Faculty of computers and artificial intelligence Helwan University

Cover sheet

AI330 Machine Learning Project

Team no. :

Name	ID
زياد محمد خضر البيومي	20210373
بدر فيصل عبدالرؤوف عبد الحليم	20210219
باسم ياسر رجب عمر	20210217
مارتينا سليمان سامي	20210704
حازم وليد عبد المنجي ابو زيد	20210271
سلمى انور انور عبد العزيز	20210410

Project Description Document

Model 1: [Linear Regression]

General Information on Numerical Dataset:

- **Project Description Document Model 1:** Linear Regression for California Housing
- **Dataset Name:** California Housing Dataset
- **Number of Classes:** Regression task (predicting median house value)
- **Total Number of Samples:** 8025
- **Training Samples:** 6420
- **Testing Samples:** 1605

Implementation Details:

Feature Extraction Phase:

Number of Features Extracted: All features in the dataset

Feature Names:

- MedInc (median income in block)
- HouseAge (median house age in block)
- AveRooms (average rooms)
- AveBedrms (average bedrooms)
- Population (block population)
- AveOccup (average house occupancy)
- Latitude (house block latitude)
- Longitude (house block longitude)

Dimension of Resulted Features: 8 features

Cross-Validation:

- **Used**: No

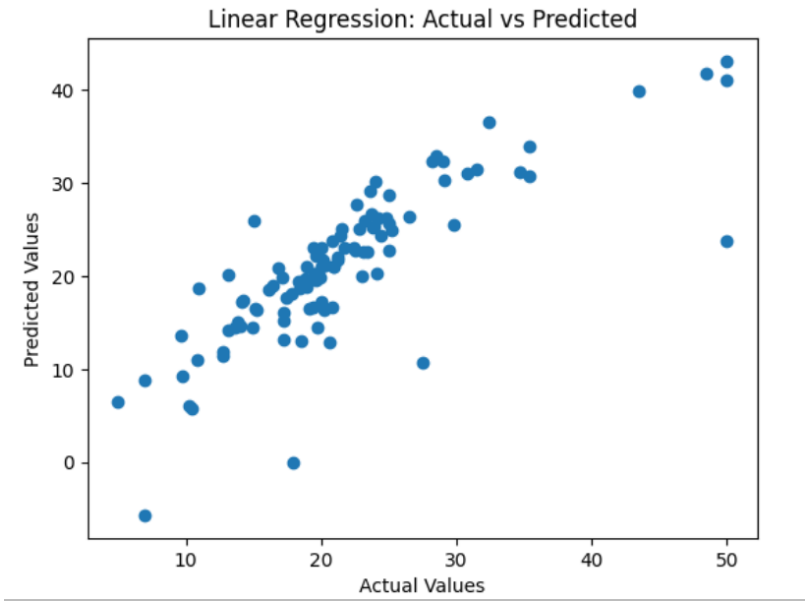
Hyperparameters:

- **Initial Learning Rate:** Not applicable (linear regression)
- **Optimizer:** Not applicable (linear regression)
- **Regularization:** Not applicable (linear regression)
- **Batch Size:** Not applicable (linear regression)
- **Number of Epochs:** Not applicable (linear regression)

Results Details:

Testing Data:

- **Loss Curve:** Not applicable (linear regression)
- **Accuracy:** Not applicable (regression task)
- **Evaluation Metric:** Mean Squared Error (MSE) , R^2 Score
- **MSE on Testing Data:** [25.017672023842596]
- **R^2 Score on Testing Data**: [0.6588520195508154]



Model 2: [KNN Regression]

General Information on Numerical Dataset:

- **Project Description Document Model 2:** KNN Regression for California Housing
- **Dataset Name:** California Housing Dataset
- **Number of Classes:** Regression task (predicting median house value)
- **Total Number of Samples:** 8025
- **Training Samples:** 6420
- **Testing Samples:** 1605

Implementation Details:

Feature Extraction Phase:

Number of Features Extracted: All features in the dataset

Feature Names:

- MedInc (median income in block)
- HouseAge (median house age in block)
- AveRooms (average rooms)
- AveBedrms (average bedrooms)
- Population (block population)
- AveOccup (average house occupancy)
- Latitude (house block latitude)
- Longitude (house block longitude)

Dimension of Resulted Features: 8 features

Cross-Validation:

- **Used**: Yes
- **Number of folds :** 5
- **Training/Validation Ratio/Testing :** 60% / 20% / 20%

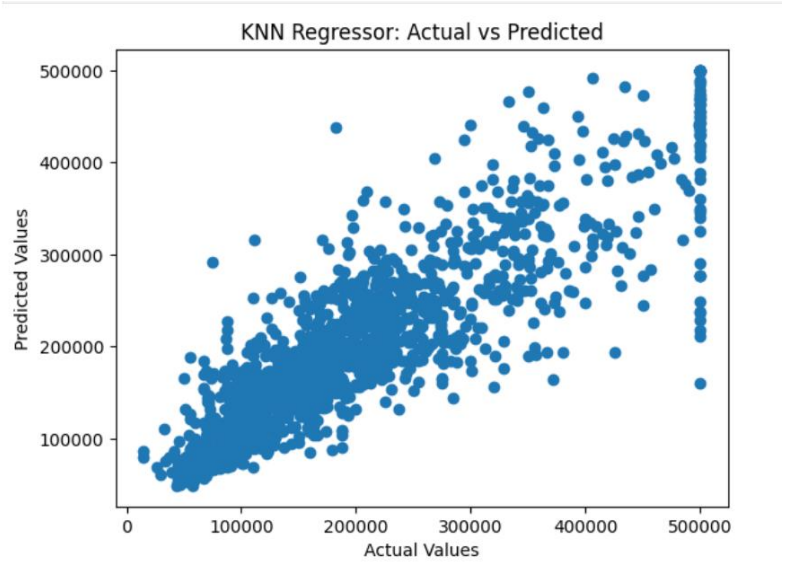
Hyperparameters:

- **Number of Neighbors (k):** [9]
- **Distance Metric:** Euclidean distance (default)

Results Details:

Testing Data:

- **Loss Curve:** Not applicable (KNN regression)
- **Evaluation Metric:** Mean Squared Error (MSE) , R^2 Score
- **`RMSE on Testing Data:`** [56438.011742719355]
- **R^2 Score on Testing Data**: [0.7545223381157126]
- **Mean MSE:** [3185249169.471328]
- **Standard Deviation MSE:** [262635941.85936466]



Model 1 : [Logistic Regression for UTKFace Dataset]

General Information on Image Dataset:

- **Project Description Document Model 1:** Logistic Regression for UTKFace Dataset
- **Dataset Name:** UTKFace Dataset
- **Number of Classes:** Gender Prediction (Regression task)
- **Total Number of Samples:** 3252
- **Training Samples:** 2602
- **Testing Samples:** 650

Implementation Details:

Feature Extraction Phase:

Number of Features Extracted: All features in the dataset

Feature Names :

- Age , index[0]
- Race , index[2]
- Date&time , index[3]

Target :

- Gender (Classify humans into 0 Male | 1 Female) , index[1]

Dimension of Resulted Features: 3 features

Cross-Validation:

- **Used**: No (Logistic Regression doesn't typically involve cross-validation during training)

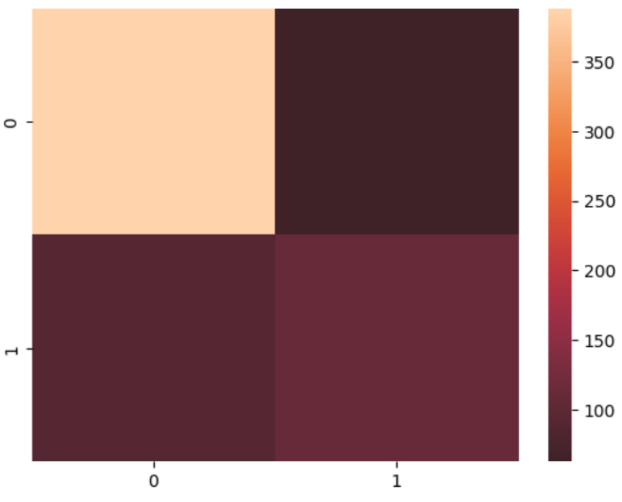
Hyperparameters:

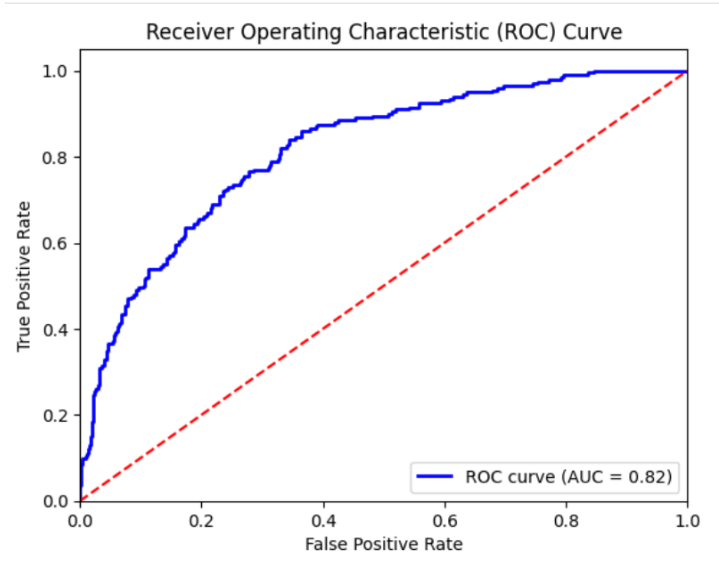
- **Regularization parameter (C) :** [1.0]
- **Solver:** ‘Saga’
- **Maximum iterations:** [100]
- **Penalty:** l2

Results Details:

Testing Data:

- **Loss Curve:** Not applicable (logistic regression)
- **Accuracy Score:** 0.7649
- **Evaluation Metric:** Mean Squared Error (MSE) , Confusion Matrix
- **Confusion Matrix :** [[388 63]
  [90 110]]
-
- **ROC AUC Score**: [0.705155]
- **ROC Curve** :





Model 2 : [K-Means Clustering for UTKFace Dataset]

General Information on Image Dataset:

- **Project Description Document Model 2 :** K-means Clustering for UTKFace Dataset
- **Dataset Name:** UTKFace Dataset
- **Number of Classes:** Unsupervised (Clustering task)
- **Total Number of Samples:** 3252
- **Training Samples:** 2602
- **Testing Samples:** 650

Implementation Details:

Feature Extraction Phase:

Number of Features Extracted: All features in the dataset

Feature Names :

- Age , index[0]
- Race , index[2]
- Date&time , index[3]

Target :

- Gender (Classify humans into 0 Male | 1 Female) , index[1]

Dimension of Resulted Features: 3 features

Cross-Validation:

- **Used**: No (K-means is unsupervised and does not involve cross-validation during training)

Hyperparameters:

- **Number of clusters (K) :** [2]

Results Details:

Testing Data:

- **Inertia:** 1120.2396
- **Silhouette:** 0.750
- **Visualize the cluster :**
- **Select K randomly and loop from 2 to 11 , in each iteration save inertia in the list after end the loop then compare all inertias and plot Elbow to select the right K**