Applying Different Preprocessing Technique To Enhance Accuracy of Prediction.

# Machine Learning

Project

# Machine Learning Project

| ID | Name | Grade |
|---|---|---|
| **20210520** | عبدالرحمن عمرو محمد محمد | |
| **20211057** | يوسف احمد عبدالرؤف احمد | |
| **20211061** | يوسف احمد محمود علي علي | |
| **20211036** | هنا محمد مصطفي | |
| **20211077** | يوسف صلاح يوسف | |
| **20210322** | رحاب ابراهيم علي | |

**Datasets:**

**Image :**
**https://www.kaggle.com/datasets/moltean/fruits**

**Numerical :**
**https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques**

## Numerical

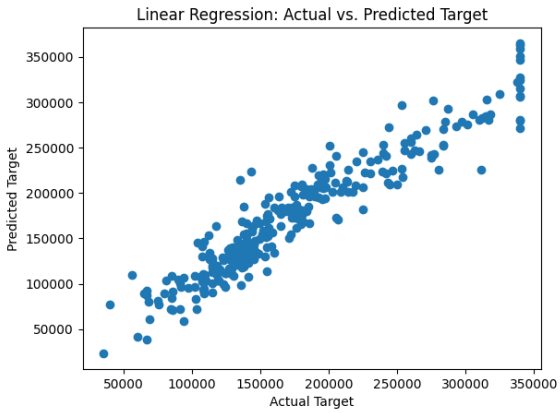
## House Price Prediction

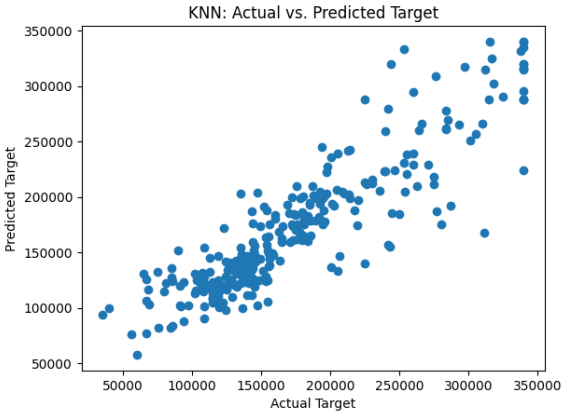
| Name | House Price Prediction |
|---|---|
| Target | SalePrice |
| No. Samples | 1460 |
| No. Training | 1168 |
| No.Testing | 292 |

**After Applying Preprocessing Techniques (Handling Missing Values- Handling Outliers -Scaling – Encoding Categorical)**

**Here is The List of Features we use:**

LotFrontage, LotArea, LotShape, Neighborhood, OverallQual, YearBuilt, YearRemodAdd, MasVnrType, MasVnrArea, ExterQual, Foundation, BsmtQual, BsmtExposure, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, HeatingQC, CentralAir, Electrical, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, KitchenAbvGr, KitchenQual, TotRmsAbvGrd, Fireplaces, GarageType, GarageYrBlt, GarageFinish, GarageCars, GarageArea, PavedDrive, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal, SaleCondition




**After Applying KNN and Linear Regression as a Regressor:**

We found that Linear Regression tends to perform better than KNN in term of Accuracy of Prediction:

Knn accuracy : 84.09%

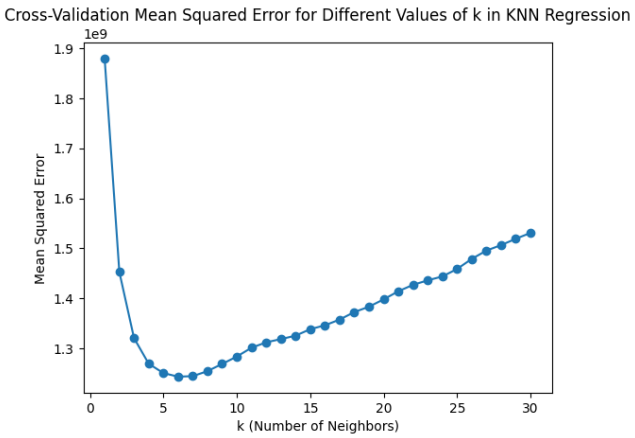
Linear Regression : 90.25%

As we can see in the given plots

In KNN we try to apply some sort of hyperparameter Tuning using cross validation to obtain the best value of k

*We use the whole dataset

**Here is what we got:**




The Best value of k is 6 which increase the accuracy of knn from 80% (k=3) to 84% (k=6)

Image

Fruits-360

| Name | Fruits-360 |
| --- | --- |
| Target | Classify Each fruit |

*The dataset include (131) classes

**Here is the 5 classes we use:**

Training Set Fruit Counts:
Apple Braeburn: 492 images
Banana: 490 images
Strawberry: 492 images
Pineapple: 490 images
Mango: 490 images

Testing Set Fruit Counts:
Apple Braeburn: 164 images
Banana: 166 images
Strawberry: 164 images
Pineapple: 166 images
Mango: 166 images

The preprocessing Technique we use :

Applying grayscale

*It helps in simplifying algorithms and as well eliminates the complexities related to computational requirements.

Resizing The Image (64,64)

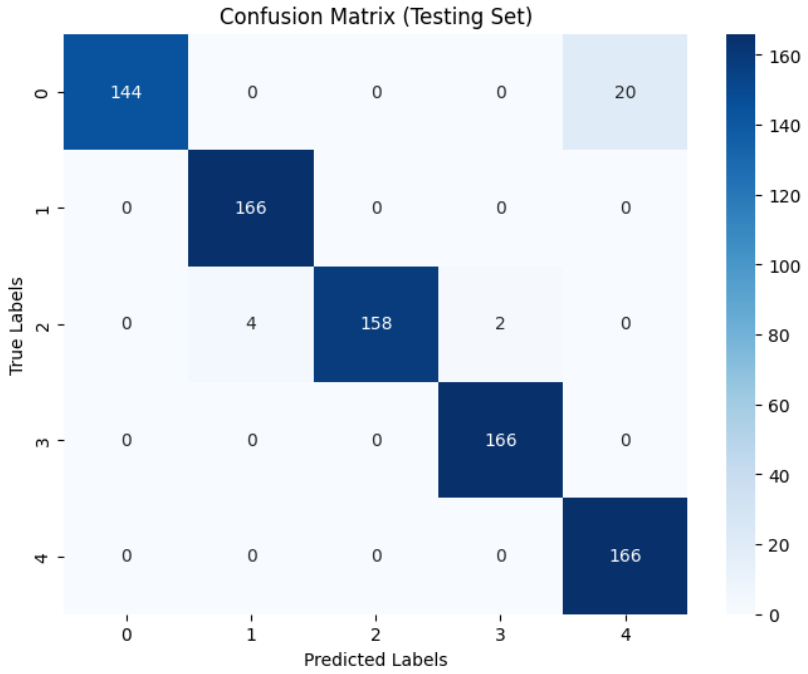
Applying HOG As a feature Extractor for The images

**Logistic Regression :**

We use Logistic Regression as a classifier

With technique OVR and Maxt_iteration=1000

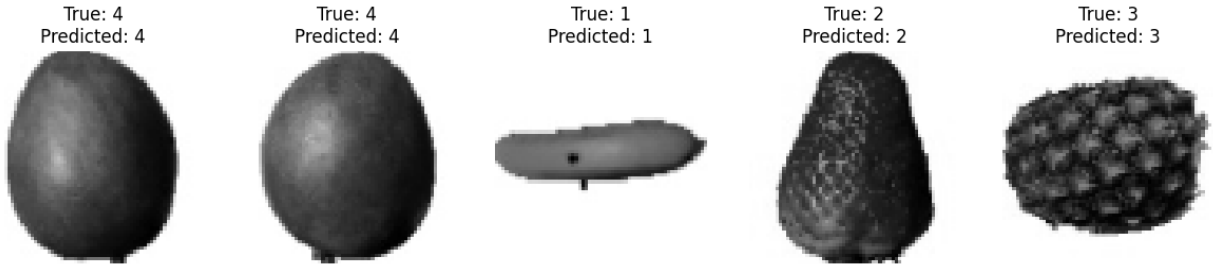
**Here is the result:**




Accuracy = 96.85%

*We Notice that if we pass the images to the model without applying HOG The accuracy decreases slightly to 92.37%

**Here is some Random Samples :**



**K-Mean :**

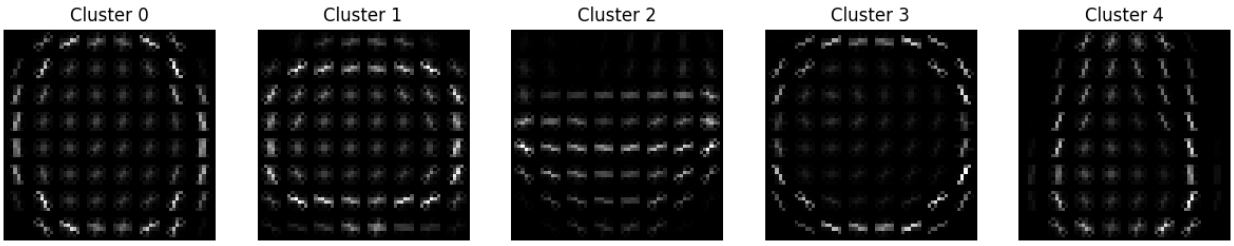
We use K-mean for Clustering

*we apply it on the whole 5 classes we use

*No. Clusters = 5

**Here is an image for each cluster :**

(HOG Feature Extracted)




- The dataset we use contain images with their labels
- We think that there is no need to apply clustering in that situation if we want to classify them using the provided labels