

Student Dropout Analysis - Summary Report

1. Data Insights

The dataset focuses on analyzing student dropout behavior based on various features such as gender, course type, session count, and satisfaction rating. Exploratory Data Analysis (EDA) revealed the following insights:

- Dropout rates differ significantly between genders and course types.
- A negative trend is observed between session count and dropout likelihood.
- Students with lower satisfaction ratings and fewer sessions are more likely to drop out.
- Correlation heatmaps identified moderately related features that influence dropout behavior.

2. Modelling Approach

The following steps were performed for model development:

- Removal of the non-informative 'student_id' column.
- Label encoding for categorical variables: 'gender' and 'course_type'.
- Standardization of features using StandardScaler.
- Dataset split into training and testing sets (80%-20% ratio).
- Two models were trained:
 - Logistic Regression (baseline linear model).
 - Random Forest Classifier (ensemble method offering improved performance).

3. Evaluation Results

The models were evaluated using accuracy, precision, recall, and F1-score metrics:

- Logistic Regression offered a simple yet effective baseline.
- Random Forest provided better classification results due to its ensemble nature.
- Classification reports included detailed metric values for both models (see notebook for full output).

4. Business Recommendations

- Implement early intervention strategies for students showing low satisfaction and low engagement.
- Personalize learning support based on course type and gender-specific patterns.
- Utilize predictive models to generate dropout risk alerts and target high-risk students.
- Continuously monitor and refine prediction models using new student data for better retention outcomes.