AI Project Report: Predicting Student Dropout Rates in Kenyan Universities

1. Problem Definition

Problem:

Predicting student dropout rates in Kenyan universities using academic, financial, and behavioral data.

Objectives:

1. Identify students at risk of dropping out early.
2. Improve student retention rates by informing interventions.
3. Support policy-making in higher education.

Stakeholders:

- University administration
- Ministry of Education

Key Performance Indicator (KPI):

Dropout Prediction Accuracy - Measuring the model's ability to correctly classify students who are likely to drop out, targeting accuracy above 85%.

2. Data Collection & Preprocessing

Data Sources:

1. University student records (grades, attendance, course load).
2. Financial aid and fee payment data (e.g., HELB loan status).

Potential Bias:

Socioeconomic Bias - The dataset may overrepresent students from urban areas who have consistent internet access and better educational backgrounds, potentially underrepresenting students from rural or marginalized areas.

Preprocessing Steps:

1. Handling Missing Data - Impute missing grades or income data using averages or machine learning methods.

2. Normalization - Standardize continuous features such as GPA and age for better model performance.

3. Encoding Categorical Variables - Convert non-numerical data like program or gender into numerical format using label or one-hot encoding.

3. Model Development

Model Choice:

Random Forest - A powerful ensemble model ideal for mixed data types and capable of handling nonlinear relationships. It reduces overfitting by averaging multiple decision trees.

Data Splitting:

- 70% Training Set - Used to train the model.

- 15% Validation Set - Used for hyperparameter tuning and performance validation.

- 15% Test Set - Used for final evaluation.

Hyperparameters to Tune:

1. n_estimators - Number of trees in the forest, which affects performance and overfitting.

2. max_depth - The maximum depth of each tree to control model complexity and prevent overfitting.

4. Evaluation & Deployment

Evaluation Metrics:

1. Precision - Ensures students falsely flagged as at risk are minimized.

2. Recall - Captures as many actual dropouts as possible, helping with early interventions.

Concept Drift:

Concept drift refers to changes in the data patterns over time, such as new academic policies or curriculum updates. This may reduce model accuracy if unaddressed.

Monitoring Strategy:

- Regular retraining with updated data.
- Performance tracking through dashboard metrics every academic term.

Technical Deployment Challenge:

Scalability - Deploying the model across various institutions with large datasets may strain computational resources. Solutions include cloud infrastructure and batch processing.

References

1. Breiman, L. (2001). Random Forests. Machine Learning.

2. University of Nairobi Student Records (Hypothetical source).

3. Kenya Higher Education Loan Board (HELB) - Financial Aid Data.

4. Gama, J. et al. (2014). A survey on concept drift adaptation.

5. Provost, F., & Fawcett, T. (2013). Data Science for Business.