Machine Learning

Description Document

General Information on dataset:

1. **Image Dataset:**

    - **Dataset Name**: UTKFace
    - **Number of Classes and Their Labels**:
        - **Class 0**: Ages 0-20 years (Youth)
        - **Class 1**: Ages 21-40 years (Young Adult)
        - **Class 2**: Ages 41-60 years (Middle-Aged Adult)
        - **Class 3**: Ages 61-80 years (Senior Adult)
        - **Class 4**: Ages 81-116 years (Elderly)
    - **Total Number of Samples**: 24,108 images.
    - **Image Size**: Each image is resized to 200x200 pixels**.**
    - **Data Split**:
        - **Training Set**: 75% of the dataset (~18,000 images).
        - **Testing Set**: 25% of the dataset (~6,000 images).

2. **Numerical Dataset:**

    - **Dataset Name**: Life Expectancy
    - **Number of Features/Columns**: 22 columns.
    - **Total Number of Samples**: 2,938 rows
    - **Total Number of Missing Values**: 14 columns
    - **Data Split**:
        - **Training Set**: 70% of the dataset (~2,056 rows).
        - **Testing Set**: 30% of the dataset (~881 rows).

Applied Algorithms:

1. **Image Dataset:**
    i. Logistic Regression: Predict the age class of individuals based on image data. images are normalized and dimensionality reduced using PCA**,** Logistic Regression is applied after normalization and dimensionality reduction.
    ii. **KNN Regression**: is used to classify images into 5 age classes, images resized to 28x28 pixels and flattened into 1D arrays, images are normalized and dimensionality reduced using PCA, a KNN classifier with optimal hyperparameters (e.g., k=81) is trained

2. **Numerical Dataset:**
    i. **Linear Regression**: to estimate life expectancy based on a dataset containing 2938 samples and 21 features. This algorithm establishes a linear relationship between the independent variables (features) and the dependent variable (Life Expectancy).
    ii. **KNN Regression**: implemented to predict the target variable (Life Expectancy), based on various health, economic, and demographic indicators from the dataset.

Comparative Analysis:

1. **Image Dataset:**

	Logistic Regression	KNN
Performance	0.5468724074995852	0.5503567280570765
Log Loss	1.153231177033976	1.2114021542510467

- KNN outperforms Logistic Regression on all metrics:
    - ❖ Higher Accuracy indicates better classification performance across all 5 age classes.
    - ❖ Higher ROC-AUC values for all classes show KNN's ability to differentiate between the age groups effectively.
    - ❖ Lower Log Loss indicates better confidence in predictions for KNN compared to Logistic Regression.

Thus, KNN is the better model for predicting and classifying age in this case.

2. **Numerical Dataset:**

	KNN	Linear Regression
Mean Absolute Error (MAE)	1.9036377468412986	2.9780594691160074
Mean Squared Error (MSE)	8.818238974163677	15.859259203093796
Root Mean Squared Error (RMSE)	2.969551982061213	3.98236854184715
R-squared (R^2)	0.9011764307313425	0.8222696612103129

- KNN outperforms Linear Regression on all metrics:
    - ❖ Lower error values (MAE, MSE, RMSE) indicate better predictions.
    - ❖ A higher R^2 value (closer to 1) shows KNN better explains the variance in the data.

Thus, KNN is the better model for predicting life expectancy in this case.