**Team Members1:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Team Member ID | Team member name (**in Arabic**) | Grade |
| 1 | 20210064 | أحمد صلاح الدين أحمد سيد |  |
| 2 | 20210063 | أحمد تحسين صلاح الدين عبدالمجيد |  |
| 3 | 20210124 | أحمد هاني سعد محمد علي |  |
| 4 | 20210159 | إسلام مجدي محمد الصفي |  |
| 5 | 20210074 | أحمد عبدالغني أحمد عبدالغني |  |
| 6 | 20210081 | أحمد علي إبراهيم سليمان |  |
| 7 | 20210057 | أحمد سعدالدين محمد علي المراكبي |  |

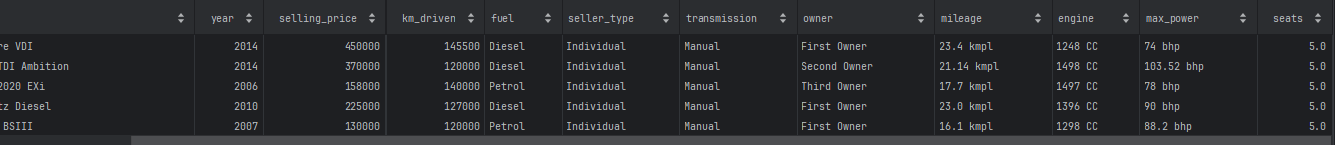
Numerical:

General information about dataset:

|  |  |
| --- | --- |
| Name: | [Vehicle dataset](https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho) |
| Total number of samples | 7530 |
| Total number of training | 6024 |
| Total number of test | 1506 |

Linear Regression

I have features [‘ km\_driven’,’ seller\_type’,’ seller\_type’,’ transmission’, ’ owner’,’ mileage’,’ engine’,’ max\_power’,’ seats’]

My dataset before cleaning   


I applied encoding to convert categorical string values, such as 'diesel' and 'petrol', into numerical representations, with a subsequent addition of 1 to all values to avoid potential model confusion with zero. I then proceeded to transform specific features like 'mileage', 'engine', and 'max\_power' into numerical values. To enhance data quality, I removed nullable values from the dataset.

Following this, I computed z-scores and selected specific rows based on the calculated scores. To enrich the dataset, I introduced a new column representing the age of the entries, I removed year and name columns.

So my dataset after cleaning is :

A black screen with numbers and letters

Description automatically generated

I have split my dataset to get train and test data then I applied standard scaler to normalize data and make\_pipeline to make PolynomialFeatures and linear regression to work as one model then I fit my training data then predict   
I have calculated R-squared score :

A number on a black background

Description automatically generated

The linear regression plot:

A graph with a line and a red line

Description automatically generated with medium confidence

The cross validation plot:

A graph with a line and a red line

Description automatically generated with medium confidence

Cross validation Score:



Linear regression Loss Curve:

A graph with text and numbers

Description automatically generated

K-NeighborsRegressor

I used same cleaned data I made before

R-squared score is :

A number with green numbers

Description automatically generated

K-NeighborsRegressor plot:

A graph with a line and a line

Description automatically generated with medium confidence

K-NeighborsRegressor Loss curve:

A graph with numbers and lines

Description automatically generated

Conclusion:

From all the above we found that linear regression is better than K- NeighborsRegressor according to R-squared score.

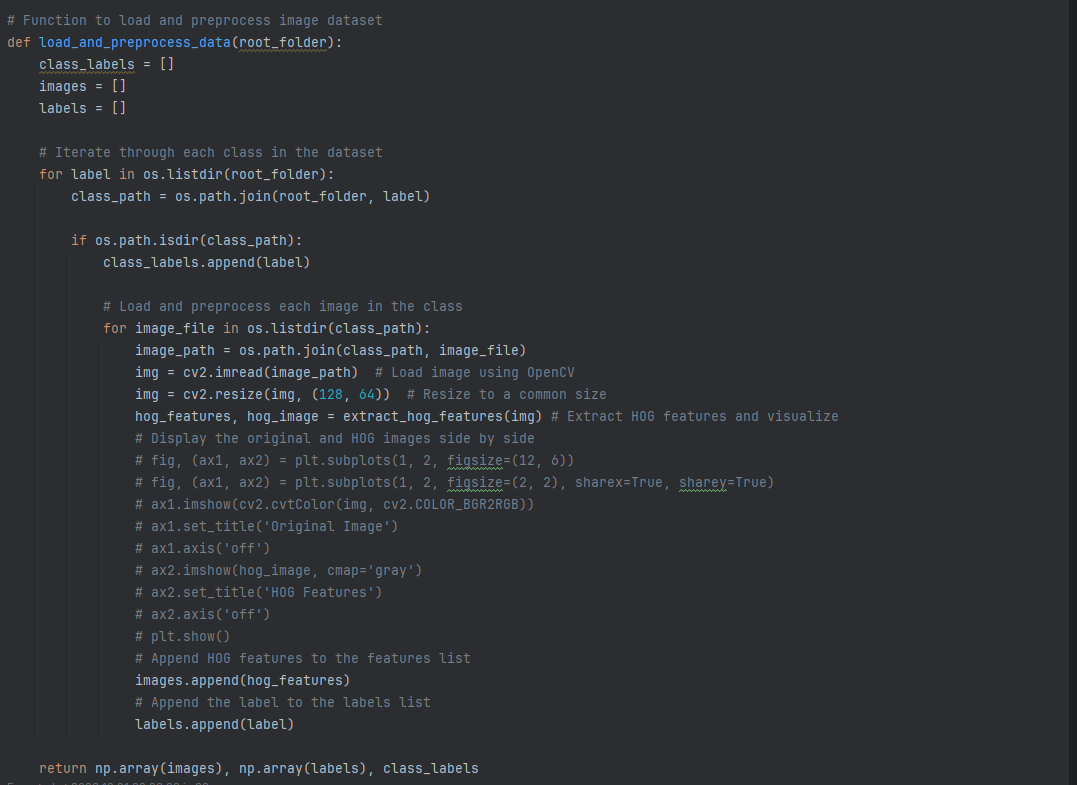
Image:

General information about dataset:

|  |  |
| --- | --- |
| Name: | [Fruit](https://www.kaggle.com/datasets/moltean/fruits)360 |
| Total number of samples | 3239 |
| Total number of training | 2072 |
| Total number of test | 648 |
| Total number of validation | 519 |

Logistic Regression

Here we preprocessing and load our data



And we use HOG function to extract features

A screen shot of a computer

Description automatically generated

I split it into train, test, validation to see the loss curve

A graph with text on it

Description automatically generated

Confusion Matrix for Test Set

A screenshot of a computer

Description automatically generated

ROC Curve on Test Set

A graph with a line and numbers

Description automatically generated

Classification Report

A screenshot of a computer screen

Description automatically generated

Predicts the test data correctly



Kmeans

Here we don’t use HOG function we just flatten our image to convert it from 2D to 1D

A computer screen shot of a black screen

Description automatically generated

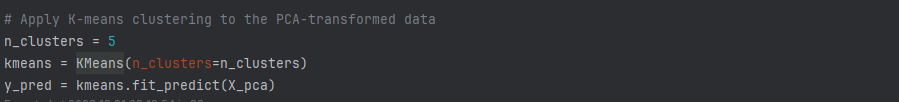
We didn’t split the data into train, and test as in logistic regression

We used the elbow method to choose optimal K

A graph with a line

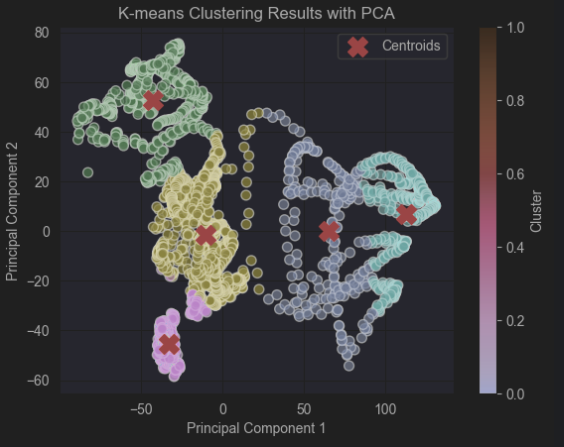
Description automatically generated

Before using Kmeans we used PCA to reduce our dimensionality (features)



We used 5 clusters as we have 5 different classes of fruits

And here is the Kmeans Clustering



Here we calculated the silhouette and intertia

A black rectangular object with a black border

Description automatically generated