International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

Grid Search Optimized Machine Learning based Modeling of CO2 Emissions Prediction from Cars for Sustainable Environment

Sagar Sidana
Principal Software Engineer, McKinsey & Company.

ABSTRACT: Carbon emissions have increased dramatically because of industrialization, trapping heat in the atmosphere and hastening climate change. This is a serious threat to the wealth, security, and well-being of the world. The effects are extensive, ranging from severe weather, disease outbreaks, and economic disruption to food insecurity and water scarcity. The World Health Organization (WHO) has determined that climate change poses the greatest threat to public health in the twenty-first century. Thus, precise CO2 emissions have emerged as a crucial concern in recent times. Several studies have tried to forecast the amount CO2 from industry and power plant using statistical analysis. Efficiency, robustness and diverse application was the limitation of the study. In this study, we have proposed an AI based model that is able to predict the amounts of CO2 emissions from cars. We applied a grid search-optimized machine learning approach using the publicly available Canadian dataset. Incorporation of different statistical analyses and preprocessing techniques such as duplicate data management, outlier rejection, scaling contributed to enhance the quality of the dataset. Later, grid search techniques were applied to tune the KNN, RF, and SVR models. The approach has enhanced the performance of CO2 emissions prediction. In the study, we further used the explainability of the random forest model to check the bias and fairness of predictability. MSE, RMSE, and R-squared metrics of the proposed approach were the highest as the state of the art.

KEYWORDS: Feature Selection, Grid search, Random Forest, Tunning.

I. INTRODUCTION

Climate change is a worldwide hazard to all living things, not only an environmental one. Its effects are widespread, complicating ecosystems, upsetting our food security [1], and making water scarcer [2]. El-Sayed and Kamel et. al. [3] have noted the development of new diseases, the increase in extreme weather occurrences, and the tremendous burden on public health systems. According to Hernández-Delgado et. al. [4], socioeconomic effects, including mass migration and unemployment, are also happening more frequently. Deforestation and the burning of fossil fuels cause the atmospheric carbon dioxide level to rise quickly [5]. The overall amount of greenhouse gas (GHG) emissions has increased by 78% since 1970 due to a 90% increase in CO2 emissions [6].

Research indicates that the global energy system is one of the primary human-caused sources of CO2 emissions. Due to the fact that countries depend heavily on energy for development, energy is a difficult issue [7]. Hence, energy system transformation to lower GHG emissions is the main focus of climate change mitigation efforts. But these emissions are also influenced by other social and economic factors [8].

A multifaceted strategy is required to solve the issue. To identify emission sources at different levels (individual, community, and national), forecast GHG concentrations, and use data science to comprehend trends and possible mitigation tactics, we require precise models [8]. We cannot hope to prevent climate change and preserve life on Earth without taking such a comprehensive approach. CO2 emissions have increased by about 90% since 1970 and now make up an astounding 78% of all greenhouse gases (GHGs). Forecasting future greenhouse gas emissions is a challenging undertaking because of numerous dynamic aspects, including socioeconomic patterns and carbon emissions. This intricacy highlights how important but difficult accurate forecasting is.

Recent years have seen the rise of artificial intelligence (AI) and machine learning (ML) as effective methods for deciphering complicated environmental events, especially those that show notable temporal and spatial fluctuations. In this research, we have proposed a grid search-optimized machine learning model for accurate forecasting of CO2 emissions from cars. We have used the publicly available Canadian CO2 emission dataset, which is prepared by the Canadian government. The limitations and scientific gap of the previous study have been resolved. We incorporated different pre-processing and statistical techniques to improve the

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

data patterns. Later, three machine learning models, such as KNN, RF, and SVR, were prioritized using grid search techniques. The proposed approach achieved the highest performance. A comparative analysis was carried out. The contribution can be highlighted as follows:-

- Local outlier factor analysis
- Duplicate and missing data handling
- Grid search optimization
- Comparative analysis
- Proposing a high performance AI model for CO2 emissions prediction.
- Application of explainability for the fairness of the model's predictionability.

The proposed approach has resolved the limitation of previous studies on the same dataset, even though the though the number of studies for CO2 emissions from cars is limited.

In Section II, review of the literature is provided. Section III provides illustrations of the materials and techniques. The discussion and conclusion are covered in Section IV. Section V wraps up the analysis and suggests further investigation.

II. LITERATURE SURVEY

With the evolution of artificial intelligence and machine learning algorithms, researchers are adopting this technology for CO2 emissions. Several studies have been conducted on CO2 emissions prediction by machine learning algorithms.

Baky et. al. [9] proposed a deep learning, SVM, and ANN approach to forecast greenhouse gas emissions from electricity production in Turkey. The study was evaluated using RMSE, MBE, RMSE, R2, and MAPE. Using data from the US Environmental Protection Agency (EPA), a model that uses artificial neural networks to predict the emissions of CO2 and NOx from heavy-duty trucks was developed. While the outcome is encouraging, CO2 needs to be considered, and this model works with gasoline-powered vehicles [10].

Madziel et al. [11] provided a computationally efficient approach for developing CO2 emission models, and the models produced produce workable outcomes for fully hybrid cars. The initial stage in developing an emission model is to review the accuracy results of methods such as neural networks (NNET), cubic support vector machines (SVM), bagged trees, linear, robust regression, fine, medium, and coarse trees, and neural machines.

Evin GARİP et. al. [12] proposed SVM and RF for estimating Turkey's CO2. It has been noted that the support vector machine approach yields more accurate forecasts. The study was only based on turkey vehicles with a small number of cars and a narrow dataset.

Ahmed et al. [13] projected Saudi Arabia's yearly CO2 emissions through 2030 using long short-term memory (LSTM), an adaptive network-based fuzzy inference system (ANFIS), and a feed-forward neural network (FFNN). In contrast to the R2 averages of 0.990985, 0.98875, and
0.9945, respectively, the RMSE averages were 19.78,
20.89505, and 15.42295, respectively.

Another study's goal [14] was to use AI and machine learning to anticipate CO2 emissions. Using SARIMA (SARIMAX), a model for prediction based on ARIMA, the researchers created four models. The COVID-19 pandemic's peak time serves as the basis for the models. The study projects the total CO2 emissions worldwide for the following three time periods: the near future, 2022 to 2027, the future, 2022 to 2054, and the long future, 2022 to 2072. For comparing accuracy, the mean absolute percentage error, or MAPE, is used. For the years 2022 to 2027, the post-COV estimates for the total world CO2 emissions are 36,218.59, 36,733.69, 37,238.29, 37,260.88, 37,674.01, and 37,921.47 million tons (MT).

Cosimo Magazzino et al.'s study [15] used data from
1970 to 2017 to investigate the connection between Russia's GDP, energy consumption, and CO2 emissions. The outcomes were compared using time-series analysis and a novel D2C technique. The study found that energy consumption and CO2 emissions are influenced by economic growth, but it also recommended using alternative energy sources to cut emissions.

To cut CO2 emissions by 40–45% by 2030, Yuvaraj et al. [16] created algorithms to estimate emissions on a sizable sensor-based dataset of 7384 light-duty cars. Even with a single car attribute as input, the suggested method, categorical boosting

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

(Catboost), successfully forecasted CO2 emissions. This method gave producers and consumers insightful information on air pollution caused by vehicles, along with suggestions. The study's MSE and R-squared values were 3.83 and 0.996, correspondingly.

The literature review showed that a few studies have been accomplished on CO2 emissions. Most of the study was carried out on industry and electrical power generationbased CO2 emissions prediction. In our study, we have addressed the CO2 emission prediction from vehicles. One study was carried out to predict CO2 emissions from Canadian vehicles. But the performance of the approach was not very impressive. There was scope to further improve the performance of the model. In our study, we focused on it to enhance the performance of CO2 emission prediction from cars. Our approach has achieved the highest level of accuracy. The pre-processing and parameter tuning have been applied to this standard. Feature importance and comparative analysis were also applied, which proved that our approach is more efficient and sustainable for CO2 emissions from cars.

III. MATERIALS AND METHODOLOY

A. Dataset Collection

The study has used a publicly available CO2 emissions dataset from Canada [17]. The dataset was prepared by the Canadian government for statistical analysis of the carbon emissions of different vehicles. It contains details of how a vehicle's CO2 emissions might change depending on several attributes. This includes information spanning seven years. In total, there are 12 columns with features and 7385 rows with instances. The features of the dataset have been shown in Table I.

Table I. Features Name and Description of Dataset.

S.N.	Features Name	Description
1	Make	Company of the vehicle
2	Model	Car model
3	Vehicle Class	Class of vehicle depending on their utility, capacity and weight
4	Engine Size	Size of engine used in Litre
5	Cylinders	Number of cylinders
6	Transmission	Transmission type with number of gears
7	Fuel Types	Type of Fuel used
8	Fuel Consumption City	Fuel consumption in city roads (L/100 km)
9	Fuel Consumption Hwy	Fuel consumption in Hwy roads (L/100 km)
10	Fuel Consumption Comb	The combined fuel consumption (55% city, 45% highway) is shown in L/100 km
11	Fuel Consumption Comb mpb	The combined fuel consumption in both city and highway is shown in mile per gallon(mpg)
12	CO2 Emissions(g/km)	Target Variable

B. Preprocessing and Exploratory Data Analysis

Preprocessing is a vital step in machine learning-based prediction model analysis [17]. Similarly, exploratory data analysis also involves conducting preliminary research to identify trends, identify abnormalities, test theories, and verify presumptions using graphical and summary statistics.

In our study, the missing values were analyzed using the pandas isnull() and sum() functions. No missing values were found. The duplicate values were investigated by the Pandas duplicated () function. The dataset had duplicate values of 979. We dropped all the duplicate values. The outliers were handled by the local outlier factor method. The distribution of “make” features is shown in Fig. 1. Normalized Fuel Consumption, Combined Mile Per Gallon, and CO2 Emissions Distribution are also depicted for easy understanding of the feature distribution in Figs. 2 and 3. A statistical analysis has been carried out on the data set that is shown in Fig. 4. The average CO2 emissions of the vehicles are 250.584699 g/km, with a maximum of 522 g/km and a minimum of 96 g/km. The correlation analysis of the features is illustrated in Fig. 5.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

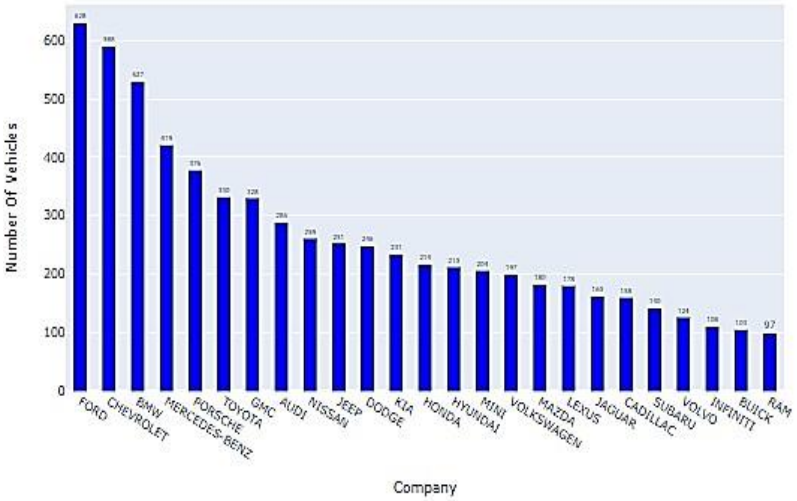

Fig.1. Top 25 companies and their number of vehicles in the dataset.

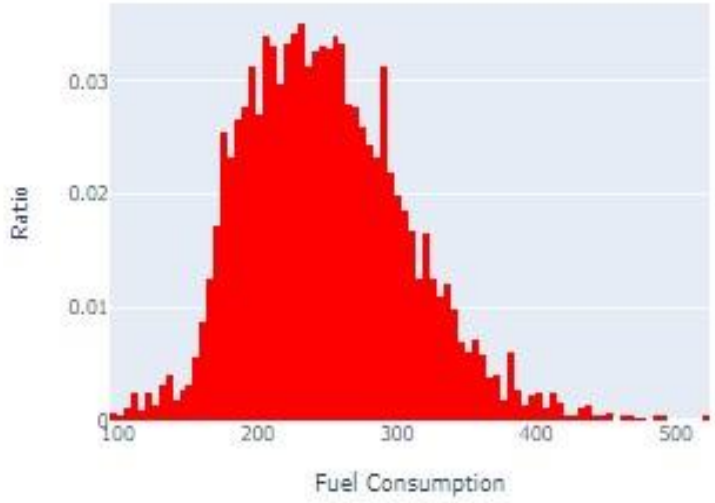

Fig.2. Fuel consumption combined per mile gallon.

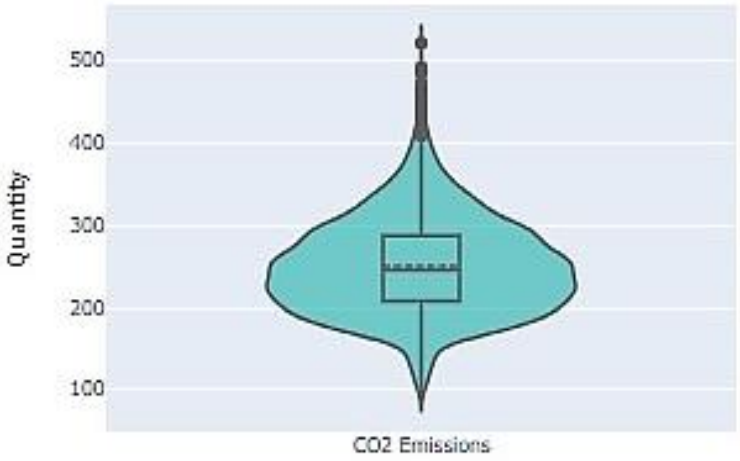

Fig.3. CO2 emissions distribution.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

	count	mean	std	min	25%	50%	75%	max
Engine Size(L)	7385.0	3.160068	1.354170	0.9	2.0	3.0	3.7	8.4
Cylinders	7385.0	5.615030	1.828307	3.0	4.0	6.0	6.0	16.0
Fuel Consumption City (L/100 km)	7385.0	12.556534	3.500274	4.2	10.1	12.1	14.6	30.6
Fuel Consumption Hwy (L/100 km)	7385.0	9.041706	2.224456	4.0	7.5	8.7	10.2	20.6
Fuel Consumption Comb (L/100 km)	7385.0	10.975071	2.892506	4.1	8.9	10.6	12.6	26.1
Fuel Consumption Comb (mpg)	7385.0	27.481652	7.231879	11.0	22.0	27.0	32.0	69.0
CO2 Emissions(g/km)	7385.0	250.584699	58.512679	96.0	208.0	246.0	288.0	522.0

Fig.4. Statistical analysis of the dataset.

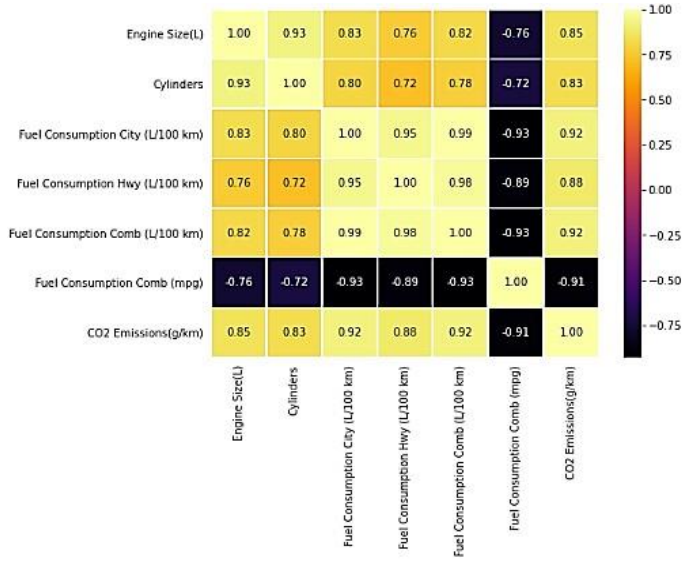

Fig. 5.Correlation analysis of the dataset.

C. Proposed Approach

In this study, we proposed an efficient approach to predicting CO2 emissions from different vehicles. The collected dataset is trustworthy and publicly available. In this approach, we introduced modern pre-processing techniques. Missing data management, duplicate data handling, outlier rejection by local outlier factor strategy, and scaling of the data are the pre-processing steps. We found 979 duplicates of data that was dropped in our study. The dataset was widely exploratory analyzed for a better understanding of the distribution of features. The correlation analysis among the features also contributed to the relationship among them. Later, the dataset was split into training and testing at a ratio of 75:25. We applied three machine learning models, such as random forest (RF), Knearest neighbor (KNN), and support vector regression (SVR). The model was hyperparameter-tuned with a grid search technique. The grid space is shown in Table II. From the grid search, we chose the best parameters. Using the parameter, we trained the models and tested them with testing data. The model was evaluated by mean square error, root mean square error, mean absolute error, and values. We carried out a comparison with the tuned and untuned versions of the model. The outcome of the study has revealed that the performance of our approach is the highest compared to previous.

D. Model Evalution

Our proposed model was evaluated by the prominent parameters. The parameters and their respective equations are given as follows as-

Mean square error: The average square error between predicted value and actual value is determined by MSE. It indicates how a model predicts accurately. It can be formulated as follows as-

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

$$MSE = \frac{1}{v}\sum_{i=1}^{v}(o_i - \hat{o}_i)^2$$

Root means square error: The square root of the average of the squared differences between paired observations that reflect the same occurrence is represented by the statistical measure known as the root mean square error (RMSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{v}(o_i - \hat{o}_i)^2}{v}}$$

Mean absolute error: Measuring the average amount of mistakes between paired observations of the same phenomena, the mean absolute error (MAE) is a statistical metric.

$$MAE = \frac{\sum_{i=1}^{n} |e_i|}{v}$$

R-squared: The percentage of the dependent variable's variance that can be predicted from the independent variables is expressed statistically as R-squared (R^2). It shows how well a regression model fits data; a number nearer 1 denotes a better fit.

$$R^2 = 1 - \frac{\sum_i (o_i - o_i)^2}{(o_i - o_i)^2}$$

Where, v is the number of measurements. o is the $i-th$ measurement and $\hat{o_i}$ is the $i-th$ prediction values.
$|o_i - \hat{o_i}| = |e_i|$

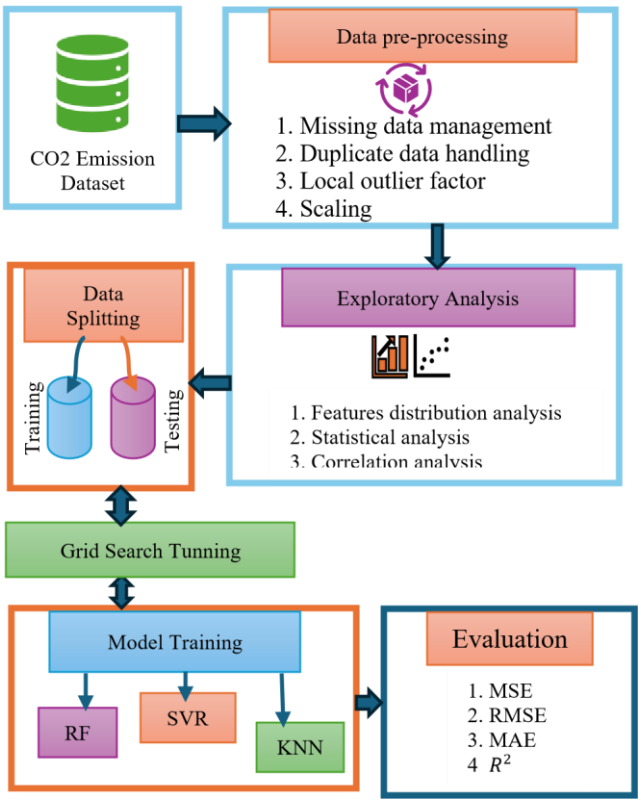

Fig.6. Proposed CO2 emissions approach.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

Table II. Grid Space of the Models.

Model Name	Tuned Parameters
KNN	'n_neighbors': array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29]
SVR	"C": [0.01, 0.1,0.4,5,10,20,30,40,50]
RF	'max_depth': list(range(1,10)), 'max_features': [3,5,10,15], 'n_estimators' : [100, 200, 500, 750

IV. RESULT AND DISCUSSION

The study has been implemented in Python using the Pandas, NumPy, Matplotlib, and machine learning libraries. The grid search optimization of the study provided the bestfitting parameters of the models. The outcome of the grid search of the models is listed in Table III. The study was conducted on both tuned and untuned models. The results of the untuned model have been shown in Table IV. The MSE, RMSE, and MAE of the SVR are very high. But KNN and RF performed well. The RF models provided the highest values of 0.995, and KNN is the second with 0.995. The outcomes of the models after tunning are listed in Table V. The performance of SVR increased remarkably. The MSE and RMSE were reduced from 16.76 to 3.86 and increased to 0.995. But the performance of RF is a is a little bit degraded.

The results of the analysis showed that the RF without hypermeter tuning performed the best. SVR and KNN, after hyperparameter tuning, also improved their performance. Advanced preprocessing techniques and hypermeter tuning contributed to enhancing the performance. A comparison with the previous study shown in Table VI also presents the superior performance of our approach. The explainability of the RF model has been shown in Fig. 7. The importance of the importance of the features also added to the fairness of our approach. Finally, we conducted a comparison with the actual and predicted values in Fig. 8. The RF model predicted more closely with the actual values.

Table III. Grid Search Parameters of the Models.

Model Name	Tuned Parameters
KNN	n_neighbors=3
SVR	C=50
RF	max_depth = 9, max_features = 5, n_estimators =750

Table IV. Performance Metrics of the Model before Tunning.

Model Name	MSE	RMSE	MAE	R^2
KNN	4.26	4.26	2.90	0.995
SVR	16.76	16.76	11.17	0.916
RF	3.13	3.13	2.33	0.997

Table V. Performance Metrics of the Model after Tunning.

Model Name	MSE	RMSE	MAE	R^2
KNN	4.05	4.05	2.81	0.995
SVR	3.86	3.86	2.63	0.995
RF	3.19	3.19	2.41	0.997

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

Table VI. Comparison with the Previous Work.

Ref.	Model	MSE	MAE	R-Square
[15]	SVR	4.14	2.64	0.994
	CatBoost	3.83	2.41	0.996
Our proposed	Tunned SVR	3.86	2.63	0.995
	RF	3.13	2.33	0.997

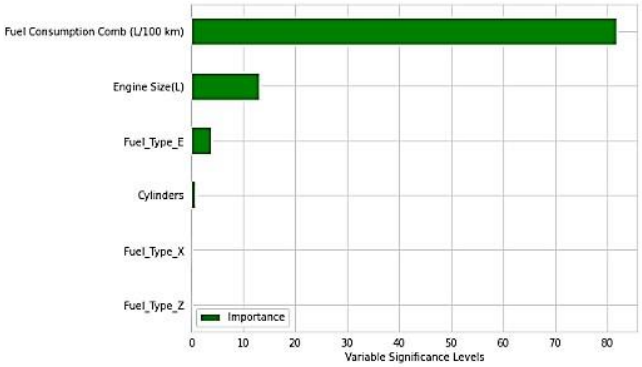

Fig.7. Feature importance of RF model.

Real Value	Pred KNN	Pred SVR	Pred RF
193,193	211	207.93	209.68
265,265	305	301.97	298.4
270,270	288.33	292.82	267.47
338,338	399	403.32	400.12
175,175	175.33	181.35	179.44
342,342	395	394.09	387.24
214,214	229.33	234.39	233.13
314,314	370	363.57	362.1
313,313	357.67	360.58	359.52
120,120	109.67	110.85	108.87
207,207	221.67	223.63	223.89
239,239	266.67	265.32	264.88
181,181	187.67	187.1	187.8
165,165	161	161.81	162.58
258,258	283.33	283.29	281.06
191,191	203.33	203.37	203.49

Fig.8. Comparison of the model in predicting CO2 emissions.

V. CONCLUSION AND FUTURE WORK

Efficient and grid-search-optimized SVR, KNN, and RF models have been implemented for CO2 emissions prediction from previous seven-year vehicle data. The study has addressed all the previous limitations. Integration of preprocessing strategies such as duplicate data handling, missing data management, outlier rejection, and hypermeter tuning contributed greatly to the model's performance. The MSE, MAE, and R-squared of SVR have been enhanced from the previous study. Our approach has achieved the highest R-square score of 0.997 and the lowest MSE of 3.13, with a with a MAE of 2.33.

However, the study will contribute greatly to the prediction of CO2 emissions from different cars. It will help in building a sustainable environment and a living earth.

The study has several limitations, for instance, the lack of a dataset and the fact that the fact that all models of cars are not considered. The future researcher can extend the work.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 07 Issue 09 September 2024
DOI: 10.47191/ijcsrr/V7-i9-37, Impact Factor: 7.943
IJCSRR @ 2024
www.ijcsrr.org

REFERENCES

1. Wiebe, K., Robinson, S., & Cattaneo, A, "Climate change, agriculture and food security: impacts and the potential for adaptation and mitigation", Sustainable food and agriculture, 55-74, 2019.
2. Ismail, Z., & Go, Y. I., "Fog-to-Water for Water Scarcity in Climate-Change Hazards Hotspots: Pilot Study in Southeast Asia. Global Challenges", Vol. 5, pp. 2000036, 2021.
3. El-Sayed, A., & Kamel, M, "Climatic changes and their role in emergence and re-emergence of diseases". Environmental Science and Pollution Research, Vol. 27, pp. 22336-22352, 2020.
4. Hernández-Delgado, E. A, "The emerging threats of climate change on tropical coastal ecosystem services, public health, local economies and livelihood sustainability of small islands: Cumulative impacts and synergies", Marine Pollution Bulletin, Vol. 101, pp. 5-28, 2015.
5. Dyson, F. J., "Can we control the carbon dioxide in the atmosphere?". Energy, Vol. 2, pp. 287-291, 1977.
6. McKinnon, A., & Piecyk, M., "Measuring and managing CO2 emissions. Edinburgh" European Chemical Industry Council, 2010.
7. Lin, B., & Li, X..,"The effect of carbon tax on per capita CO2 emissions. Energy policy", Vol. 39, pp. 5137-5146, 2011.
8. Böttcher, H., Eisbrenner, K., Fritz, S., Kindermann, G., Kraxner, F., McCallum, I., & Obersteiner, M., "An assessment of monitoring requirements and costs of'Reduced Emissions from Deforestation and Degradation", Carbon balance and management, Vol. 4, pp. 1-14, 2009.
9. Bakay, M. S., & Ağbulut, Ü., "Electricity production based forecasting of greenhouse gas emissions in Turkey with deep learning, support vector machine and artificial neural network algorithms", Journal of Cleaner Production, Vol. 285, pp. 125324, 2021.
10. Tóth-Nagy, C., Conley, J. J., Jarrett, R. P., & Clark, N. N., " Further validation of artificial neural network-based emissions simulation models for conventional and hybrid electric vehicles", Journal of the Air & Waste Management Association, Vol. 56, pp. 898-910, 2006.
11. Mądziel, M., Jaworski, A., Kuszewski, H., Woś, P., Campisi, T., & Lew, K., "The development of CO2 instantaneous emission model of full hybrid vehicle with the use of machine learning techniques" Energies, Vol. 15, pp. 142, 2021.
12. E. GARİP and A. B. OKTAY, "Forecasting CO2 Emission with Machine Learning Methods," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), pp. 1-4 Malatya, Turkey, 2018.
13. Nassef, A. M., Olabi, A. G., Rezk, H., & Abdelkareem, M. A., "Application of Artificial Intelligence to Predict CO2 Emissions: Critical Step towards Sustainable Environment. Sustainability", Vol. 15, pp. 7648, 2023.
14. Meng, Y., & Noman, H., "Predicting co2 emission footprint using ai through machine learning". Atmosphere, Vol. 13, pp. 1871, 2023.
15. Magazzino, C., & Mele, M., "A new machine learning algorithm to explore the CO2 emissions-energy use-economic growth trilemma" Annals of Operations Research, pp. 1-19, 2022.
16. Natarajan, Y., Wadhwa, G., Sri Preethaa, K. R., & Paul, A., " Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms. *Electronics"* Vol. *12*, pp. 2288, 2023.
17. "CO2 Emission by Vehicles", Available link: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-byvehicles/data (Accessed on 24 June, 2024)
18. Talukder, M. S. H., & Sarkar, A. K., "Nutrients deficiency diagnosis of rice crop by weighted average ensemble learning", Smart Agricultural Technology, Vol. 4, pp. 100155, 2023.
19. Singla, P., & Verma, V., "Towards Personalized Job Recommendations: A Natural Language Processing Perspective", In 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), pp. 768-773, 2023.

Cite this Article: Sagar Sidana (2024). Grid Search Optimized Machine Learning based Modeling of CO2 Emissions Prediction from Cars for Sustainable Environment. International Journal of Current Science Research and Review, 7(9), 7199-7207. https://doi.org/10.47191/ijcsrr/V7-i9-37