# Phase-2 Submission Template

## Phase-2 Submission Template

## Student Name

Abdul Rahman M.S

## Register Number

422223243002

## Institution

SURYA GROUP OF INSTITUTIONS

## Department

B.TECH ARTIFICIAL INTELLIGENCE & DATA SCIENCE

## Date of Submission

22/04/2025

## Github Repository Link

<https://github.com/steeveabdul/Abdul-Rahman>

## 1. Problem Statement

Topic: Predicting Student Performance using Machine Learning

This project aims to predict whether a student will pass or fail based on their academic and socio-demographic features. The problem is a binary classification task where the target variable is the final student result (pass/fail). The problem is important for early identification of students at risk, enabling timely intervention and academic support.

## 2. Project Objectives

The goal of this project is to build a classification model that predicts student performance accurately.

- Improve model accuracy through preprocessing and feature engineering.

- Ensure interpretability of the results.

- Apply the model to identify patterns influencing academic success.

## 3. Flowchart of the Project Workflow

1. Data Collection

2. Data Preprocessing

3. Exploratory Data Analysis (EDA)

4. Feature Engineering

5. Model Building

6. Model Evaluation

7. Conclusion and Insights

## 4. Data Description

Dataset Name: Student Performance Dataset

Source: UCI Machine Learning Repository / Kaggle

Type: Structured CSV file

Records: 1044 rows

Features: 33 columns

Static Dataset

Target Variable: Final Result (Pass/Fail)

## 5. Data Preprocessing

- Missing values handled using mean/mode imputation

- Duplicate records were removed

- Outliers detected using IQR method and capped

- Data types were corrected (e.g., categorical to category type)

- Label encoding used for binary categoricals, One-hot encoding for nominal data

- Features standardized using MinMaxScaler

## 6. Exploratory Data Analysis (EDA)

- Univariate analysis showed that absences and study time had significant variation

- Bivariate analysis: Higher study time correlated with better grades

- Parental education level and previous failures influenced outcomes

- Visualizations: Histograms, box plots, pairplots, correlation heatmap

- Key Insight: Students with higher family support and study time tend to perform better

## 7. Feature Engineering

- Created a new feature: total support (combining family and school support)

- Extracted week vs weekend study habits

- Applied polynomial features on 'study time'

- Feature selection using SelectKBest to keep top predictors

- PCA used to reduce dimensionality, retaining 95% variance

## 8. Model Building

- Models Used: Logistic Regression and Random Forest Classifier

- Train-Test Split: 80% training, 20% testing with stratification

- Logistic Regression: Accuracy = 78%, F1-Score = 0.75

- Random Forest: Accuracy = 85%, F1-Score = 0.82

- Random Forest performed better due to handling of non-linearity and feature interactions

## 9. Visualization of Results & Model Insights

- Confusion Matrix: Showed better true positive rate in Random Forest

- ROC Curve: AUC = 0.91 for Random Forest

- Feature Importance: 'Study time', 'Failures', 'Parental Education' ranked high

- Residual analysis showed well-distributed errors, indicating a good fit

## 10. Tools and Technologies Used

- Programming Language: Python

- IDE: Google Colab

- Libraries: pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost

- Visualization: Plotly and seaborn for visual exploration

## 11. Team Members and Contributions

- ABDUL RAHMAN.M S: Data Preprocessing and Cleaning

- AJAY.K: Model Development and Evaluation

- BALA SUNDAR.R: EDA and Feature Engineering

- EZHILARASAN.K: Documentation and Visualization