Machine Learning

Titanic Dataset

- **Instructions:**
  - Use the **Titanic dataset** (train.csv from Kaggle: https://www.kaggle.com/c/titanic/data)
  - Answer the following questions by applying appropriate techniques in Python (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, etc.)
  - Provide code, output, and explanations for each question.
- **The Role of Statistics in Machine Learning**
  - Explain the role of statistics in Machine Learning with examples from the Titanic dataset.
  - Compute basic statistical measures (mean, median, mode, variance, standard deviation) for numeric columns like Age and Fare.
  - What insights can be drawn from these statistics?
- **2. Population vs Sample**
  - Define and explain the difference between population and sample.
  - Extract a sample (30% of the dataset) and compare its statistical properties with the full dataset.
- **3. Sampling Data Techniques**
  - Perform the following sampling techniques and explain:
    - Random Sampling
    - Stratified Sampling (based on Survived column)
    - Systematic Sampling
  - Compare the distributions of sampled data with the original dataset.
- **4. Correlation Matrix and Heatmap**
  - Compute the correlation matrix for numeric features.
  - Visualize the correlation matrix using a heatmap.
  - Identify which features are highly correlated and explain their impact on model performance.
- **5. Pre-processing**
  - Identify categorical and numerical columns in the dataset.
  - Perform encoding on categorical columns (Sex, Embarked).
  - Scale numeric features using StandardScaler or MinMaxScaler.

- **6. Feature Engineering**
    - Create new meaningful features, such as:
        - FamilySize = SibSp + Parch
        - Title extraction from Name
        - IsAlone (indicating if the passenger is alone or not)
    - Explain how these features can improve model performance.
- **7. Feature Transformation**
    - Apply log transformation on Fare to reduce skewness.
    - Normalize or standardize numerical columns.
- **8. Binning and Binarization**
    - Perform binning on Age into categories (child, teenager, adult, senior).
    - Convert Fare into bins (low, medium, high).
    - Convert Survived into binary values if necessary.
- **9. Handling Mixed-Type Values**
    - Identify any columns with mixed data types.
    - Convert mixed-type columns into a uniform format.
- **10. Handling DateTime**
    - If DateTime data was available (e.g., travel date), how would you extract useful features such as:
        - Day of the week
        - Month
        - Is weekend or not
- **11. Feature Construction**
    - Create interaction features, such as combining Pclass and Fare.
    - Generate polynomial features and assess their usefulness.
- **12. Handling Missing Data**
    - Identify missing values in the dataset.
    - Apply different techniques to handle missing values:
        - Mean/Median/Mode Imputation
        - KNN Imputation
        - MICE
    - Compare how different imputation methods affect model training.
- **13. Handling Outliers**
    - Detect outliers in Age and Fare using:
        - Boxplot
        - Z-score method

    - IQR method
  - Remove or cap outliers and compare results.
- **14. Convert Non-Normally Distributed Column into Normal Distribution**
  - Identify non-normally distributed columns.
  - Apply transformations (log, Box-Cox, Yeo-Johnson) to make the distribution normal.
  - Compare before and after distributions using histograms.
- **15. Principal Component Analysis (PCA)**
  - Apply PCA on numeric features after scaling.
  - Determine the optimal number of components based on explained variance.
  - Visualize PCA components.
  - Interpret PCA results and discuss their impact on dimensionality reduction.

Note:

1. Dummy data can be generated as needed for the questions.
2. Chatbot assistance (ChatGPT) is available for support, but final solutions should be provided independently.

Submission Date: **20-3-25**

Note:

- **Assignment Submitted in PDF form.**
- **In PDF form, you must include the screenshot of the code and its output.**
- **If not mentioned, then write a Python script that includes the code for each task.**
- **Include comments in your code to explain the purpose and functionality of each step.**
- **Send it to this email id: jtechsolution93@gmail.com**
- **You can get help from ChatGPT or any Chatbot but at the end.**

- **If Any, then It is highly recommended that you read research papers for assignment.**

  Helping Websites for research papers are:

  https://scholar.google.com/

  https://sci-hub.hkvisa.net/