Forecasting Student Dropouts and Academic Success – Summary Report

1. Problem Statement

Student dropout is one of the major challenges faced by educational institutions globally. It affects students' futures and institutional performance. Early identification of at-risk students enables universities to provide timely support. This project aims to predict student dropout likelihood and forecast academic success using demographic, socio-economic, and educational data.

2. Objectives

- Clean and preprocess raw student data for modeling.
- Perform exploratory data analysis (EDA) to uncover factors associated with dropout and academic success.
- Train classification models to predict dropout risk and regression models to forecast academic performance.
- Identify key features (e.g., parental education, study time, previous grades) influencing student outcomes.
- Build a system that can assist administrators and counselors in proactive student support.

3. Dataset Overview

The dataset includes over 4,000 student records with 35+ features covering academic, demographic, financial, and regional data. The target variable 'Outcome' indicates whether the student dropped out, continued, or graduated.

4. Methodology

1. Data Preprocessing – handled missing values, encoded categorical variables, normalized data.
2. Exploratory Data Analysis – identified correlations and trends affecting dropout.
3. Model Training – applied Logistic Regression, Random Forest and XGBoost for classification; Linear, Ridge/Lasso, and Decision Tree models for regression.
4. Model Evaluation – used Accuracy, F1-score, Precision/Recall, RMSE for performance metrics.

5. Results

• XGBoost achieved the best accuracy (~76%) for dropout prediction.
• Random Forest showed balanced precision and recall.

• Logistic Regression achieved the lowest RMSE for grade forecasting.
• Combining academic and socio-economic factors improved accuracy.

6. Key Insights

• Academic performance in early semesters strongly predicts dropout risk.

• Parental education and financial stability are major success factors.

• Outstanding debts and low attendance correlate with higher dropout likelihood.

• Scholarship recipients show better retention rates.

7. Conclusion

Integrating academic, financial, and socio-economic data into predictive models helps identify students at risk early. These insights allow educational institutions to implement targeted interventions that improve retention and success rates.

8. Tools and Technologies

• Python (Pandas, NumPy, Scikit-learn, XGBoost)

• Data Visualization (Matplotlib, Seaborn, Plotly)

• Model Evaluation Metrics: Accuracy, F1-score, Precision/Recall, RMSE