Name:Vaibhavi Vinayak Patki

Documentation for Project 1

Approach and Methodologies:

Data Science Task (Classification):

1. **Algorithm Selection:**
   - Chose Logistic Regression as the classification algorithm due to its simplicity, interpretability, and effectiveness for multiclass classification tasks.
   - Logistic Regression is suitable for the Iris dataset, providing a good balance between model complexity and interpretability.
2. **Feature Selection:**
   - Utilized all four features (sepal length, sepal width, petal length, petal width) from the Iris dataset for classification.
   - These features are standard in the Iris dataset and are known to be relevant for distinguishing between different species.
3. **Evaluation Metrics:**
   - Selected accuracy, precision, recall, f1 score as evaluation metrics.
   - Accuracy provides an overall measure of model performance.
   - Precision and recall are crucial in a multiclass setting, offering insights into class-specific performance.
4. **Model Training and Evaluation:**
   - Split the dataset into training (80%) and testing (20%) sets to train and evaluate the model's generalization performance.
   - Trained the Logistic Regression model on the training set.
   - Evaluated the model on the testing set using accuracy, precision, recall, confusion matrix, and classification report.

Exploratory Data Analysis (EDA):

1. **Visualization Techniques:**
    - Utilized histograms to visualize feature distributions, providing insights into the spread and central tendencies of data.
    - Employed box plots to identify potential outliers and summarize the distribution of features.
    - Utilized scatter to visualize relationships between features and species.
2. **Libraries:**
    - Relied on Pandas for data manipulation and summary statistics.
    - Used Matplotlib for basic visualizations such as histograms and box plots.
    - Incorporated Seaborn for enhanced visualizations.

Challenges Faced:

1. **Visualization Variety:**
    - While Seaborn enhances visualization, achieving diverse visualizations, especially with more intricate styles, may still require additional effort.
2. **Model Choice:**
    - Logistic Regression is suitable for this task, but exploring other algorithms like Decision Trees or Random Forests might be necessary for more complex datasets. The choice of model depends on the trade-off between model complexity and dataset size.