Final Report

Project Title: Analyzing Educational Inequality: A Machine Learning Approach to Predicting Dropout Risks in U.S. Schools

Author: Farida Kendi Nkatha

Date: October 2025

Executive Summary

This project aims to identify students at risk of dropping out using machine learning techniques applied. This was achieved through a dataset from Kaggle which explores education inequality across 1,000 schools in the United States.

Use of engineering features that ensure funding efficiency and applying resampling strategies like

SMOTE, improved the model's ability to detect vulnerable students. The findings highlight the importance of balancing predictive accuracy with ethical responsibility in educational interventions.

Introduction

Education inequality remains a pressing issue, especially in under-resourced schools. Early identification of dropout risk is critical for targeted support

The objective was to build a classification model that predicts dropout risk using school-level indicators.

The dataset included indicators such as funding per student, average test scores, student-teacher ratios, low-income and minority student percentages, internet access levels, and dropout rates. It focused on US; public, private and charter schools across selected states in US

Methodology

Data Preprocessing

- My dataset did not have any missing values.
- Converted categorical variables using label encoding (grade_level) and one-hot encoding (school_type, state).
- Scaled numerical features using StandardScaler.

Feature Engineering

• Created funding_per_score to measure funding efficiency.

• Defined binary target dropout_risk based on dropout rate threshold (>10%).

• The dropout_risk portrayed the class distribution. Having a moderately imbalanced dataset of 65% no risk vs 35% at risk of dropout, made my model susceptible to favor the majority class (0) and miss true dropout cases (1). Applying the use of SMOTE enabled my model to learn from more balanced data.

Model Selection

To train the classification model, I used the processed dataset df_model, where the target variable

dropout_risk was moderately imbalanced (645 no-risk vs 355 at-risk cases). The features were separated from the target, and the data was split into training and testing sets using an 80/20 ratio with a fixed random seed for reproducibility.

To mitigate bias toward the majority class, I applied SMOTE (Synthetic Minority Over-sampling Technique) to the training data. SMOTE generates synthetic examples of the minority class, which

helps the model learn more diverse patterns and improves recall for dropout risk cases. SMOTE was applied only to the training set to prevent data leakage and preserve the integrity of the test evaluation.

I selected Random Forest as the primary classifier due to its robustness, interpretability, and ability

to handle both numerical and categorical features. Random Forest is an ensemble method that builds multiple decision trees and aggregates their predictions, reducing overfitting and improving

generalization. I also enabled class_weight='balanced' to also help address class imbalance by penalizing misclassification of the minority class.

Model Evaluation

To assess the performance of the trained Random Forest model, I used standard classification metrics provided by sklearn, including the confusion matrix, precision, recall, F1-score, and accuracy. These metrics offered an understanding of how well the model distinguished between students at risk of dropping out (1) and those not at risk (0).

The confusion matrix and classification report on the model’s predictions on the test set, helped quantify both correct predictions and misclassifications for the minority class.

Confusion matrix

```
[[99 25]
[62 14]]
```

- True Negatives (99): Correctly predicted no-risk students.
- False Positives (25): Predicted at-risk but were not.
- False Negatives (62): Missed actual at-risk students.
- True Positives (14): Correctly identified at-risk students.

Classification Report

- Accuracy: 56%- overall correct predictions.
- Macro Avg F1: 0.47 -treats both classes equally.
- Weighted Avg F1: 0.52- accounts for class imbalance.

The model performs reasonably well for the majority class (no risk), with 80% recall and 69% F1 score. However, performance on the minority class (dropout risk) remains limited, with only 18% recall and 24% F1-score. This indicates that while the model has improved after applying SMOTE, it still struggles to identify all at-risk students.

Model optimization

Hyperparameter Tuning and Cross-Validation

To optimize model performance, I conducted a grid search over key hyperparameters of the Random Forest classifier, including; max_depth, min_samples_split, and n_estimators. The model was evaluated using 5-fold cross-validation, with F1-score as the scoring metric to balance precision and recall.

The best-performing configuration was: max_depth = 20, min_samples_split = 5, n_estimators = 100 and Mean F1-score = 0.144

This configuration was selected for final evaluation, as it offered the highest sensitivity to dropout

risk cases while maintaining generalizability across folds.

Discussion

- Model Strengths: Improved recall for dropout risk cases after SMOTE; better balance between precision and recall.
- Limitations: Overall accuracy remains modest; model still struggles with minority class.
- Ethical Considerations: False negatives (missed dropout risks) carry high social cost. Prioritizing recall is justified in this context.
- Schools flagged by the model can be targeted for early intervention, funding audits, or support programs.

Conclusion

The model demonstrates that machine learning can support educational equity by identifying at risk students. While performance is imperfect, strategic resampling and feature engineering significantly improve sensitivity to dropout risk. Future work should explore ensemble methods, cost-sensitive learning, and stakeholder-informed thresholds.

Recommendations

- Incorporate qualitative data (e.g., teacher feedback, student surveys).
- Use model outputs to inform targeted interventions, not punitive measures.