Student Dropout Prediction Report

Tools Used: Python (Pandas, Scikit-learn, Seaborn, Matplotlib, Joblib), SQL Server, Power BI

1. Project Overview

This project is designed to identify and prevent student dropouts by using machine learning classification techniques on educational and behavioral data. The pipeline includes:

- Data cleaning and transformation
- Exploratory data analysis and visualization
- Feature engineering and encoding
- Model training using Logistic Regression and Random Forest
- Evaluation and interpretation of models
- SQL-based insight mining
- Dashboards for stakeholders using Power BI

Project Goal: Predict whether a student will drop out based on available features and derive actionable insights to support early intervention strategies.

---

2. Dataset Overview

The dataset includes demographic, academic, and behavioral variables:

- **Source:** Encoded and cleaned dataset from Kaggle
- **Total Records:** 649 students
- **Target Variable:** `Dropped_Out` (0 = Stayed, 1 = Dropped Out)
- **No missing or duplicate records**
- **Key Feature Categories:**
  - Demographics: Gender, Parental Education
  - Academic: GPA, Study Time, Number of Failures
  - Behavioral: Internet Access, Extra-curricular Activities, Absences

---

3. Data Cleaning & Preprocessing

- **Encoding:**
  - Binary columns (Yes/No) converted to 0/1
  - Multi-class categorical features one-hot encoded
- **Outlier Handling:**
  - Capped `Number_of_Absences` at 95th percentile

- **Target Conversion:** `Dropped_Out` cast to integer binary
- **Train-Test Split:**
  - Used `StratifiedShuffleSplit` to maintain dropout distribution
  - 80% training, 20% testing

---

4. Model Building & Evaluation

- **Logistic Regression**
  - Accuracy: 97%
  - Strengths: Simple, interpretable, fast
  - Use Case: When explainability is critical
- **Random Forest**
  - Accuracy: 100% (suspected overfitting)
  - Parameters: `n_estimators=50`, `max_depth=5`
  - Strengths: Robust to outliers, handles complex relationships
- **Evaluation Metrics:**
  - Confusion Matrix, Accuracy Score, Classification Report
  - Confusion matrix visualized via heatmap

---

5. Feature Importance

Top features from Random Forest:

1. Final Grade / GPA
2. Number of Absences
3. Mother's Education Level
4. Number of Past Failures
5. Weekly Study Time

---

6. Model Deployment

- **Saved using `joblib`:**
- **Cleaned dataset saved as:** `cleaned_student_data.csv`

7. Stakeholder Problems & Data-Driven Solutions

Stakeholder	Problem	Data Insight	Recommendation
School Admins	High dropout rates	15.41% overall dropout	Implement targeted retention programs
Teachers	Can't identify at-risk students	Male students, fewer study hours, more absences	Use dashboards and predictive model alerts
Parents	Unaware of influence	Mother's low education links to dropouts	Parent workshops, increase communication
Counselors	Need behavioral flags	High absences, non-participation, low grades	Monitor and flag high-risk profiles

---

8. Key Insights (SQL-Based)

1. Dropout Rate: 15.41% of students dropped out
2. Gender Impact: Males 18.8%, Females 13.05%
3. Absenteeism: Dropout rate rises after 8+ absences
4. Extracurriculars: Non-participants had 17.07% dropout
5. Mother's Education: Basic education students = 25.87% dropout
6. Study Time: <2 hours/week = 23.58% dropout
7. Final Grade: <10 = 100% dropout

---

9. Dashboards (Power BI)

Dashboard 1: Demographics

- Pie Chart: Gender Distribution
- Bar Chart: Father's Education vs. Dropouts
- KPI Card: Dropout Rate
- Waterfall Chart: Dropout by Gender

Dashboard 2: Behavioral Factors

- Stacked Column: Absences vs. Dropout Rate
- Stacked Column: Family Support vs. Dropout Rate
- Stacked Column: Internet Access vs. Dropout Rate
- Stacked Bar: Relationships vs. Dropout
- Stacked Bar: Extra Curricular Participation vs. Dropout
- Stacked Bar: School Support vs. Dropout Rate

Dashboard 3: Predictive Modeling

- Bar Chart: Feature Importance
- Heatmap: Confusion Matrix
- KPI Cards: Accuracy, Precision, Recall

---

10. SMART Recommendations

Goal	Specific	Measurable	Achievable	Relevant	Time-Bound
Reduce Dropouts	Target at-risk students	Monitor dropout quarterly	Via model alerts	Supports retention objectives	Reduce by 5% in 6 months

---

11. Conclusion & Recommendations

Key Actions:

- Use predictive model alerts to flag at-risk students
- Intervene early with male and low-performing profiles
- Promote extracurricular participation
- Engage parents with low education levels
- Enforce attendance monitoring policies