**UCI Student Performance Dataset: Project Requirements Document**

**Dataset Overview**

The UCI Student Performance dataset includes information on students’ academic, social, and demographic details, aiming to predict the students' final grade based on these factors. The goal of this project is to implement data cleanup, exploratory data analysis (EDA), hypothesis testing, and machine learning modeling to predict student performance.

**Dataset Description**

**Features Overview:**

1. **School** - The school the student is attending (GP for Gabriel Pereira, MS for Mousinho da Silveira).
2. **Sex** - Gender of the student (M for male, F for female).
3. **Age** - Age of the student in years.
4. **Address** - Student’s residence type (U for urban, R for rural).
5. **Family Size** - Family size (LE3 for less than or equal to 3, GT3 for greater than 3).
6. **Parent Status** - Parents' cohabitation status (T for together, A for apart).
7. **Mother's Education** - Mother’s education level (0 = none, 4 = higher).
8. **Father's Education** - Father’s education level (0 = none, 4 = higher).
9. **Mother's Job** - Mother’s job type (e.g., teacher, health, services).
10. **Father's Job** - Father’s job type (e.g., teacher, health, services).
11. **Reason** - Reason for choosing the school (e.g., course, reputation).
12. **Guardian** - Student's primary guardian (e.g., mother, father).
13. **Travel Time** - Time to reach school (1 = <15 min, 4 = >1 hr).
14. **Study Time** - Weekly study time (1 = <2 hours, 4 = >10 hours).
15. **Failures** - Number of past class failures.
16. **School Support** - Extra educational support (yes or no).
17. **Family Support** - Family support for studies (yes or no).
18. **Paid Classes** - Extra paid classes in subject (yes or no).
19. **Activities** - Participation in extracurricular activities (yes or no).
20. **Nursery** - Attended nursery school (yes or no).
21. **Higher Education** - Aspires for higher education (yes or no).
22. **Internet** - Access to internet at home (yes or no).
23. **Romantic** - In a romantic relationship (yes or no).
24. **Family Relations** - Quality of family relationships (1 = very bad, 5 = excellent).
25. **Free Time** - Amount of free time after school (1 = very low, 5 = very high).
26. **Go Out** - Frequency of going out with friends (1 = very low, 5 = very high).
27. **Workday Alcohol** - Workday alcohol consumption (1 = very low, 5 = very high).
28. **Weekend Alcohol** - Weekend alcohol consumption (1 = very low, 5 = very high).
29. **Health** - Current health status (1 = very bad, 5 = very good).
30. **Absences** - Number of school absences.
31. **Grades** - G1, G2, G3 (first, second, and final grades).

**Project Objectives**

1. **Data Preparation and Cleanup**
   * Handle missing values and choose appropriate imputation strategies.
   * Convert categorical variables into numerical values where necessary.
   * Address outliers, especially in the target variable (G3).
2. **Exploratory Data Analysis (EDA)**
   * Generate descriptive statistics for each feature.
   * Visualize key features to understand distributions (histograms, box plots).
   * Create correlation matrices to identify relationships between features and the final grade (G3).
3. **Hypothesis Testing**
   * Perform statistical hypothesis tests to investigate:
     + Whether there is a significant difference in G3 based on School or Gender (use t-tests or ANOVA).
     + The correlation between Study Time and Absences with G3 using correlation tests.
     + The association between Family Size and Family Relations using a chi-square test.
4. **Machine Learning Modeling**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Build and evaluate multiple machine learning models to predict G3 (final grade):
     + **Linear Regression** – Set a baseline for grade prediction.
     + **Random Forest Regressor** – Use an ensemble model and fine-tune hyperparameters.
   * Evaluate models using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared metrics.
   * Choose the best-performing model and interpret the results, do a cross validation if necessary
5. **Model Interpretation and Insights**
   * Examine feature importance to determine which attributes impact the final grade most.
   * Create partial dependence plots for the top influential features on the predicted grade.
   * Summarize findings and insights in a report format.
6. **Python API using Flask**
   * Develop a simple Flask REST API server for predicting the values
   * Use POSTMAN or similar technique to test it

**Project Timeline**

**Total Duration**: 4 hours

| **Task** | **Description** | **Estimated Time** |
| --- | --- | --- |
| **Data Preparation and Cleanup** | Clean missing data, handle data types and outliers | 45 minutes |
| **Exploratory Data Analysis** | Visualize and analyze feature relationships | 45 minutes |
| **Hypothesis Testing** | Conduct statistical tests on selected features | 30 minutes |
| **Machine Learning Modeling** | Build, train, and evaluate ML models | 1 hour |
| **Model Interpretation and Insights** | Analyze model importance and create final report | 30 minutes |
| **API Development** | Python FLASK API | 30 mins |

**Submission Requirements**

1. **Python code and Jupyter Notebook** with all code, visualizations, and explanations.
2. **Summary Report** detailing:
   * Key findings from EDA and hypothesis tests.
   * Comparison of model evaluation metrics.
   * Interpretation of the best-performing model and significant features.
   * Presentation for 10 minutes