Technical Report

☑ Abstract

This project aims to predict student academic success using machine learning techniques. The dataset includes demographic, social, and academic features that influence student performance. Several supervised learning models—Logistic Regression, Decision Tree, and Random Forest—were trained and evaluated under both data-leakage and no-leakage scenarios. The results show that Random Forest without leakage provides the best balance between accuracy, recall, and F1-score, while models with grade information (leakage) achieve higher but unrealistic performance. Insights from feature importance suggest that failures, absences, and parental involvement strongly impact student outcomes. These findings can support schools in early interventions, personalized tutoring, and attendance monitoring to improve overall success rates.

☑ Problem & Value

- **Problem:** Many students are at risk of failing or dropping out due to poor academic performance, absenteeism, or lack of support. Teachers and schools often lack predictive tools to identify these students early, which delays effective interventions.
- **Value:** By building predictive models on student data, schools can:
  - Detect at-risk students before final exams.
  - Target interventions such as tutoring, counseling, or attendance monitoring.
  - Allocate resources more effectively to students who need the most support.
  - Improve overall academic performance and reduce dropout rates.

Dataset (Source, Schema, and Limits)

Source:

The dataset is the **Student Performance Dataset** from the **UCI Machine Learning Repository** (ID = 320). It contains data collected from two Portuguese secondary schools (Math and Portuguese courses).

Schema (Features):

The dataset includes **student demographic, social, and academic attributes**:

- **Demographics & Background:**
  - `school` (binary: GP = Gabriel Pereira, MS = Mousinho da Silveira)
  - `sex` (binary: M/F)
  - `age` (numeric: 15–22)
  - `address` (binary: U = urban, R = rural)
  - `famsize` (binary: LE3 = ≤3, GT3 = >3)
  - `Pstatus` (binary: T = living together, A = apart)
  - `Medu`, `Fedu` (numeric: parental education, 0–4)

    - `Mjob`, `Fjob` (categorical: teacher, health, services, at_home, other)
    - `guardian` (categorical: mother, father, other)
    - `reason` (categorical: home, reputation, course preference, other)
- **Study & Support:**
    - `studytime` (1–4: weekly study hours)
    - `traveltime` (1–4: home–school travel time)
    - `failures` (numeric: past class failures)
    - `schoolsup`, `famsup`, `paid` (binary: yes/no, support programs)
    - `activities`, `nursery`, `higher`, `internet`, `romantic` (binary: yes/no indicators)
- **Lifestyle & Behavior:**
    - `famrel` (family relationship quality, 1–5)
    - `freetime` (1–5: free time after school)
    - `goout` (1–5: going out with friends)
    - `Dalc`, `Walc` (1–5: weekday/weekend alcohol use)
    - `health` (1–5: self-reported health)
    - `absences`, `absences_capped` (school absences, raw and capped)

- **Performance Indicators:**
  - `G1`, `G2`, `G3` (first, second, final period grades, 0–20 scale)

Target Variables (defined for classification):

- **Binary target:** `pass` (1 if final grade ≥ 10, 0 otherwise)
- **Alternative target:** `risk` (3-class risk grouping: low, medium, high, based on failures/grades)

Limits:

- **Size:** ~650 records (small dataset → limited generalization).
- **Geographic bias:** Only from Portuguese schools, not globally representative.
- **Imbalance:** Majority of students pass, fewer fail (class imbalance issue).
- **Data leakage risk:** Including G1 and G2 (early grades) strongly predicts G3 (final grade).

- **Sensitive variables:** Some features (e.g., romantic relationships, alcohol consumption) may raise ethical/privacy concerns.

Methods

We followed a structured approach to analyze the dataset.

1. **Data Cleaning & Preprocessing**: We handled missing values, converted categorical variables into numerical form (one-hot encoding), and engineered new features such as `pass/fail`, `risk`, and `absences_capped`.
2. **Exploratory Data Analysis (EDA)**: We explored distributions, correlations, and patterns (e.g., absences vs. grades, failures vs. outcomes).
3. **Clustering**: We applied K-Means to group students into behavioral/risk clusters.

4. **Supervised Learning**: We trained three algorithms—Logistic Regression, Decision Tree, and Random Forest—using two setups: with data leakage (including G1, G2) and without leakage (excluding them).
5. **Evaluation**: We used both hold-out validation and 5-fold cross-validation. Hyperparameter tuning was conducted to optimize performance.

Results

- With leakage, models achieved **very high performance** (Logistic Regression F1 ≈ 0.95, Random Forest F1 ≈ 0.96).
- Without leakage, performance dropped but remained strong (Random Forest F1 ≈ 0.90).
- Random Forest consistently outperformed other models, especially in recall (identifying at-risk students).

- Logistic Regression provided interpretable coefficients, while Decision Tree offered simple rules but slightly lower accuracy.
- ROC-AUC scores confirmed that models with leakage were artificially strong, while no-leakage results reflected realistic generalization.

Ethics

- **Privacy**: Student data must remain anonymized, as it includes sensitive educational and family background details.
- **Fairness**: Predictions should not discriminate by gender, socioeconomic status, or family situation.
- **Responsible Use**: Models should guide **supportive interventions** (tutoring, mentoring), not punitive actions.
- **Transparency**: Educators and students should understand the model's purpose and limitations.

Recommendations

1. Students with **high absences and ≥2 failures** are at the greatest risk → recommend early tutoring + attendance monitoring.
2. **Low study time** is linked with poor performance → provide study skills workshops.
3. **Parental education level** strongly correlates with outcomes → schools could involve families in academic planning.
4. **Weekend alcohol use** negatively impacts grades → awareness and counseling programs should be considered.
5. Students with **strong family relationships** perform better → encourage family-school partnerships.
6. **Internet access** is a positive factor → ensure digital resources are available to all students.

Limitations

- The dataset comes from Portugal only, so generalization to other countries is limited.
- Some features are **self-reported** (study time, alcohol use), which may be biased or inaccurate.
- The dataset size (~650 students) is relatively small, which limits robustness.
- Including G1/G2 grades leads to **data leakage**, inflating model accuracy unrealistically. For real-world use, these should be excluded.