Support Vector Regression (SVR) Model to Predict Car Price

Edwin Ario Abdiwijaya
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
edwin.abdiwijaya@binus.ac.id

Nathaniel Orion
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
nathaniel.orion@binus.ac.id

Peter Samuel Lim
Japanese Department
Faculty of Humanities
Bina Nusantara University
Jakarta, Indonesia 11480
peter.lim001@binus.ac.id

Johannes Simatupang
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
johannes.simatupang@binus.ac.id

***Abstract*—Many technologies have aided humans in their everyday lives. Artificial Intelligence (AI) has become a trend in the technology industry, especially Machine Learning (ML). ML has capabilities to predict things such as predict price. The model utilizes the Min-Max scaler used to scale data, then fitted by Support Vector Regression (SVR). The model then calculates the model's error. The calculations show that the model performs greatly.**

Keywords—Machine Learning, Car, Predict, Support Vector Regression, Radial Basis Function

I. INTRODUCTION

Technology has become an essential part of everyday life. Many new technologies are emerging that can aid human activities in everyday life and at work. Phones, heart rate monitors, and other similar technologies are examples. Over time, technology evolves and becomes more capable of assisting humanity, one of which is artificial intelligence, also known as AI. The presence of AI in humanity plays a role in solving problems that humans face, such as predicting car prices. This problem can be solved using one of AI's domains, namely Machine Learning (ML).

In this paper, the authors used Machine Learning to create a model to predict the price of a car. The model is implemented in Python version 3.10.2. Next will be explained the methodology, result, and conclusion.

II. METHODOLOGY

A. Dataset Preparation

The data can be found on Kaggle. The data was downloaded to a local computer from https://www.kaggle.com/datasets/adhurimquku/ford-car-price-prediction [2].

The model, year, price, transmission, mileage, fuel type, tax, mpg (miles per gallon), and engine size are all included in the dataset. The dataset is shown in Figure 1.

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
0	Fiesta	2017	12000	Automatic	15944	Petrol	150	57.7	1.0
1	Focus	2018	14000	Manual	9083	Petrol	150	57.7	1.0
2	Focus	2017	13000	Manual	12456	Petrol	150	57.7	1.0
3	Fiesta	2019	17500	Manual	10460	Petrol	145	40.3	1.5
4	Fiesta	2019	16500	Automatic	1482	Petrol	145	48.7	1.0
...	...	...	...	...	...	...	...	...	...
17961	B-MAX	2017	8999	Manual	16700	Petrol	150	47.1	1.4
17962	B-MAX	2014	7499	Manual	40700	Petrol	30	57.7	1.0
17963	Focus	2015	9999	Manual	7010	Diesel	20	67.3	1.6
17964	KA	2018	8299	Manual	5007	Petrol	145	57.7	1.2
17965	Focus	2015	8299	Manual	5007	Petrol	22	57.7	1.0

Fig. 1. Dataset

The authors add one more piece of data to the dataset that serves as a placeholder for the user's input when they use the website. The updated dataset is shown in Figure 2

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
0	Fiesta	2017	12000	Automatic	15944	Petrol	150	57.7	1.0
1	Focus	2018	14000	Manual	9083	Petrol	150	57.7	1.0
2	Focus	2017	13000	Manual	12456	Petrol	150	57.7	1.0
3	Fiesta	2019	17500	Manual	10460	Petrol	145	40.3	1.5
4	Fiesta	2019	16500	Automatic	1482	Petrol	145	48.7	1.0
...	...	...	...	...	...	...	...	...	...
17962	B-MAX	2014	7499	Manual	40700	Petrol	30	57.7	1.0
17963	Focus	2015	9999	Manual	7010	Diesel	20	67.3	1.6
17964	KA	2018	8299	Manual	5007	Petrol	145	57.7	1.2
17965	Focus	2015	8299	Manual	5007	Petrol	22	57.7	1.0
17966	Fiesta	2017	12000	Automatic	15944	Petrol	150	57.7	1.0

Fig. 2. Updated Dataset

B. Data Processing

In the data preprocessing section, the authors do data cleaning by checking if there is any null data in the dataset. The authors use a heatmap to determine the data correlation. The authors print the correlation value to be more precise. The updated dataset heatmap is shown in Figure 3.

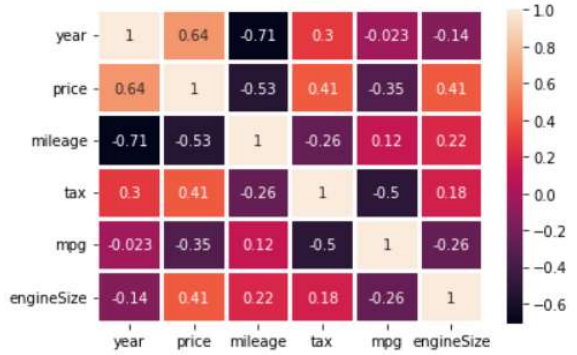

Fig. 3. Updated Dataset Heatmap

And the updated dataset correlation values are shown in Figure 4.

	year	price	mileage	tax	mpg	engineSize
year	1.000000	0.636009	-0.707816	0.298505	-0.022967	-0.137312
price	0.636009	1.000000	-0.530656	0.406851	-0.346419	0.411173
mileage	-0.707816	-0.530656	1.000000	-0.260460	0.120075	0.215047
tax	0.298505	0.406851	-0.260460	1.000000	-0.502976	0.184332
mpg	-0.022967	-0.346419	0.120075	-0.502976	1.000000	-0.260516
engineSize	-0.137312	0.411173	0.215047	0.184332	-0.260516	1.000000

Fig. 4. Updated Dataset Correlation Values

The authors will then look at the data types to see whether the authors can transform the object type into a category type. To scale the dataset, the authors utilize a min-max scaler. The data is scaled from 0 to 1. The scaled dataset is shown in Figure 5.

	model	year	price	transmission	mileage	fuelType	\
0	0.208333	0.328125	0.211101	0.0	0.089747	1.0	
1	0.250000	0.343750	0.247798	0.5	0.051125	1.0	
2	0.250000	0.328125	0.229450	0.5	0.070113	1.0	
3	0.208333	0.359375	0.312018	0.5	0.058877	1.0	
4	0.208333	0.359375	0.293670	0.0	0.008337	1.0	
...	...	...	...	...	...	...	
17962	0.000000	0.281250	0.128514	0.5	0.229106	1.0	
17963	0.250000	0.296875	0.174385	0.5	0.039456	0.0	
17964	0.458333	0.343750	0.143193	0.5	0.028180	1.0	
17965	1.000000	0.296875	0.143193	0.5	0.028180	1.0	
17966	0.958333	0.328125	0.211101	0.0	0.089747	1.0	

	tax	mpg	engineSize
0	0.258621	0.203867	0.20
1	0.258621	0.203867	0.20
2	0.258621	0.203867	0.20
3	0.250000	0.107735	0.30
4	0.250000	0.154144	0.20
...	...	...	...
17962	0.051724	0.203867	0.20
17963	0.034483	0.256906	0.32
17964	0.250000	0.203867	0.24
17965	0.037931	0.203867	0.20
17966	0.258621	0.203867	0.20

Fig. 5. Scaled Dataset

The authors then split the scaled data into X and y for the independent and dependent variables. X contains independent variables, whereas y contains dependent variables. The split method does not randomize the data. The scaled dataset heatmap is shown in Figure 6.

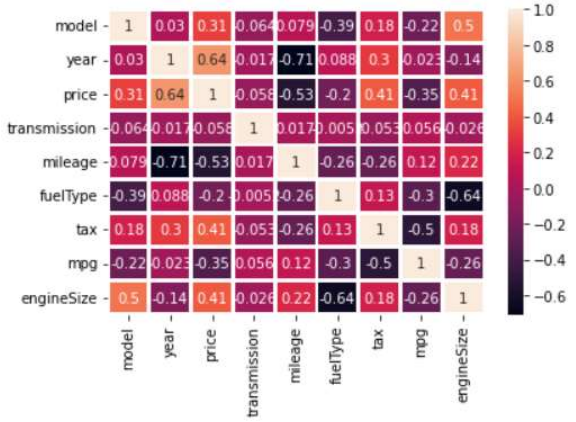

Fig. 6. Scaled Dataset Heatmap

And the scaled dataset correlation values are shown in Figure 7.

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
model	1.000000	0.030228	0.314551	-0.063947	0.078626	-0.388338	0.183507	-0.221851	0.499909
year	0.030228	1.000000	0.636009	-0.017127	-0.707816	0.087955	0.298505	-0.022967	-0.137312
price	0.314551	0.636009	1.000000	-0.058013	-0.530656	-0.202855	0.406851	-0.346419	0.411173
transmission	-0.063947	-0.017127	-0.058013	1.000000	0.016702	-0.005218	-0.053382	0.056076	-0.026360
mileage	0.078626	-0.707816	-0.530656	0.016702	1.000000	-0.257567	-0.260460	0.120075	0.215047
fuelType	-0.388338	0.087955	-0.202855	-0.005218	-0.257567	1.000000	0.129605	-0.297793	-0.644630
tax	0.183507	0.298505	0.406851	-0.053382	-0.260460	0.129605	1.000000	-0.502976	0.184332
mpg	-0.221851	-0.022967	-0.346419	0.056076	0.120075	-0.297793	-0.502976	1.000000	-0.260516
engineSize	0.499909	-0.137312	0.411173	-0.026360	0.215047	-0.644630	0.184332	-0.260516	1.000000

Fig. 7. Scaled Dataset Correlation Values

C. Support Vector Regression

Support Vector Regression (SVR) is from a Support Vector Machine (SVM). SVR is useful for predicting the regression of linear and non-linear data [1]. The authors use the sklearn.svm library to import SVR, and with SVR, the authors use the kernel Radial Basis Function (RBF). The regressor variable contains the SVR and belows are how the regressor defined,

regressor = SVR(kernel='rbf', degree=3, gamma='scale', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200, verbose=False, max_iter=- 1)

The regressor variable will be used to fit the processed dataset. The authors then anticipate the test set data and store it in a separate dataframe, allowing them to compare actual and predicted data. The actual data price values and predicted data price values are shown in Figure 8.

	Actual	Prediction
0	0.064294	0.114764
1	0.247211	0.231302
2	0.103633	0.090138
3	0.228385	0.256441
4	0.119339	0.153405
...	...	...
3589	0.128514	0.142317
3590	0.174385	0.242808
3591	0.143193	0.230201
3592	0.143193	0.516369
3593	0.211101	0.454186

Fig. 8. Actual Data Price Values and Predicted Data Price Values

D. Model Performance

To evaluate the model performance, the authors calculate the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. The result calculations are shown in table 1.

TABLE I. RESULT CALCULATIONS

Data	MAE	MSE	RMSE	R^2 Score	Accuracy
Price	0.0336	0.0021	0.0457	0.7245	72.45%

The authors also plot the data to show visually the comparison between actual data and the predicted data. The plot is shown in Figure 9.

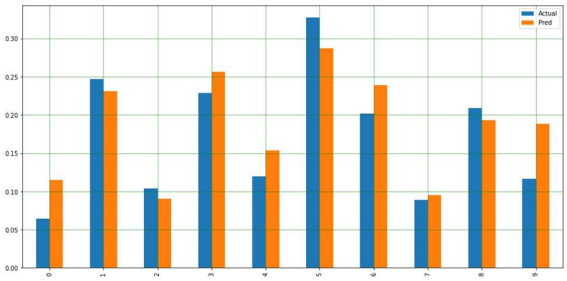

Fig. 9. Comparison Between Actual Data and The Predicted Data

E. Predicting

The authors create an interactive website with Gradio so that people can try the website and predict the car based on user inputs. Users can also change the train set percentage, from default 0.2 to range from 0.01 to 0.99. The website is shown in Figure 10.

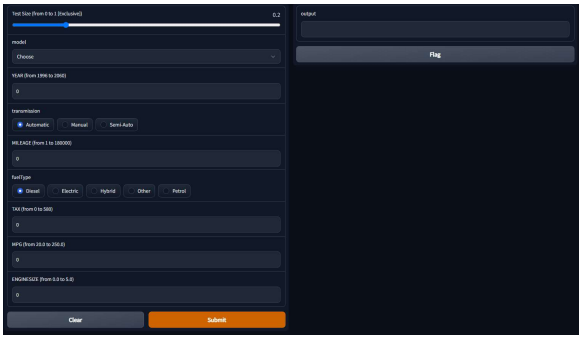

Fig. 10. Comparison Between Actual Data and The Predicted Data

All the necessary code and dataset can be accessed through https://colab.research.google.com/drive/1Pzs5fUQTWh9lgpPJn6p3VAvXsRUvmTuj?usp=sharing

III. RESULT AND DISCUSSION

The model's performance results demonstrate that it performed well in predicting price because MAE, MSE, and RMSE are all less than 0.05 or 5%, and the R^2 score or accuracy is 0.7245 or 72.45%. Figure 9 demonstrates that out of the first ten data points, about two do not predict well. The authors additionally tested the website's operation and discovered that all of the functions work as expected.

IV. CONCLUSION

The proposed methodology is used to predict the price of cars. To scale the data, the approach utilizes a Min-Max scaler. The scaled data was then fitted by using the SVR kernel RBF. Finally, MAE, MSE, RMSE, and R^2 values are used to assess the model. The authors find that the model does an amazing job. The authors aim to obtain new datasets with more data in the future so that the authors can improve the model.

REFERENCES

[1] Parbat, D., & Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. Chaos, Solitons & Fractals, 138, 109942. Available from https://doi.org/10.1016/j.chaos.2020.109942

[2] Quku, Adhurim. (2022, April). Ford Car Price Prediction, Version 1. Retrieved May 7, 2022 from https://www.kaggle.com/datasets/adhurimquku/ford-car-price-prediction