Computer Science Department

Report About :

Dataset and Model Documentation

Team Members :

1- Aya Ahmed Abdelsatar
2- Sama sameh abdelal
3- Rehap abdelghani mohamed
4- Mena esmat abdelghani
5- Sara omar mohamed
6- Sama haitham ezzat

Table of Contents

Dataset Overview

Numeric Dataset

- **Description:**
  - This dataset contains information related to life expectancy metrics for various countries over multiple years.
- **Number of Records:**
  - 2,342 entries (before preprocessing).
- **Number of Features:**
  - 21 features.
- **Missing Data:**
  - Yes, missing data was handled using imputation techniques during preprocessing.

Image Dataset

- **Description:**
  - This dataset contains images of food items categorized into 101 distinct classes. However, only 5 classes were used for this project.
- **Total Classes:**
  - 101 classes.
- **Used Classes:**
  - 5 classes.
- **Names of Used Classes:**
  - Pizza, French Fries, Ice Cream, Donuts, and Hamburger. **Missing Data**: No missing images; all images used were intact.

Preprocessing Steps

Numeric Dataset

1. Handling Missing Values:

- Used `SimpleImputer` to fill missing values.
- Numerical columns: Imputed with the mean.
- Categorical columns: Imputed with the most frequent value.

2. Feature Scaling:

- Applied `StandardScaler` to normalize numerical features.

1. **One-Hot Encoding**:
   - Categorical features were encoded using `OneHotEncoder`.
2. **Train-Test Split**:
   - Split data into 80% training and 20% testing sets.

Image Dataset

1- **Resizing:**
   - All images resized to a uniform size for consistency.

2- **Filtering:**
   - Applied Gaussian blur to reduce noise.

3- **Augmentation:**
   a. Applied random flips, rotations (±30°), brightness adjustments, contrast changes, and Gaussian noise.

4- **Normalization:**
   - Normalized pixel values to a range of [0, 1].

5- **Feature Extraction:**
   a. **Color Histogram**.
   b. **Local Binary Patterns (LBP)**.
   c. **Edge Detection** using Canny edge detector.

Models and Algorithms

Numeric Dataset Models

1- Linear Regression:

- Used for predicting life expectancy.
- **Evaluation Metrics**:
  - Mean Squared Error (MSE): 4.32
  - Root Mean Squared Error (RMSE): 2.08
  - R-Squared (R^2): 0.952

2- K-Nearest Neighbors (KNN):

- Used for predicting life expectancy.
- **Evaluation Metrics**:
  - Mean Squared Error (MSE): 9.55
  - Root Mean Squared Error (RMSE): 3.09
  - R-Squared (R^2): 0.893

Comparison of Numeric Models

Metric	Linear Regression	KNN Mean
Squared Error (MSE)	4.32	9.55
Root Mean Squared Error (RMSE)	2.08	3.09
R-Squared (R^2)	0.952	0.893

Conclusion:

Linear Regression outperformed KNN in all metrics for the numeric dataset.

Logistic Regression (Classification)

Purpose:

Binary classification of life expectancy as above or below the median value.

Evaluation Metric:

- Accuracy: 92.49%

Image Dataset Model

1. **Logistic Regression:**
   - Used features extracted from images to classify food items.
   - **Evaluation Metrics**:
     - Accuracy: Achieved high classification accuracy after augmentation and feature extraction.

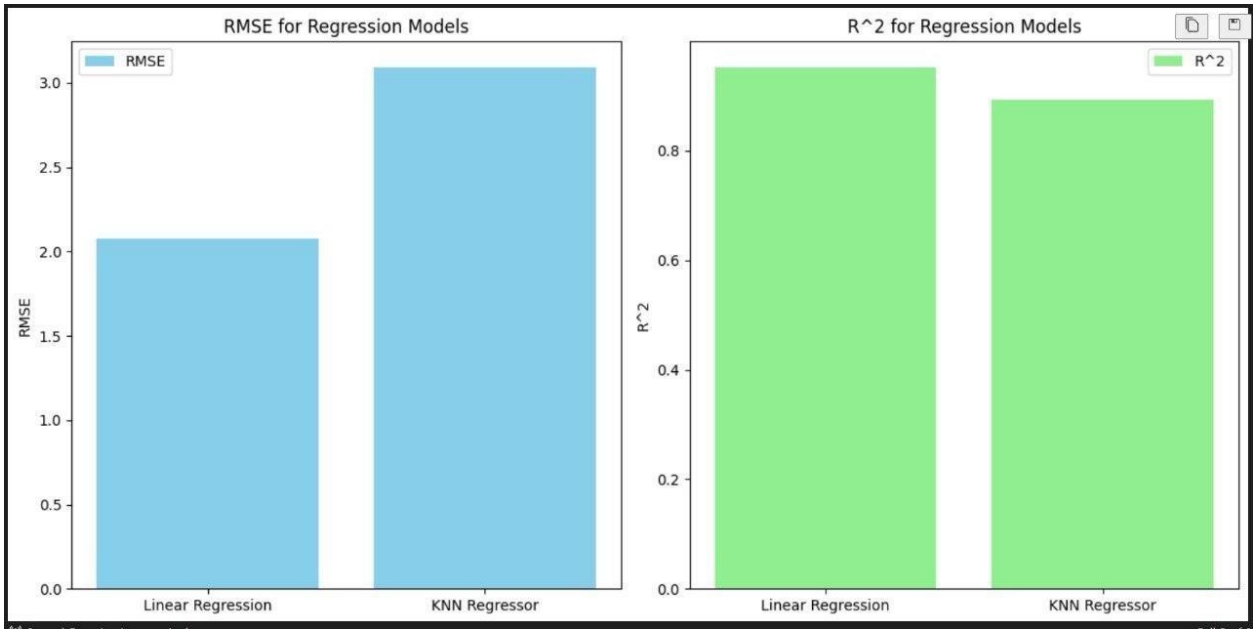
Tools and Libraries Used

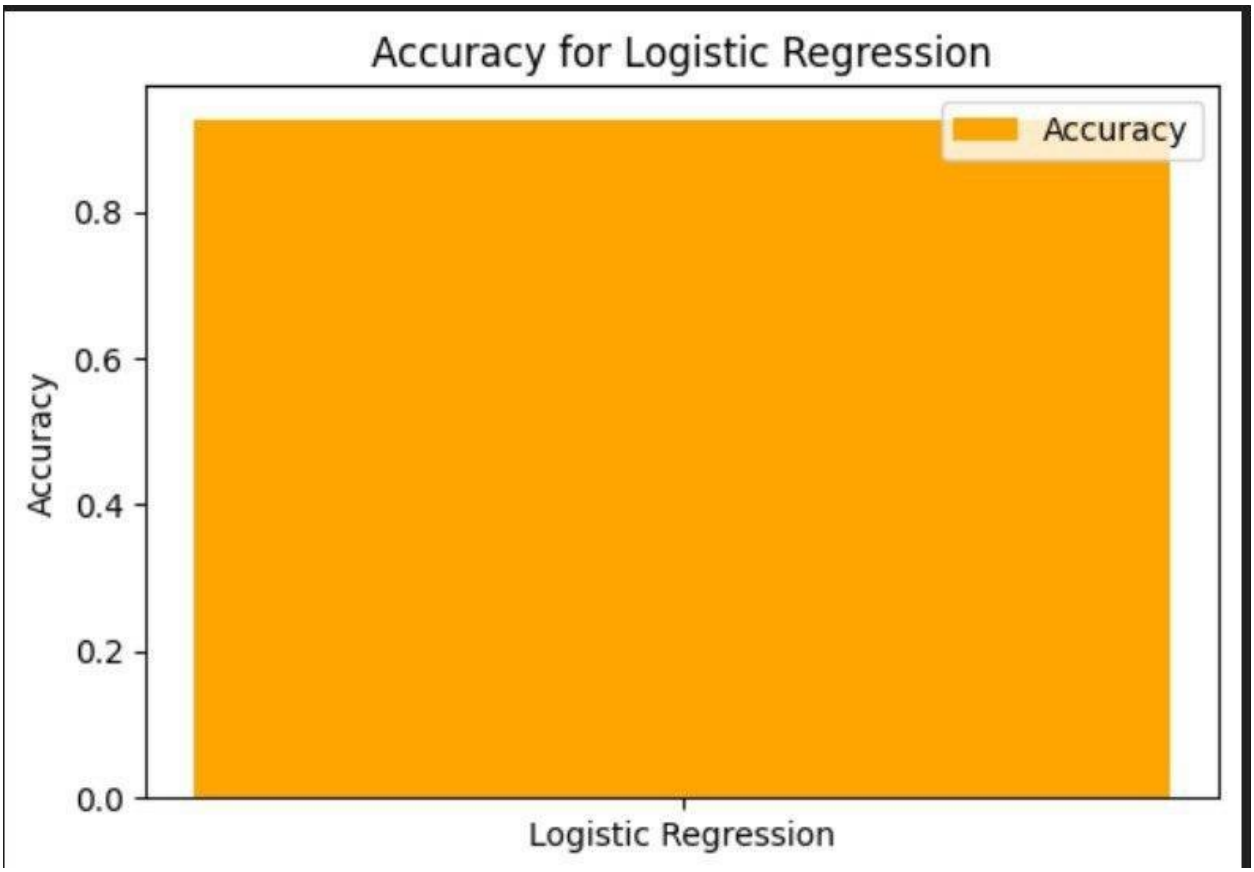
- **Libraries:**
  - Pandas, NumPy, Scikit-learn, Imgaug, OpenCV, Matplotlib.
- **Tools:**
  - Jupyter Notebook for experimentation.
  - Python for scripting and implementation.

Conclusion

- **Numeric Dataset:**
  - Linear Regression was the best-performing model for predicting life expectancy.
  - Logistic Regression effectively classified life expectancy into binary categories.
- **Image Dataset:**
  - Preprocessing and augmentation enhanced feature extraction and classification accuracy. o The chosen 5 classes (Pizza, French Fries, Ice Cream, Donuts, and Hamburger) were classified with high accuracy using Logistic Regression.

Comparison of Numeric Models





Comparison of Image Models