بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Report

Team: 11:00   TA: *Ahmed Ashraf*

Project Description Document:

General Information on dataset:

- Numeric Dataset

  Name: Cardiovascular Disease

  - aim is to predict the presence or absence of cardiovascular disease based on various patient metrics.

  - It has 17 columns and 68204 rows

  - The data types in it is float64(2), int64(13), object(2)

- Image Dataset

  Name: Fruits 360

  - A dataset of images containing fruits and vegetables

  - The total number of images: 90483.

  - Training set size: 67692 images (one fruit or vegetable per image).

  - Test set size: 22688 images (one fruit or vegetable per image).

  - The number of classes: 131 (fruits and vegetables).

- Image size: 100x100 pixels.

Implementation details:

- ***1. in Logistic Regression:***

1. Number of Features:

- For a single image: 3 * dim * dim (assuming RGB format)

- For the entire dataset: 3 * dim * dim

2. Feature Names:

Each pixel position in the flattened image array corresponds to a specific feature.

3. Dimensionality of resulted features:

-For a single image: 3 * dim * dim

-For the entire dataset: 3 * dim * dim

- ***2. in K-Means:***

1. Number of Features:

After applying PCA, the number of features is equal to the number of principal components selected. num_components = 30, so the number of features is 30.

2.Feature Names:

After PCA, the features are the principal components, and they don't have specific names related to the original pixel values. They are often referred to as PC1, PC2, ..., PCn, where n is the number of components.

3. Dimensionality of resulted features: 30

- ***3. in Numeric Dataset:***

1. Number of Features: 15

2. Feature Names: (id   age   gender   height   weight   ap_hi ap_lo cholesterol gluc   smoke   alco   active cardio age_years   bmi   bp_category   bp_category_encoded)

3. Dimensionality of resulted features: 13

- **Cross Validation is <u>not </u>used in any of our models.**

- **Used Hyperparameters:**

  1. in KNN: number of neighbors (n_neighbors).

  2. in K-means: number of clusters (num_clusters).

Results details:

- *Linear Regression :*

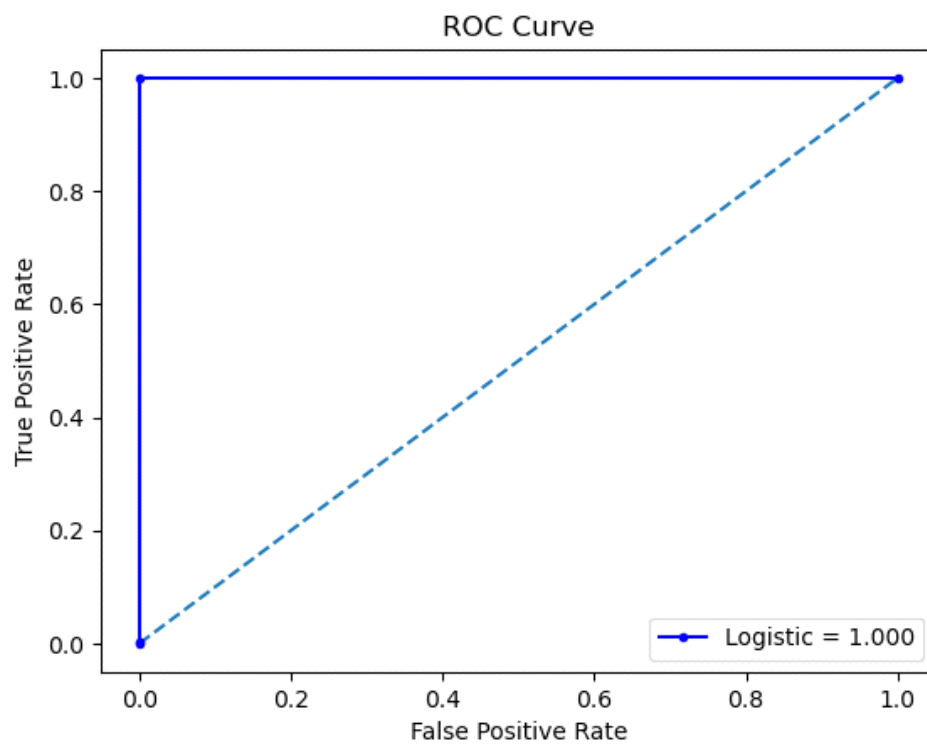True vs Predicted Digits values in Multiple Linear Regression
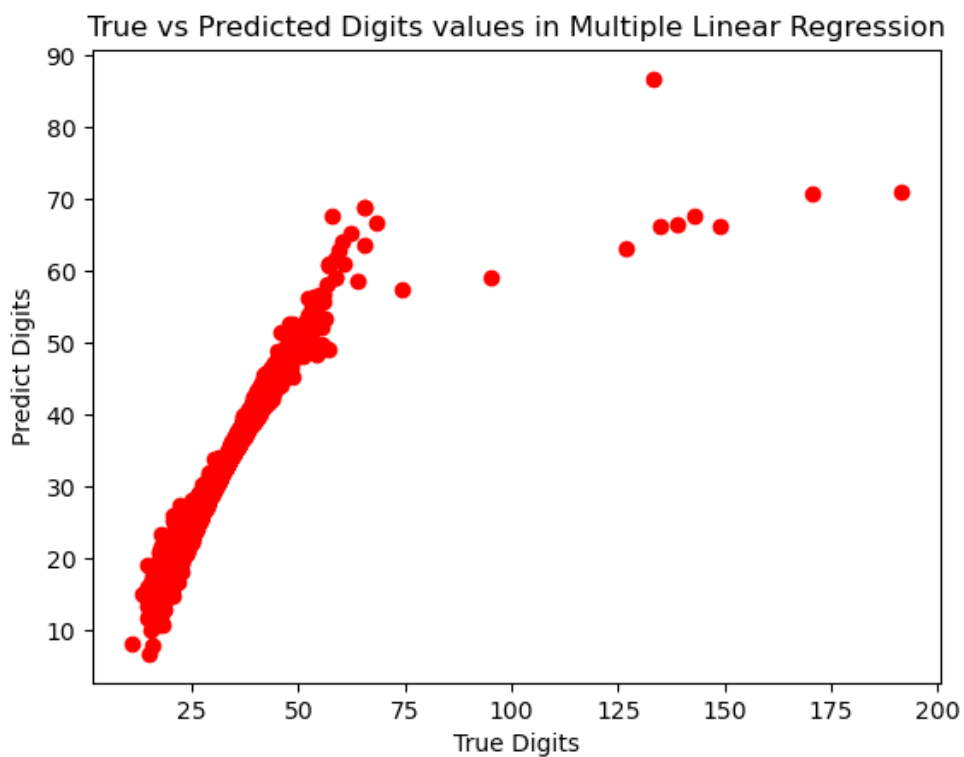
Accuracy: 89.99%

- KNN Regression:

  Accuracy: 91.38%

- Logistic Regression:

Accuracy with Logistic: 100.00%

- K-Means:

Cumulative Explained Variance