AI330: Machine Learning – Project Documentation

Numerical Dataset:

- **Dataset Name**: Car Prices Dataset
- **No. of samples**: 19,237
- **Samples used for training/testing**: 8,719 / 2,180
- **No. of features/attributes**: 18
- **Missing values**: Yes
- **Algorithms used:**
    - Linear Regression:
        - Linear regression is a method used to predict a numerical outcome by modeling the relationship between the target variable and one or more input features as a straight line. It assumes a linear relationship in the data and minimizes the error between predicted and actual values.
    - KNN Regression:
        - K-Nearest Neighbors (KNN) regression predicts a numerical value by averaging the target values of the KKK-closest data points based on distance. It makes no assumptions about data distribution and works well for non-linear relationships but can be computationally expensive for large datasets.

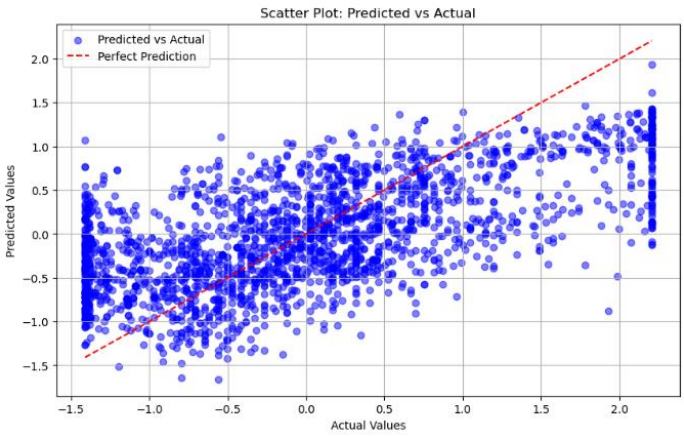
Evaluation Metric	**Linear Regression**	**KNN Regression**	**Xgboost Regressor**
Mean Absolute Error	6991	3879	2994
Mean Squared Error	78776322	33706767	20948431
R-squared	0.40	0.745	0.84

Comparison

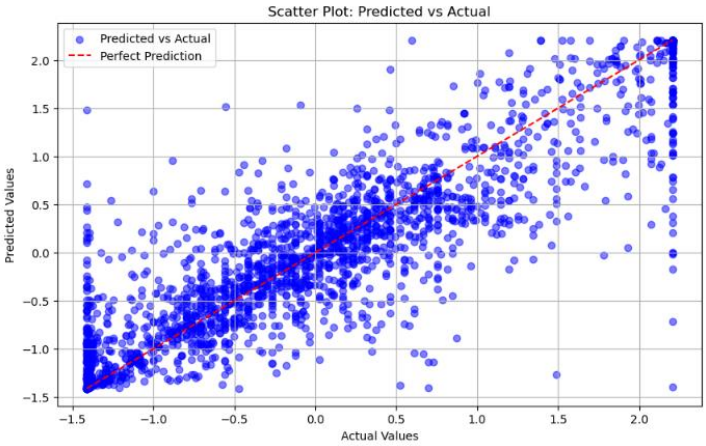
Our KNN Regression model is more accurate than linear regression as it has a lower MAE and MSE, and a higher R-squared. We also used a third built-in model called Xgboost Regressor that is better than both Linear and KNN regression.

Visualisations

Scatter Plot for Linear Regression



Scatter Plot for KNN Regressor



Scatter Plot for Xgboost Regressor

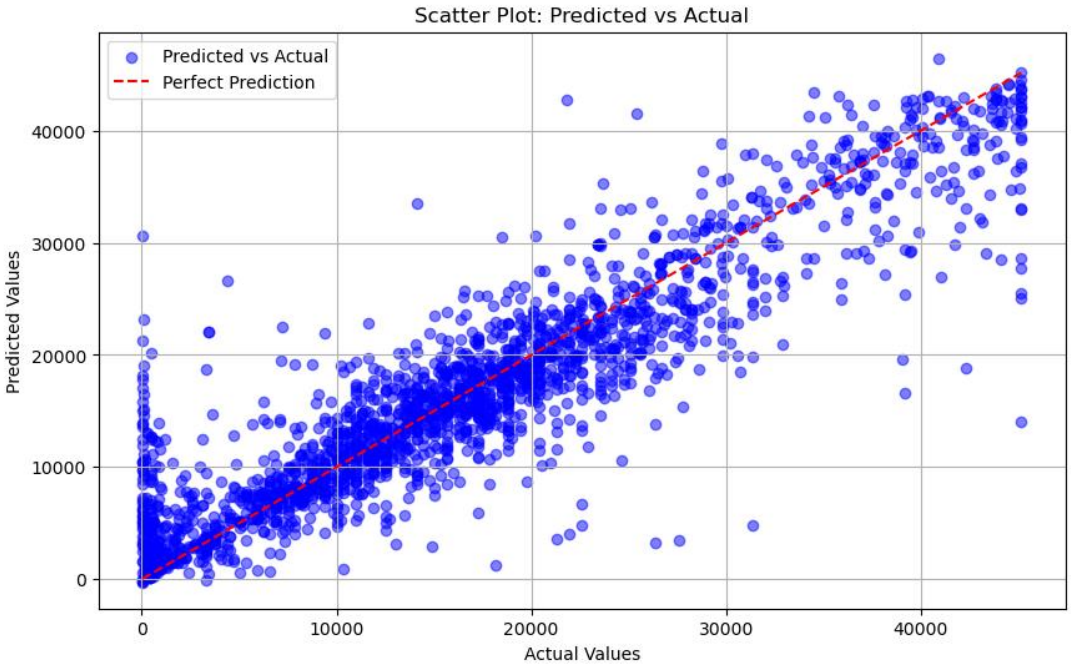


Image Dataset:

- **Dataset Name**: Cell Images For Detecting Malaria
- **No. of Classes**: 2 (Parasitized, Uninfected)
- **No. of Samples**: 27,600 (13,800 per class)
- **No. of Samples used in Training/Testing**: 22,046 / 5,512
- **Size of sample**: Approx. 120x120 pixels
- **Missing Values**: No

- **Algorithms used**:

    - Logistic Regression
        - Logistic regression is used for classification tasks, such as identifying the category of an image (e.g., parasitized/uninfected cell). It models the relationship between input features (like pixel values) and a binary or multi-class output by estimating the probability that an image belongs to a particular class.
    - KNN Regression
        - K-Nearest Neighbors (KNN) regression for image datasets predicts a numerical output (like a pixel value or image score) by averaging the target values of the KKK-closest images. Similarity between images is determined using a distance measure, such as Euclidean distance, across pixel or feature values.

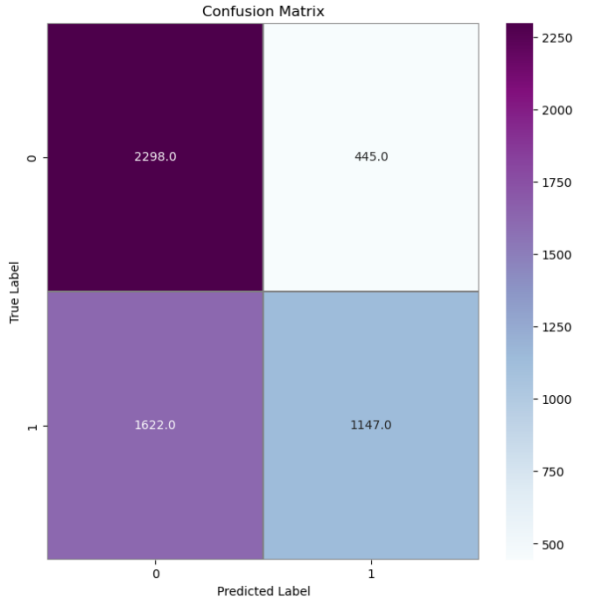
Evaluation Metric	**Logistic Regression**	**KNN Regression**
Accuracy Score %	80.95%	62.5%
Area-under-curve (AUC)	0.81	0.63
Log-loss	1.2	0.98

Comparison

Logistic Regression is generally more accurate as it has a higher accuracy score and AUC. It does, however, have a lower log-loss.

Visualisations

- Confusion Matrix for KNN



Confusion Matrix for Logistic Regression