**Faculty of computers and artificial intelligence**

**Course name: AI330-ML-Projects\_Fall2023**

|  |  |
| --- | --- |
| **names** | **id** |
| **مروان شامل محمد فكرى** | **20210895** |
| **مريم سامى فاروق** | **20210913** |
| **مريم فتحى احمد** | **20210921** |
| **مروان مجدى محمد** | **20210898** |
| **عمرو بهيج عبد العزيز** | **20210635** |
| **عبد الله احمد عوض** | **20210540** |

***ML project***

***Numerical dataset :***

**a. General Information on Dataset:**

**Dataset Information:**

**Name of Dataset:** Insurance Dataset

**Size of Each Feature (X):**

'age': Age of the insured

'sex': Gender of the insured (encoded as 0 or 1)

'bmi': Body Mass Index of the insured

'children': Number of children/dependents covered by the insurance

'smoker': Smoking status of the insured (encoded as 0 or 1)

'region': Region of the insured (encoded as integers)

Size of Target Variable (y):

'charges': Insurance charges

**Data Split:**

The dataset is split into features (X) and the target variable (y).

80% of the data is used for training, and 20% is used for testing.

**b. Implementation Details:**

**Feature Extraction:**

**Features Extracted:**

For feature 'sex,' 'smoker,' and 'region,' label encoding is applied.

**Cross-Validation:**

Cross-validation is used during the model evaluation phase.

Number of folds: 10

Models and Hyperparameters:

**1. Linear Regression:**

Features Used:

'age', 'sex', 'bmi', 'children', 'smoker', 'region'

Hyperparameters:

No explicit hyperparameters mentioned in the code.

Evaluation Metrics:

Mean Squared Error (MSE)

R-squared (R2)

**2. K-Nearest Neighbors Regression:**

Features Used:

'age', 'sex', 'bmi', 'children', 'smoker', 'region'

**Hyperparameters:**

Number of Neighbors: 5

**Evaluation Metrics:**

Mean Squared Error (MSE)

R-squared (R2)

Mean Absolute Error (MAE)

Model Evaluation and Plots:

The models are evaluated using cross-validation and various regression metrics.

Comparison plots are created between actual and predicted values for Linear Regression and K-Nearest Neighbors.

Accuracy scores for Linear Regression and K-Nearest Neighbors are provided.

Receiver Operating Characteristic (ROC) curve is plotted for Linear Regression (threshold-based classification).

Additional Information:

**Data Preprocessing:**

**Imputation: Simple** mean imputation for handling missing values.

**Standard Scaling:** StandardScaler is applied to features and target variables.

**Visualization:**

Scatter plots for linear regression predictions.

Comparison plots for Linear Regression and K-Nearest Neighbors.

Loss Curve for K-Nearest Neighbors.

ROC Curve:

ROC curve is plotted for Linear Regression predictions.

=====================================================================

***Image dataset :***

**a. General Information on Dataset:**

**Name of Dataset:** Fashion MNIST

**Number of Classes and Labels:** There are 10 classes in Fashion MNIST , used 3 of them

**Total Number of Samples:** Fashion MNIST typically contains 60,000 training samples and 10,000 testing samples.

**Size of Each Image**: The size of each image is 28x28 pixels.

**Number of Samples Used:**

**Training Set:** A subset of the original training set is used based on the specified class labels and further split into training and validation sets.

**Testing Set:** A subset of the original testing set is used based on the specified class labels.

**b. Implementation Details:**

Feature Extraction Phase:

Number of Features Extracted: The number of features extracted is 3 of 10 but HOG (Histogram of Oriented Gradients) features are extracted for each image.

Names of Features: HOG features.

Dimension of Resulted Features: The dimension of the HOG features is not explicitly mentioned, but it's typically a one-dimensional array.

Cross-Validation:

Cross-validation is not explicitly used in this code. The dataset is split into training and testing sets.

**Visualization:**

Roc\_curve  
confussion matrix