House Prices Dataset:

Overview:

The House Price Dataset contains variables describing (almost) every aspect of residential homes in Ames, Iowa.

Dataset size:

Initially 2919 row x 75 column.

Goal:

Predict sales price for each house based on its features.

Training and Testing:

**Training:** Used 2335 rows of 10 selected best features. (80% of samples)

**Testing:** Used 584 rows of 10 features. (20% of samples)

Description of the selected features:

OverallQual: Overall material and finish quality.

GrLivArea: Above grade (ground) living area square feet.

GarageCars: Size of garage in car capacity.

ExterQual: Exterior material quality.

GarageArea: Size of garage in square feet.

TotalBsmtSF: Total square feet of basement area.

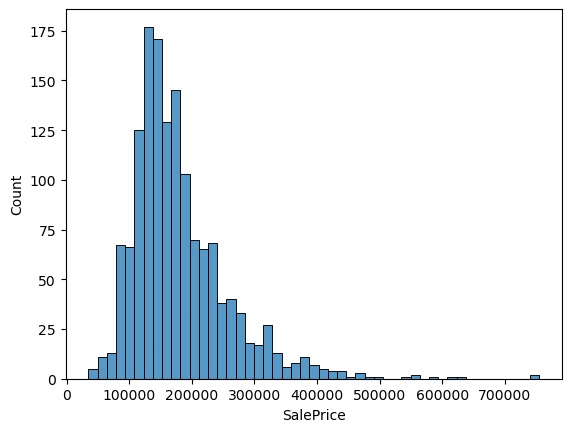
1stFlrSF: First floor square feet.

BsmtQual: Height of the basement.

KitchenQual: Kitchen quality.

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms).

Distribution of SalesPrice:



Summary:

First we handled missing values by dropping the columns with too many missing values and filling columns with moderate amount of missing values with column mean for numerical columns and mode for categorical ones. then normalized categorical values using Label Encoder. Then selected the 10 most relevant features to train the model. Then applied Linear Regression and KNN to the training data. At the end compared the performance of each model.

Performance Comparison:

|  |  |  |
| --- | --- | --- |
|  | Linear Regression | KNN |
| R2 | 0.408 | 0.318 |
| MAE | 31253.27 | 32615.22 |
| MSE | 2179459590.9 | 2512043957.0 |
| RMSE | 46684.7 | 50120.3 |

Overall:

Linear Regression performs better than KNN by 28%.

Plant Disease Dataset:

Overview:

The PlantVillage Dataset is a collection of images of healthy and diseased plant leaves, widely used for training machine learning models in plant disease classification.

Dataset description:

No. Samples: 20.6k images.

Number of Classes: 15 classes.

Image Dimensions: 256x256.

Goal:

Classify the plant disease based on the leaf image.

Classes:

Pepper\_\_bell\_\_\_Bacterial\_spot

Pepper\_\_bell\_\_\_healthy

Potato\_\_\_Early\_blight

Potato\_\_\_Late\_blight

Potato\_\_\_healthy

Tomato\_Bacterial\_spot

Tomato\_Early\_blight

Tomato\_Late\_blight

Tomato\_Leaf\_Mold

Tomato\_Septoria\_leaf\_spot

Tomato\_Spider\_mites\_Two\_spotted\_spider\_mite

Tomato\_\_Target\_Spot

Tomato\_\_Tomato\_YellowLeaf\_\_Curl\_Virus

Tomato\_\_Tomato\_mosaic\_virus

Tomato\_healthy

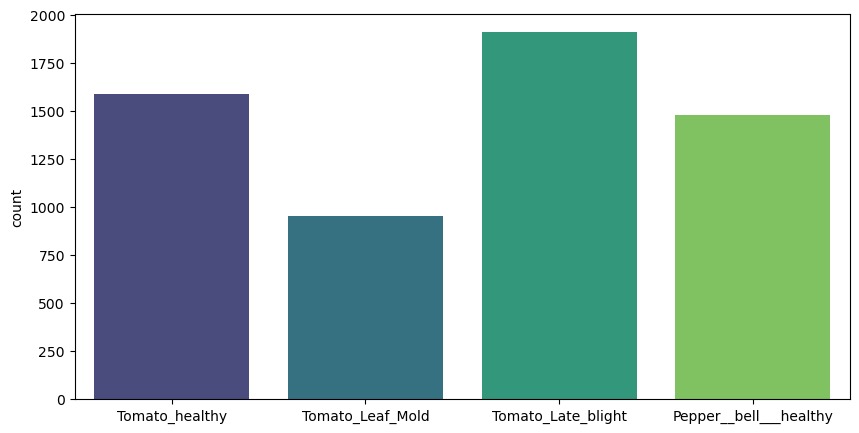
Selected classes:

Tomato\_healthy: 1591 images.

Tomato\_Early\_blight: 1000 images.

Tomato\_Late\_blight: 1909 images

Pepper\_\_bell\_\_\_healthy: 1478 images.



Training and Testing:

**Training:** 4782 samples. (80%)

**Testing:** 1196 samples. (20%)

Performance Comparison:

|  |  |  |
| --- | --- | --- |
|  | Logistic Regression | KNN |
| Accuracy | 0.8 | 0.68 |

Logistic Regression performs better than KNN by 18%.

Logistic Regression:

A graph with a line

Description automatically generated

A graph with a line and numbers

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph of a logistic regression

Description automatically generated

KNN:

A graph of a curve

Description automatically generated

A screenshot of a graph

Description automatically generated