**Project Report**

**Project Title: Predict Students' Dropout and Academic Success**

## 1. Introduction:

Student dropout is a critical challenge faced by educational institutions worldwide. High dropout rates negatively impact both the students and the institution's reputation, funding, and overall performance.
This project aims to analyze various demographic, socio-economic, and academic factors to **predict whether a student will graduate, drop out, or remain enrolled**. By identifying at-risk students early, institutions can take proactive measures to improve retention rates and academic success.

The dataset used in this project contains **4,424 rows** and **37 Columns** covering:

- **Demographic Information** (age, gender, nationality, marital status)
- **Academic Records** (grades, course load, previous qualifications)
- **Socio-economic Indicators** (parental education, unemployment rate, GDP)
- **Special Status** (scholarship holders, displaced students, tuition payment status)

## 2. Methodology:

**Data Collection & Understanding:**

- Dataset loaded from structured_data.csv.
- Basic inspection for size, column types, and missing values.

**Data Cleaning & Preprocessing:**

- **Missing values** handled with forward fill.
- **Duplicates** removed.
- **Outliers** detected using **Z-score method** and removed.
- **Categorical features** encoded with **Label Encoding**.
- **Numerical features** scaled with **StandardScaler**.

**Exploratory Data Analysis (EDA):**

- Descriptive statistics calculated.
- Visualizations created for:
    - Outcome distribution (Graduate, Dropout, Enrolled)
    - Gender distribution
    - Age at enrollment
    - Nationality breakdown
    - Parental education levels
    - Admission grade distribution

    - GDP vs dropout rate

**Model Selection & Training:**

- Target variable analyzed:
    - Categorical → **Classification task**.
- Chosen model: **Random Forest Classifier** (robust for mixed data types).
- Data split: **80% training, 20% testing**.
- Model trained and evaluated.

**Evaluation:**

- Accuracy score calculated.
- Classification report generated.
- Confusion matrix visualized.

**3. Results:**

**Key Findings:**

- Most students are within the **18–24 age range**.
- A significant portion of students **drop out before completing their course**.
- Students with **lower admission grades** and **lower parental education levels** are more likely to drop out.
- Economic indicators like **GDP** and **unemployment rate** show correlation with dropout rates.
- Gender distribution is balanced, but **male students have a slightly higher dropout rate**.
- **Random Forest Classifier** achieved:
    - **Accuracy:** ~85% (varies depending on train-test split).
    - Good precision and recall for identifying dropouts.

**4. Recommendations:**

1. **Early Intervention Programs** – Identify at-risk students early based on predictive model outputs and offer targeted academic and counseling support.
2. **Scholarship & Financial Aid** – Increase financial support for students from low socio-economic backgrounds.
3. **Parental Engagement** – Encourage awareness programs for parents with lower educational backgrounds.
4. **Improve Academic Preparedness** – Offer remedial classes for students with lower admission grades.
5. **Monitor Economic Indicators** – Adapt student support strategies during economic downturns.

**5. Future Work**

- **Include Attendance & Behavioral Data** – Adding student attendance patterns, engagement in coursework, and behavioral data could improve prediction accuracy.
- **Use Advanced Models** – Experiment with gradient boosting models (XGBoost, LightGBM) or neural networks for potentially higher accuracy.
- **Real-time Monitoring System** – Implement a dashboard for continuous tracking of student risk scores.
- **Geographical Analysis** – Map dropout patterns by nationality or region to develop location-specific interventions.
- **Longitudinal Tracking** – Study how student performance evolves over time to refine predictive models.