-aware learning algorithms.

Novelty:

- Addresses a critical gap in current research by explicitly evaluating fairness in educational ML models.
- Demonstrates whether ML models reinforce or reduce existing inequalities, offering evidence-based guidance for inclusive educational interventions.

Guidance for Students

1. Select one novelty angle to ensure a focused and comprehensive research project. Clearly document all assumptions, data preprocessing steps, and the rationale behind your chosen ML/DL algorithms.
2. The submission deadline is **23:59 on 25 October 2025.** Late submissions will not be accepted.
3. Submit the following as separate files**: Final PDF mini research report, Turnitin similarity report, and appendix** containing your code.
4. Save your final report using the following naming convention: firstname_surname.pdf
5. The content to be examined, exempting the ad hoc pages (cover page, appendix, references list, etc), should not be more than 12 pages. Please adhere strictly to this limit, as excessively long reports will be penalised.

The mini research report should be structured as follows:

Title
Abstract
Introduction
Literature review (brief)
Methodology
Application
Discussion
Conclusion

Assessment Objective

This assessment aims to demonstrate your ability to design, implement, and evaluate advanced predictive analytics models that are both technically sound and socially responsible.

Data

A Python file has been provided to generate the data