SECTION 1

A Scientific report for Analysing Second-hand Car Sales Data with Supervised and Unsupervised Learning Models

ABSTRACT

The purpose of this report is to explore how supervised learning models can be used to predict the price of a second-hand car, based on some parameters contained in this dataset.

Keywords: Random Forest, Artificial Neural Network (ANN), k-Means clustering, Agglomerative Hierarchical, DBSCAN, Simple Linear Regression (SLR), Davies Bouldin (DB), Silhouette Score (SS), Hyper Parameter Loss Plot (HPLP).

1. INTRODUCTION

Artificial Intelligence plays a significant role in the automobile industry because of predictive analytics that enable better forecasting for the selling of cars for manufacturers and retailers and this can greatly contribute to the global economic system. This report aims to analyse different Artificial Intelligence models which can be used to accurately predict car sales using different criteria and evaluation metrics.

2. RELATED WORKS

Research and studies on car price prediction has been used in a lot of Artificial intelligence methodologies. Agrahari et al., (2021) explored optimal models for estimating used car values. Gegic et al., (2019) used machine learning models to predict used car prices in Bosnia and Herzegovina. Furthermore, Dnyaneshwar et al., (2023) found out more parameters that could affect car prices in certain demographics like age, income which was considered to make their models more accurate. These findings, studies the effectiveness of supervised learning, neural networks, and regression analyses, factoring in vehicle specifics, year, model market conditions, and historical data and more to build a suitable model for their predictions. (Gegic et al., 2019; Agrahari et al., 2021; Abishek R, 2022; Dnyaneshwar et al., 2023).

3.0 METHODOLOGY

3.1 Data Description

The data used for this report as shown in (Figure 1.0) is a 50,000-car sales dataset containing ["Mileage", "Year of manufacture", "Fuel type", "Engine size", "Model", "Manufacturer"] which will be used to perform all analysis and models.

	Manufacturer	Model	Engine size	Fuel type	Year of manufacture	Mileage	Price
0	Ford	Fiesta	1.0	Petrol	2002	127300	3074
1	Porsche	718 Cayman	4.0	Petrol	2016	57850	49704
2	Ford	Mondeo	1.6	Diesel	2014	39190	24072
3	Toyota	RAV4	1.8	Hybrid	1988	210814	1705
4	VW	Polo	1.0	Petrol	2006	127869	4101
...	...	...	...	...	...	...	...
49995	BMW	M5	5.0	Petrol	2018	28664	113006
49996	Toyota	Prius	1.8	Hybrid	2003	105120	9430
49997	Ford	Mondeo	1.6	Diesel	2022	4030	49852
49998	Ford	Focus	1.0	Diesel	2016	26468	23630
49999	VW	Golf	1.4	Diesel	2012	109300	10400

50000 rows × 7 columns

Figure 1.0 Dataset

3.2 Models Used

The Supervised and Unsupervised Learning regression models were used in this report to predict car prices are Simple Linear Regression (SLR), Polynomial 2nd degree, Random Forest, Artificial Neural Network (ANN), k-Means clustering, Agglomerative Hierarchical, DBSCAN clustering.

3.3 Evaluation Metrics

Mean absolute error (MAE), Mean squared error (MSE), Root mean squared error (RMSE), Coefficient of determination (R2), Davies Bouldin index and Silhouette Coefficient, were used to evaluate the model.

4.0 RESULTS

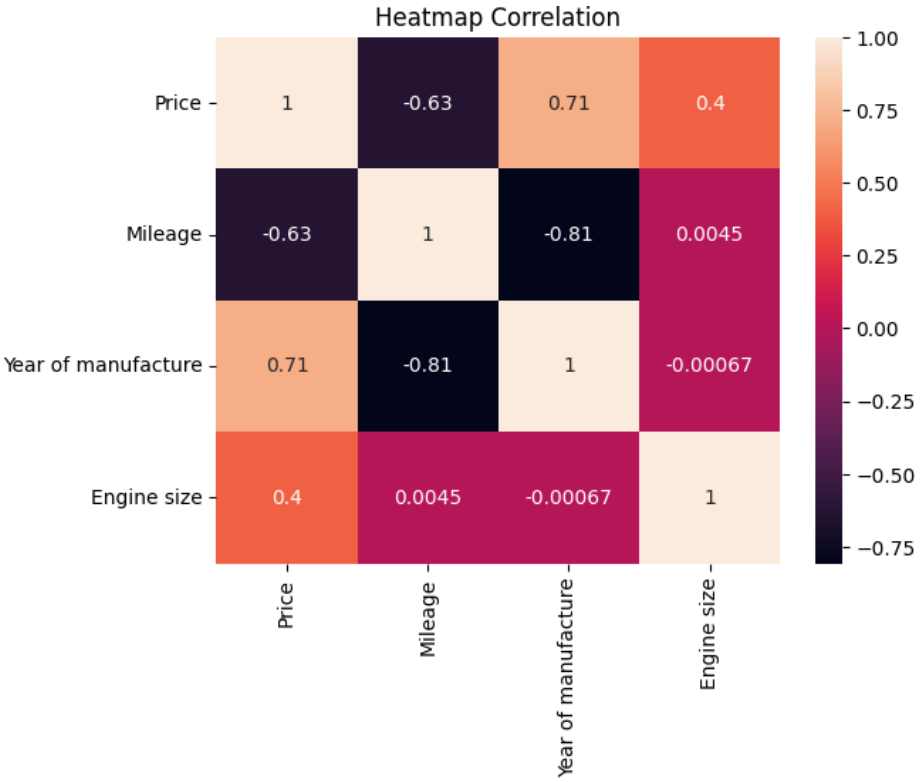

Figure: 2.0 Heatmap of corelation

4.1 Simple Linear Regression Model Result

Table 1.0 Simple Linear Regression Result

Feature names	Gradient	Intercept	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R2 (R squared)
Mileage	-10383.851	13814.105	7964.785	162468566.873	12746.316	0.401
Year of Manufacture	11715.456	13825.710	7031.039	132678999.948	11518.637	0.511
Engine size	6538.709	13853.141	10817.492	230499154.453	15182.199	0.151

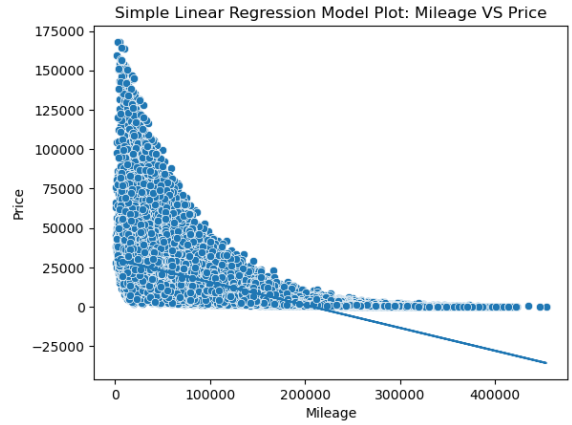

Figure 3.0 SLR(Millage vs Price)plot

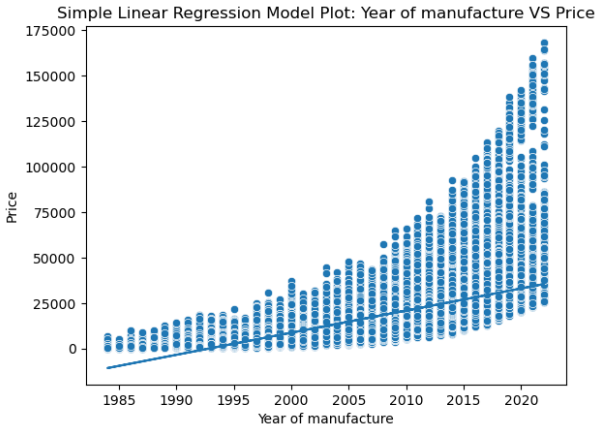

Figure 3.1 SLR(Year of Manufacture vs Price) plot

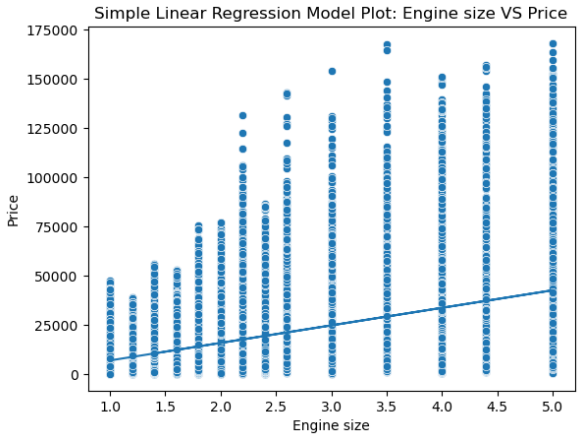

Figure 3.2 SLR(Engine vs Price) plot

4.2 Non-Linear Model: Polynomial 2nd Degree

Table 1.1 Polynomial 2nd Degree Result

Feature names	Coefficients	Intercept	Mean Absolute Error	Mean Squared Error	Root Mean Squared error	R Squared (R2)
Mileage	[-13608.66073765, 4484.590535 14]	9325.301	6409.912	129620312.163	11385.092	0.522

Year of Manufacture	[12043.66326258, 5285.35400178]	8542.972	5387.109	105993894.202	10295.334	0.609
Engine size	[6045.42772837, 236.02040276]	13614.558	10807.262	230326165.999	15176.500	0.151

4.3 Multiple Linear Regression Model Result

Table 1.2 Multiple Linear Regression Result

Feature names	**Gradient**	**Intercept**	**Mean Error**	**Mean Squared Error**	**Root Mean Squared error**	**R Squared (R2)**
Mileage, Engine size, Year of Manufacture	-2722.981	13819.002	6091.458	89158615.760	9442.384	0.671

4.4 Random Forest regression Model

Table 1.3 Random Forest result

Feature names	**Mean Absolute Error**	**Mean Squared Error**	**Root Mean Squared error**	**R Squared (R2)**
Mileage, Engine size, Year of Manufacture	2294.500	20156062.711	4489.550	0.926

Mileage, Year of manufacture, Fuel type, Engine size, Model, Manufacturer	333.272	480136.153	692.919	0.998

4.5 ANN Model

Table 1.4 ANN Model Results

Feature names	**Mean Absolute Error**	**Mean Squared Error**	**Root Mean Squared Error**	**R Squared (R2)**
Model 1: 6 layers and default learning rate	2367.301	7529814.555	2744.051	0.972
Model 2: 7 ANN layers and default learning rate	2680.830	9117991.614	3019.601	0.966
Model 3: Learning rate 0.0001	2088.120	5881353.961	2425.150	0.978
Model 4: Dropout rate = 0.1	1118.075	2175148.228	1474.838	0.992

4.5.1 ANN Architecture

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	448
dropout (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 64)	4160
dense_2 (Dense)	(None, 64)	4160
dense_3 (Dense)	(None, 64)	4160
dense_4 (Dense)	(None, 64)	4160
dense_5 (Dense)	(None, 1)	65

Total params: 17153 (67.00 KB)
Trainable params: 17153 (67.00 KB)
Non-trainable params: 0 (0.00 Byte)

Figure 3.3 *ANN Architecture (model 1)*

Model: "sequential_1"

Layer (type)	Output Shape	Param #
dense_6 (Dense)	(None, 64)	448
dropout_1 (Dropout)	(None, 64)	0
dense_7 (Dense)	(None, 64)	4160
dense_8 (Dense)	(None, 64)	4160
dense_9 (Dense)	(None, 64)	4160
dense_10 (Dense)	(None, 64)	4160
dense_11 (Dense)	(None, 64)	4160
dense_12 (Dense)	(None, 1)	65

Total params: 21313 (83.25 KB)
Trainable params: 21313 (83.25 KB)
Non-trainable params: 0 (0.00 Byte)

Figure 3.4 *ANN Architecture (model 2)*

Model: "sequential_2"

Layer (type)	Output Shape	Param #
dense_13 (Dense)	(None, 64)	448
dropout_2 (Dropout)	(None, 64)	0
dense_14 (Dense)	(None, 64)	4160
dense_15 (Dense)	(None, 64)	4160
dense_16 (Dense)	(None, 64)	4160
dense_17 (Dense)	(None, 64)	4160
dense_18 (Dense)	(None, 1)	65

Total params: 17153 (67.00 KB)
Trainable params: 17153 (67.00 KB)
Non-trainable params: 0 (0.00 Byte)

Figure 3.5 *ANN Architecture (model 3)*

Model: "sequential_3"

Layer (type)	Output Shape	Param #
dense_19 (Dense)	(None, 64)	448
dropout_3 (Dropout)	(None, 64)	0
dense_20 (Dense)	(None, 64)	4160
dense_21 (Dense)	(None, 64)	4160
dense_22 (Dense)	(None, 64)	4160
dense_23 (Dense)	(None, 64)	4160
dense_24 (Dense)	(None, 1)	65

Total params: 17153 (67.00 KB)
Trainable params: 17153 (67.00 KB)
Non-trainable params: 0 (0.00 Byte)

Figure 3.6 *ANN Architecture (model 4)*

4.5.2 ANN Model Prediction Plot with 6 layers

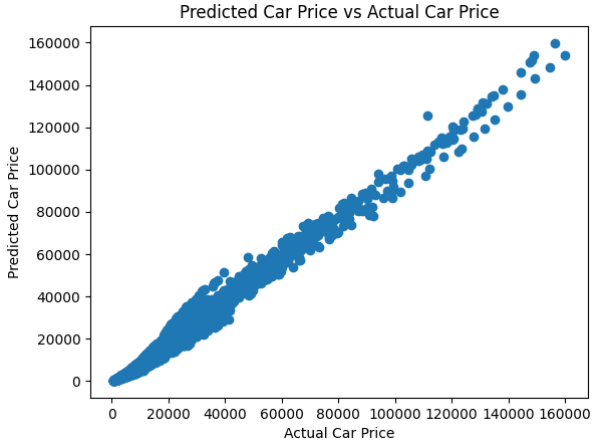

Figure 3.7 ANN Prediction on Car price and Actual Car Price (model 1)

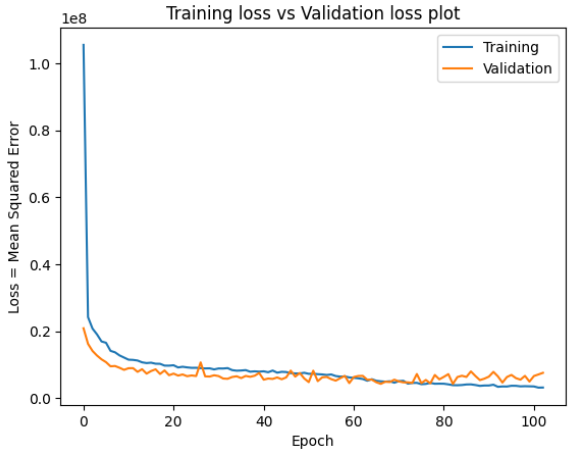

Figure 3.8 ANN Hyper Parameter Loss Prediction Plot (model 1)

4.5.3 ANN Hyperparameter Tunning 7 Layers Car price Prediction Plot

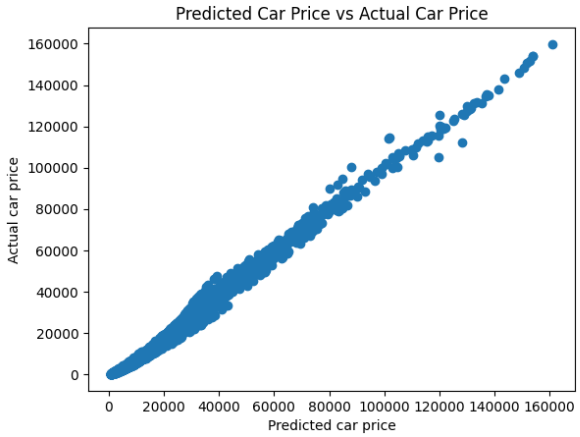

Figure 3.9 ANN Hyper Parameter Prediction Plot (model 2)

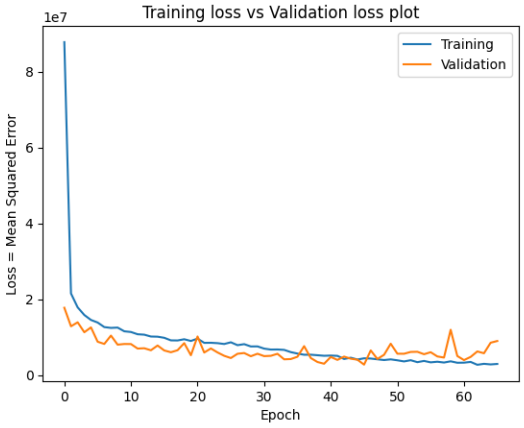

Figure 3.10 ANN HPLP (model 2)

4.5.4 ANN Hyperparameter Tunning 7 Layers Car price Prediction Plot

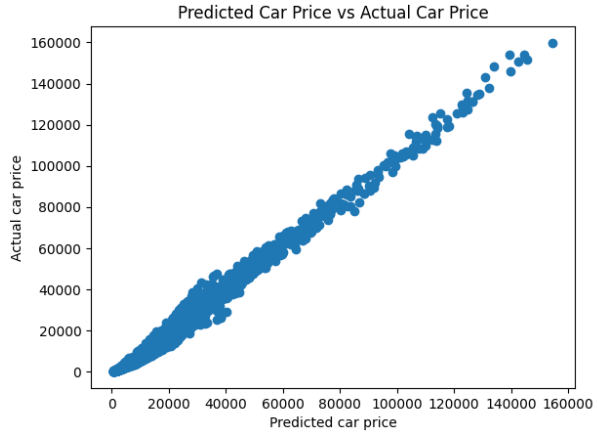

Figure 3.11 ANN Hyper Parameter Prediction Plot (model 3)

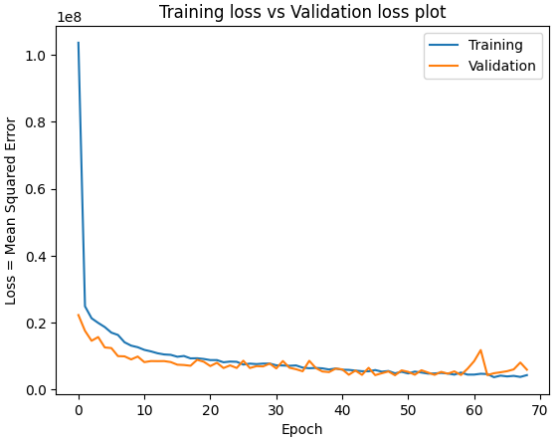

Figure 3.12 ANN HPLP (model 3)

4.5.5 ANN Hyperparameter (10% Dropout rate) Plot

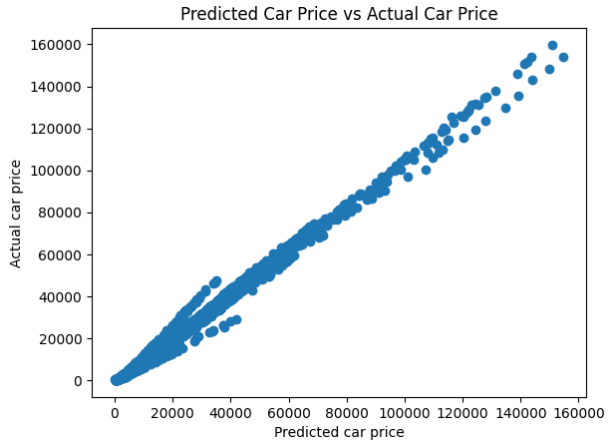

Figure 3.13 ANN Hyper Parameter Prediction Plot (model 4)

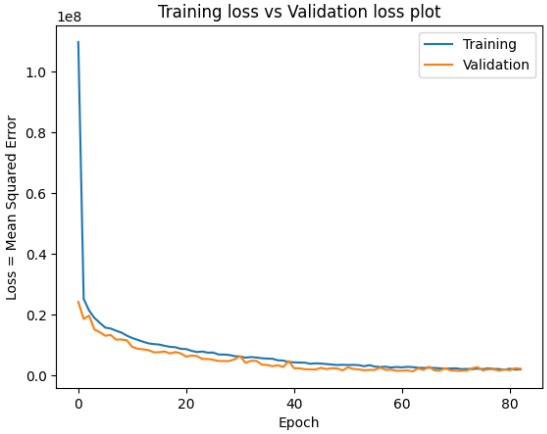

Figure 3.14 ANN HPLP (model 4)

4.6 K-Means Clustering Model

4.6.1 K-Means Elbow Point (Category 1)

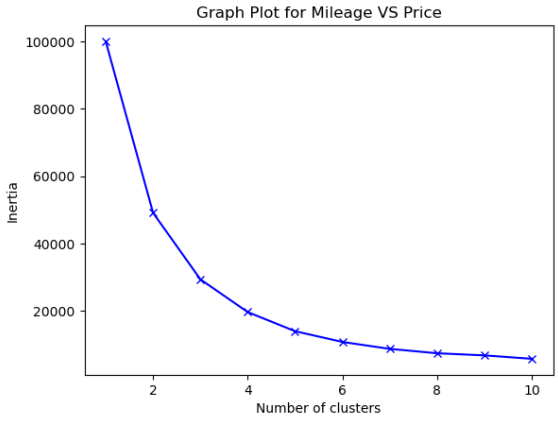

Figure 3.15 KNN Means (Category 1)

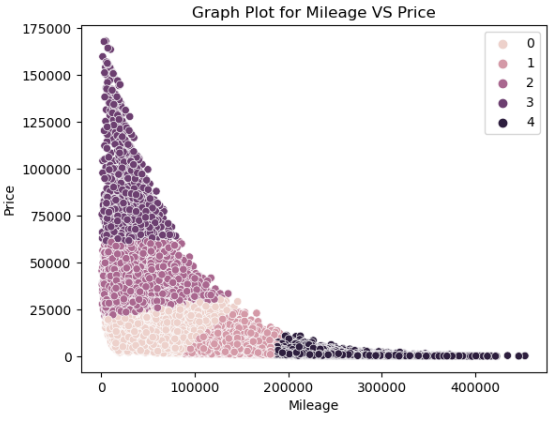

Figure 3.16 KNN Means Cluster Plot (Category 1)

4.6.2 K-Means Elbow Point (Category 2)

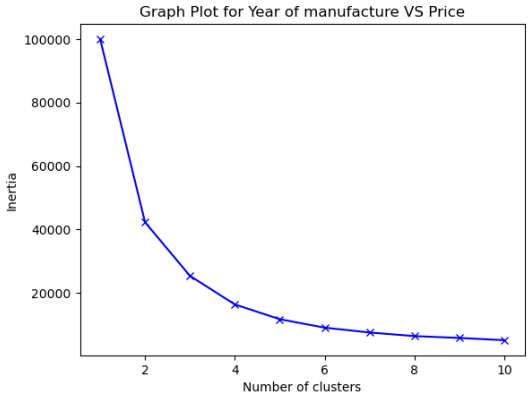

Figure 3.17 KNN Means (Category 2)

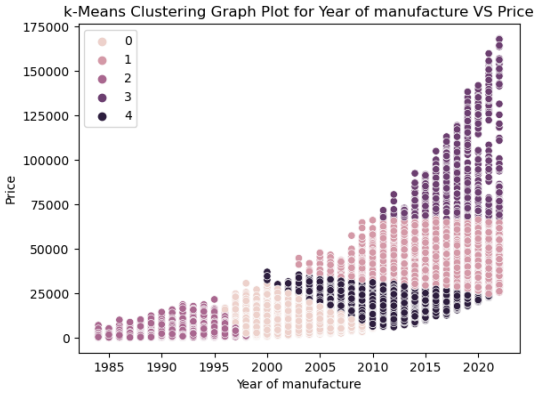

Figure 3.18 KNN Means Cluster Plot (Category 2)

4.6.3 K-Means Elbow Point (Category 3)

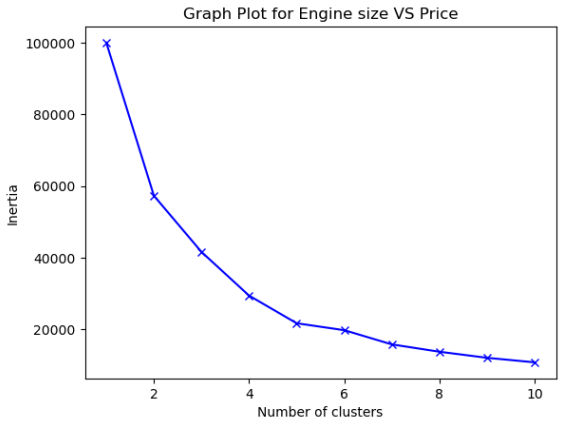

Figure 3.19 KNN Means Cluster Plot (Category 3)

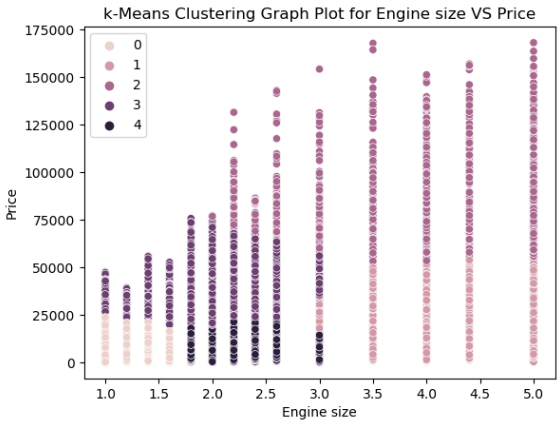

Figure 3.20 KNN Cluster Plot (Category 3)

Table 1.5 k-Means Clustering Results towards Price Prediction

Feature names	**K(n-clusters)**	**Davies Bouldin Score**	**Silhouette Score**
Mileage	5	0.648	0.453
Year of Manufacture	4	0.651	0.472
Engine size	5	0.755	0.432

4.7 Agglomerative Clustering Model

Table 1.6 Agglomerative Results towards Price Prediction

Feature names	**K**	**Davies Bouldin Score**	**Silhouette Score**
Mileage	5	0.551	0.398
Year of Manufacture	4	0.587	0.498
Engine size	5	0.690	0.586

4.8 DB Scan Clustering Model

Table 1.7 DB Scan Clustering Price Prediction Results

Feature names	**K**	**Davies Bouldin Score (DB)**	**Silhouette Score (SS)**
Mileage	5	0.584	0.809
Year of Manufacture	4	0.778	0.708
Engine size	5	2.723	0.060

5.0 DISCUSSION

From figure 2.0, the heat map shows the corelation between the numerical features, it shows that Mileage (-0.63), Year of Manufacture (0.71) and Engine size (0.4) corelates similarly to the car price.

Table 1.0 shows the result of the Simple linear regression model to predict car prices using single numerical input, the Year of Manufacture has the highest R^2 value at 0.5111. This is significantly higher than the R^2 values for Mileage and Engine Size, making the Year of Manufacture a more reliable indicator for predicting car prices with this model. From the non-linear 2nd degree polynomial model Table 1.1, the Year of Manufacture is the best predictor for the car price, as it has the highest R^2 value of 0.6094, compared to the simple linear regression models, this indicates that the relationship between the input features and to predict the car price is better modelled with a non-linear approach.

Following the Multiple Linear Regression (MLR) model in Table 1.2 using all three inputs Mileage, Engine Size, and Year of Manufacture appears to be a good predictor of the car price, with an R^2 of 0.6715. This suggests that the combination of these three features provides a more accurate prediction than any of the individual features for the prediction. Table 1.3 compares the Random Forest (RF) model using all 3 numerical features. This model is a good predictor of the car price with an R^2 value of 0.9257(92.5%) compared to single and multiple linear models. The RF was also compared using numerical and categorical variables and with an R^2 value of 0.9982(99.8%) which shows a better accurate prediction than all the other models and the previous numerical RF model.

Table 1.4 shows the ANN comparison and model 4 with a 10% dropout rate performed better in the car prediction with a high R^2 value of 0.992(99.2%) than the other models, not much overfitting was observed using this model as seen in section 4.5 in the ANN module. For the clustering models the Elbow point method was used to find the best k section 4.6 shows the visualization of the elbow plots. In Table 1.5, the k-Means clustering evaluation metrics scores show that Year of Manufacturing has a low DB score of 0.65 suggesting better clustering and the highest SS of 0.47 indicating tighter clustering and Mileage also has a low DB of 0.65 with an SS of 0.45 which is in close range with the year of manufacturing. In the Agglomerative Clustering Table 1.6 Year of Manufacture has the better SS of 0.4983 while Engine size has the highest DB score of 0.690. Table 1.7 shows the DB Scan's result showing Mileage has the better clustering with a DB score of (0.584) and SS of (0.809). Overall, the DB Scan clustering model performs well across all 3 clustering models.

6.0 CONCLUSIONS

The 2nd Random Forest Regression model stands out as the best predictor, with the highest R^2 value of 99.8% indicating a strong ability to capture both linear and non-linear relationships. The 4th ANN Model with a 10% dropout rate also does well with an R^2 value of 99.2%, especially in dealing with complicated data sets that are rather lacking behind the high performance of Random Forest. K-Means and Agglomerative clustering provides valuable data insights, especially with Year of Manufacture and Mileage, where it effectively identifies distinct clusters. But the DB Scan performs better with the Mileage as the better car price predictor.

REFERENCES

1. Abishek R (2022) *CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES. Www.Irjmets.Com @International Research Journal of Modernization in Engineering*, 45. Available online: www.irjmets.com.

2. Agrahari, K., Chaubey, A., Khan, M. & Srivastava, M. (2021) *Car Price Prediction Using Machine Learning*.

3. Dnyaneshwar, K., Tushar, J., Shivam, K. & Sagar, T. (2023) *Analysis of Car Selling Prediction Based On AIML. International Journal of Innovations in Engineering and Science*, 8(2). Available online: https://doi.org/10.46335/ijies.2023.8.2.3.

4. Gegic, E., Isakovic, B., Keco, D., Masetic, Z. & Kevric, J. (2019) *Car price prediction using machine learning techniques*. *TEM Journal*, 8(1), 113–118. Available online: https://doi.org/10.18421/TEM81-16.