**Machine Learning Test**

**Instructions:**

1. This test consists of various sections, including data exploration, preprocessing, regression models, logistic regression, and advanced concepts.
2. Ensure that your internet connection is turned off during the entire duration of the test.
3. You were provided with libraries.txt. Make sure all the libraries listed in this file are installed before beginning the test.
4. You are required to use Jupyter Notebook for coding and analysis.
5. Make sure to save your work.
6. Install the necessary libraries (scikit-learn, pandas, NumPy, matplotlib, seaborn) before starting the test.
7. Write clear and concise code, and provide explanations where necessary.
8. Each section clearly outlines the requirements. If need be, the instructor will guide you.
9. **ALL THE SECTIONS WILL USE TRAIN.CSV EXCEPT FOR SECTION 03 PARTS 02 AND 03 IN WHICH YOU SHOULD USE TEST.CSV AS WELL.**
10. **SUBMISSIONS SHOULD CONTAIN IPYNB AND PDF FILE.**
11. **EACH CELL MUST BE EXECUTED BEFORE SUBMISSION.**
12. **FAILURE TO FOLLOW THE INSTRUCTIONS CAN AND WILL CAUSE NEGATIVE MARKINGS.**

**Section 1: Data Exploration (Total: 20 points)**

1. Import Libraries (2 points):

* Import the required libraries: pandas, NumPy, matplotlib, and seaborn.

1. Load Dataset (4 points):

* Load the provided train.csv dataset.
* Display the first few rows of the dataset.

1. Explore Features (6 points):

* Describe the key features of the dataset.
* Identify any potential challenges or patterns in the data.

1. Handle Missing Data (8 points):

* Identify and handle missing data appropriately.
* Provide a summary of how you dealt with missing values.

**Section 2: Data Preprocessing (Total: 20 points)**

1. Feature Engineering (8 points):

* Apply at least two feature engineering techniques to enhance the dataset.
* Justify your choice of feature engineering techniques.

1. Encode Categorical Variables (6 points):

* Encode categorical variables using suitable methods.
* Discuss the impact of encoding on model performance.

1. Split Data (6 points):

* Split the data into training and testing sets.
* Specify the ratio used for the split and explain your choice.

**Section 3: Regression Models (Total: 40 points)**

1. Linear Regression (10 points):

* Implement a linear regression model.
* Evaluate the performance using appropriate metrics.
* Provide insights into the model's predictions.

1. Advanced Regression (15 points):

* Implement a more advanced regression algorithm (Ridge, Lasso, or ElasticNet).
* Tune hyperparameters if necessary.
* Compare the performance with linear regression.

1. Ensemble Method (15 points):

* Implement a Random Forest regression model.
* Discuss the advantages of using an ensemble method in this context.
* Evaluate and compare the model's performance with previous models.

**Section 4: Logistic Regression (Total: 20 points)**

1. Transform into Classification Task (6 points):

* Transform the problem into a binary classification task.
* Define the target variable and threshold for classification.

1. Logistic Regression (8 points):

* Implement a logistic regression model.
* Evaluate the classification performance using appropriate metrics.

1. Discussion (6 points):

* Discuss the implications of transforming a regression problem into a classification problem.
* Provide insights into the challenges and advantages of the classification approach.

**Section 5: Advanced Concepts (Total: 40 points)**

1. Regularization (15 points):

* Apply regularization techniques (L1, L2) to one of the regression models.
* Evaluate the impact on model performance and interpret the results.

1. Cross-Validation and Hyperparameter Tuning (15 points):

* Implement cross-validation for one of the models.
* Perform hyperparameter tuning and discuss the impact on model performance.

1. Bias-Variance Tradeoff (10 points):

* Discuss the concept of bias-variance tradeoff in the context of your models.
* Provide recommendations for reducing bias or variance.