# Machine Learning Term Project

**Term Project**: Smart Student Performance Prediction System

**Deadline: 31/12/2025**

**Objective:**

The goal of this project is to build a predictive analytics system that estimates a student's final academic performance using historical, behavioral, and demographic data. You will:

1- Preprocess & analyze a 20,000-sample educational dataset.
2- Build classification and regression models for performance prediction.
3- Perform model selection, hyperparameter tuning, and error analysis.
4- Compare the interpretability and results across classical ML models.

This project simulates a real educational analytics pipeline for university use.

**Dataset:** Term_Project_Dataset_20K

## *Dataset Description*

A dataset of **20,000 rows** and **38 features** (excluding the targets)

### Target Variables

- final_score (0–100, regression target)
- final_grade (A / B / C / D / F, classification target)
- pass_fail (binary: pass/fail)

## Feature Categories

### A. Demographic Features (7 features)

- Age
- Gender
- Parent_income
- Num_siblings
- family_support (0–5 scale)
- Commute_time_min
- part_time_job (yes/no)

### B. Academic History (10 features)

- Previous_gpa
- Num_failed_courses
- High_school_grade
- Math_background_score
- Language_background_score
- Science_background_score
- Prior_semester_credits
- Study_hours_last_semester
- Past_attendance_rate
- Academic_warnings_count

### C. Behavioral & Engagement Data (10 features)

- Lecture_attendance_rate
- Assignment_submission_rate
- Quiz_avg_score
- Midterm_score
- Lab_participation_rate
- Online_portal_usage_minutes
- Group_project_activity
- Library_visits_per_month
- Discussion_forum_posts
- Lateness_count

### D. Psychological / Self-Report Factors (6 features)

- stress_level (0–10)
- Sleep_hours
- motivation_level (0–10)
- Study_time_per_week
- concentration_level (0–10)

- exam_anxiety_level (0–10)

### E. Institutional Data (5 features)

- Course_difficulty_rating
- Teacher_experience_years
- Class_size
- Num_prerequisites
- course_type (mandatory/elective)

## ***Requirements***

### A. Data Understanding & Exploration

- Perform descriptive statistics
- Visualize distribution of scores
- Analyze correlations
- Identify noisy or irrelevant features

### B. Data Preprocessing

- Handle missing values
- Identify and remove outliers
- Encode categorical features
- Feature scaling (MinMax or StandardScaler)
- Train/validation/test split
- Address class imbalance (SMOTE, undersampling, oversampling)

### C. Model Development

- You must train at least 5 classical ML models, such as:
    - Logistic Regression
    - k-Nearest Neighbors
    - Random Forest
    - Gradient Boosting / XGBoost
    - Naïve Bayes
    - Support Vector Machine (SVM)
    - Decision Tree
    - Linear Regression / Lasso / Ridge
    - …..

### D. Model Evaluation

- Classification metrics:
    - Accuracy
    - Precision, Recall, F1

    - Confusion matrix
    - ROC curve + AUC
- Regression metrics:
    - MSE
    - MAE
    - $R^2$
    - Residual error plots

### E. Conclusion

- What factors most influence performance?
- Which features are the most important ones?

## ***Project Deliverables***

### A. Technical Report (PDF)

- Abstract
- Dataset description
- EDA (Exploratory Data Analysis)
- Data preprocessing steps
- Model design & justification
- Results, plots, metrics
- Discussion & error analysis
- Conclusion
- Appendix (Code Snippets)

### B. Python Source Code

- Google Colab or Jupyter Notebook or Python scripts
- Clean, modular, documented

### C. Presentation Slides

- Approach
- Models
- Results
- Interpretability
- Final conclusions