Maximum Speed Prediction of Porsche Cars Using Vehicle Specifications and Machine Learning Models

Aswin Sathyan[†], Debopriya Das[†], Rand Kuoutaly[†,*], Ahmed Hassan[‡], Meerah Karunanithi[‡] and Syed Arslan Abbas Rizvi[‡]

[†] Dept. of Business, University of Europe for Applied Sciences, 14469 Potsdam, Germany.
[‡] Dept. of Computer Science and Information Technology, Berlin School of Business and Innovation(BSBI), Berlin, Germany
[*] Corresponding Author: student.uepotsdam@gmail.com

***Abstract*—The maximum speed of a vehicle especially cars is one the most important key parameters of interest for car manufacturers, researchers as well as enthusiasts. With this study, we aim to develop a predictive model to accurately predict the maximum speed that a Porsche 911 model can produce. For an accurate evaluation, we use some specifications of the vehicle such as engine displacement, Horsepower, year of manufacturing, acceleration, power-to-weight ratio, emission Standards, year of manufacturing, and many more from the models ever produced. Several predictive algorithms are used *, Linear Regression, Decision Tree, Gradient Boost Regression, and KNN model. Through this analysis it revealed that horsepower plays the most significant predictor of maximum speed. The outcome of this predictive analysis reflects valuable and important insights that help the automotive industry and designers, It also highlights the benefits of machine learning techniques for vehicle performance optimization.**

I. Introduction

In the automotive world, The performance ratings of super-powered cars are always at the top of the conversation for automotive enthusiasts and auto journalists [1], [2]. When we look into the automotive industry there is a tight competition among the car manufacturers to produce exotic cars with most maximum speed [3], [4]. So using a predictive tool for predicting the maximum speed would play a vital role among them [5], [6]. Porsche is among one of the car manufacturers who produce cars with extreme power that produce a great maximum speed which holds a place in the heart of every car enthusiast [7], [8]. One of the most iconic cars ever built is the Porsche 911 and here we are going to use a speed predictive model to predict and analyze the maximum speed of every Porsche 911 ever built [9], [10].

In this analysis, we are going to analyze the trends and patterns in Porsche's design and engineering over the years and compare and contrast different models and generations of Porsche cars. Going to build a predictive model for car valuation as well as performance. Here we will also create stunning visualizations to showcase the results as well. It is very significant to work on this project because it helps the car manufacturers and car designers to build the most efficient and low-emission vehicles without investing much on the car. And therefore it directly or indirectly helps the company to make gain more profits and utilize the resources and energy wisely.

II. Literature Review

There are many wide varieties of techniques that can be used for predicting analysis. For this project, the techniques that are used are Linear Regression, Decision Tree Regression, Random Forest, Gradient Boosting, and last but not the least KNN Algorithm.

The Linear Regression Model is a statistical technique used in industries to make predictions, business-decisions, and many more. It is a mathematical model and it describes a straight-line relationship between independent variables and dependent variables.

The Decision Tree Regression is a technique that predicts continuous values. It is a tree-structural classifier with three types of nodes. The initial node is the root Node and it represents the entire sample and also it splits into further nodes.

The Random forest is a type of machine learning algorithm and this is a commonly used algorithm.it is trademarked by Leo Breiman and Adele Cutler. This algorithm combines the multiple decision trees and displays a single result in the output. It is an extension of the bagging methods as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. It is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from the training set with replacement, called the bootstrap sample. It reduces the risk of over-fitting, provides flexibility, easy to determine feature importance.

The Gradient Boost is a machine-learning algorithm that works on the ensemble technique called boosting. Gradient boost sequentially combines many weak learners to form a stronger learner. It is one of the most powerful techniques for building predictive models for both classification and regression problems.

The KNN stands for the K-nearest neighbor algorithm. It is a non-parametric as well as supervised learning classifier. For classification or prediction of the grouping of an individual data point it uses proximity. In the year 1951, the KNN was developed by Evelyn and Joseph and then it was expanded by Thomas Cover. Since it is non-parametric it is widely used in real-life scenarios.

III. Our Contribution

A. Gap Analysis

As a car company or manufacturer, it is important to maintain their car quality, features, and specifications. As the increasing demand from the customers, the car companies are trying to produce cars with good specifications. There is also another main thing which is emission control. The CO2 emission control is also one of the important things to look at during the production of a car. With the help of this predictive model, it is also possible to predict the emission values produced by the Porsche car in all generations.

B. Research Questions

1) Which car produces the least CO2 emissions and which produces the most? Among all the Porsche generations we will find the model that produces less CO2 emission as well as the model that produces the most emissions.
2) The car that goes fast, does it produce more emissions? Among all the Porsche generations which model will go faster and also checks if it goes that fast does it emit more CO2 emissions as compared to other models.
3) If a car have better specification than the other cars does it goes more fast? Here we checks the specifications of each Porsche car and determine that the car with better specification will also go more fast.
4) Is there any change in fuel consumption over generations? By analyzing the history of Porsche car's generation, we will find out is there any increase in fuel consumption in newer generation models as compared to older generation models.
5) The car that consumes the most fuel also produces the most CO2 emissions? By analyzing the data we will also predict and check the Porsche car that consumes the most fuel will also produce the most CO2 emission.

C. Problem Statement

Making the most efficient and reliable car in today's world is more hard than we can imagine. Nowadays there exist many barriers and issues that need to be overcome by car manufacturers. The speed predictive model will help to sort the problems more efficiently and accurately. By predicting the speed of the vehicle, it is easy to understand many things that affects it directly such as body-to-weight ratio, horsepower, aerodynamics, emissions, and much more. The project aims to enhance the overall outcome of a car manufacturer and also to help many stakeholders for better predictions as well as decision-making by utilizing many machine learning techniques such as Linear Regression Model, KNN model, Decision Tree Regression, Random Forest, and Gradient Boosting. In order to lower the risk of emissions and increase the efficiency of cars, the ultimate goal is to give car companies and automotive designers a reliable, understandable, and useful instrument for evaluating and predicting the speed of cars.

D. Significance of Our Work

The significance of our work lies in its future of improving the automotive Industry by providing better results with a precise, efficient, and fair approach to speed prediction. Traditional ways for predicting the speed prediction applications mostly finds it difficult to manage the complexity and volume of data involved, leading to sub-optimal predictions and increased problems and issues. By leveraging advanced machine learning techniques, our study mentions these challenges and offers substantial improvements in predictive accuracy, operational efficiency, and decision-making transparency.

IV. Methodology

Importing the necessary libraries, such as Scikit learn, Pandas, and Numpy, to process data and create a prediction model. Upload a pandas Data Frame with the data and here we used Porsche 911 dataset. The predictive model will be trained using the training set, and its performance will be assessed using the testing set. Select a suitable machine learning algorithm, such Decision Tree regression, Random Forest, Decision trees Regression, to predict the speed of Porsche 911. Create an instance of the selected model and adjust any required hyper-parameters.

A. Dataset

We downloaded the data set from Kaggle. The Kaggle is known for its machine learning competitions. For this project, we have used the Porsche 911 dataset for predicting the values. The dataset contains all information regarding every generation of Porsche 911 model ever built. We imported the data and assigned data frame with the following code.

```
[ ] # Save current display options
    current_max_rows = pd.get_option('display.max_rows')
    # Set the display option to None to show all rows
    pd.set_option('display.max_rows', None)
    pd.set_option('display.max_columns', None)

[ ] df = pd.read_excel('/content/porsche_911_final1.xlsx')

[ ] df.drop_duplicates(inplace=True)

[ ] df.head()
```

	link	generation	engine	start_of_production	body_type	seats	acceleration_0-100km/h	acceleration_0-62mph	maximum_speed	weight-to-power_ratio_kg/hp	weight-to-power_ratio_hp/kg	we torque_rat
0	https://www.auto-data.net/en/porsche-911-dakar-...	911 Dakar	3	2022	Off-road vehicle and Coupe	2	3.4	3.4	240.0	3.3	299.1	
1	https://www.auto-data.net/en/porsche-911-targa-...	911 Targa 4S 3.0		2020	Targa	4	3.8	3.8	304.0	3.7	268.7	
2	https://www.auto-data.net/en/porsche-911-targa-...	911 Targa 4S 3.0		2020	Targa	4	4.4	4.4	304.0	3.6	274.4	

Fig. 1. The Dataset

B. Detailed Methodology

Linear Regression Main method in predicting modeling is linear regression and helps determine the coloration between CO2 emissions and other vehicles such as engine size, weight, fuel type. Linear regression algorithms provide an effortless way of predicting emissions through input by fitting a linear

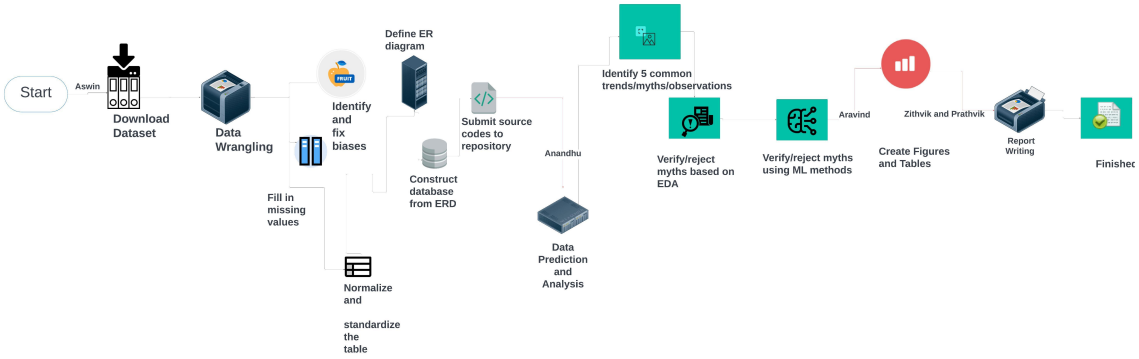

Fig. 2. Flowchart

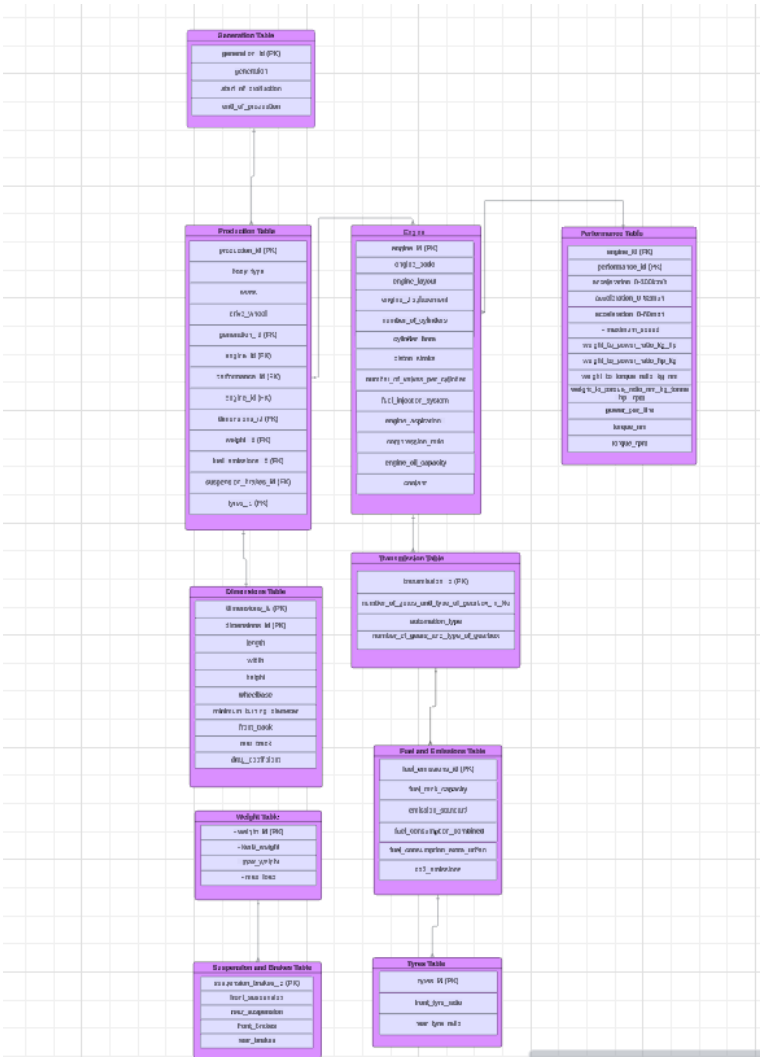

Fig. 3. Entity Relationship

equation to real-world data. For instance, researchers have shown that when there is an inherent linear relationship between variables, this may be using linear regressions as good predictors of motor vehicle emissions. The fact remains that even through the Porsche 911 used for this analysis is a complex model, there are several steps involved in developing a valid prediction using linear regression. These include data collection and preparation, model building and result assessment. Data preparation involves getting the clean data from the data set, which involves managing missing values and outliers to ensure all variables are numerical. In addition to that normalization of data may be done to maintain each attribute related to the status of residual plots, mean squared error, and coefficient of determination can also be applied after constructing models for evaluating their accuracy. On examining the determination of this model which gave the activity done researching effective model based on the complexity of data set or linearity between variables can be easily understood.

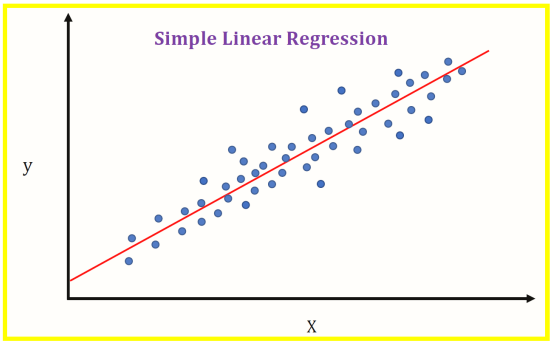

Fig. 4. Linear Rgression

Random forest Flavored algorithm for machine learning. A component of the supervised learning technique is random forest. It will be used for ML (Machine Learning) problems involving both classification and regression. It is based on concept of an ambulance learning which is technique for integrating many classifiers to handle tough jobs and problems to develop the performance of the model. Its name suggests that "Random Forest is a classifier that contains a number of decision trees on various subsets of the given data set and takes the average to improve the productivity accuracy of that data set." This uses predictions, from each decision tree and predicts, that the outcome depends on votes of a majority of projections rather than relying on one decision tree. The algorithm's ability to handle large data sets and numerous input features makes it particularly suitable for predicting vehicle emissions. It constructs multiple decision trees during training and merges their outputs to improve the accuracy and robustness of the production. Each tree in the forest is trained on a random subset of the data set and the final production is obtained by averaging the outputs of all individual trees.

Decision Tree This prediction model is known as a decision tree this flow chart, is a structure for the basis decisions on incoming data. Decision trees were used to provide models

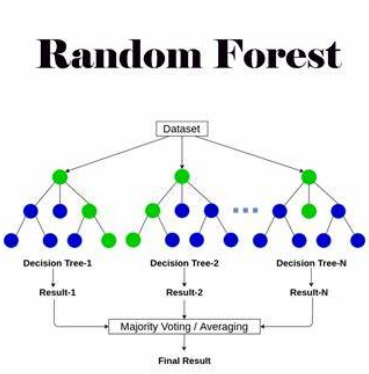

Fig. 5. Random Forest

that are simple to comprehend regression and classification problems. Data branches are built, and the results are placed at nodes of leaves. In decision support, decisions and their potential outcomes including chance occurrences, resource cost, utility are represented here by hierarchical models known as decision trees. The tree structure is made of root node, branches, internal nodes, and leaf nodes and has the appearance of a hierarchical tree. A prediction model known as decision tree uses flow charts like structure for base decision on incoming data. Decision trees were used to provide models that are simple to for classification and regression problems as shown. It is a non-parametric supervised learning method used for predicting continuous variables in this case we can make them a valuable tool. A decision tree can show how a combination of high horsepower, larger engine size, and vary fuel types may result in higher.

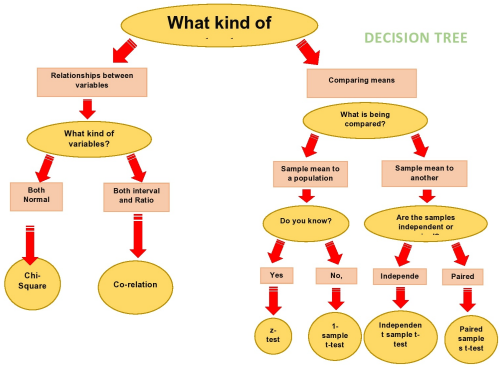

Fig. 6. Decision Tree

KNN algorithm K nearest neighbor is one of the basic supervised learning-based machine learning algorithms. These algorithms place good instances, in a category that resembles the current categories of the most pre-assuming the new cases and the previous cases are comparable. After sorting all the previous data, a new data point is categorized using the K-NN algorithm based on the similarity. Although this technique

is mostly repeatable work to solve classification problems, it can also be used for solving regressions and difficulties. K-NN Algorithms is a non-parametric method that makes no assumptions about the underlying data. Instead, it performs an action while classification of data by using the data set this approach is simple to store the data during phases of training and categorizing fresh data. It will be highly effective in capturing local patterns and situations where the relationship between the futures and emission is complex and nonlinear.

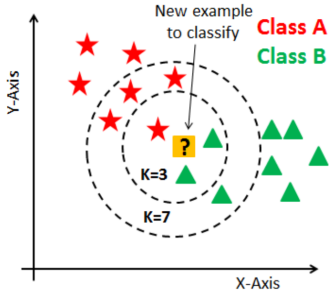

Fig. 7. KNN Algorithm

V. Results and Discussion

For the final result, we will start from the root. First and foremost we need to add the required libraries such as Pandas, Numpy, Seaborn, Matplotlib. And then we need to dataset into the software. The code used for loading the dataset will be shown below. Here we us the function head() for preview the data set and it will print the first 5 set of rows from the dataset. A short information of a data frame structure and column, including the type of data and the memory utilization by the df.info() method which is included in the Python Pandas package. After printing it comes up with column data-type, non-null count and also the usage of the memory. The df.isnull() code and the sum() function determines that how many columns are there in the data frame as null values or the missing values. It will give the complete list of the missing data. These line of codes are used for training the model for predicting the speed. The feature (X) is selected as the 'Regression Model' column, and the target variable (y) is chosen as the 'Evaluation Matrix's' column from train df.

After the accurate analysis using machine learning we have came with results for the questions that we discussed earlier. As programmers, we are satisfied with results. Here we will discuss the questions once again along with the results.

1. Which car produces the least CO2 emissions and which produces the most? Among all the Porsche generations we will find the model that produces less CO2 emission as well as the model that produces the most emissions. For this evaluation, we used the columns generation and CO2-emissions from the data set and plotted a (12,6) graph for predicting the value. From the result we can understand that the Porsche 911 speedster is producing the most carbon emission and the Porsche 911 Targa emits the least carbon emission. The graph is shown below.

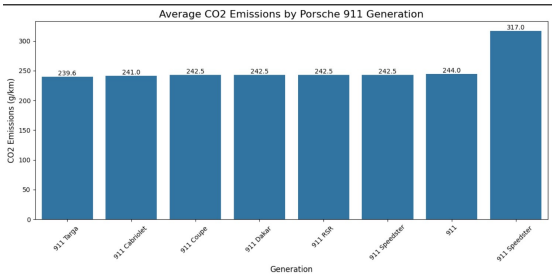

Fig. 8. Graphs of CO2 emission

2. The car that goes fast, does it produce more emissions? Among all the Porsche generations which model will go faster and also checks if it goes that fast does it emit more CO2 emissions as compared to other models. For this evaluation we compared the columns maximum-speed and co2-emissions and plotted a (12,6) graph for predicting the value. From the analysis it is clear that the car that goes faster does not produce the more emissions as compared to other models. In fact there exist some other models that produces more emission than the fastest car. The graph is shown below.

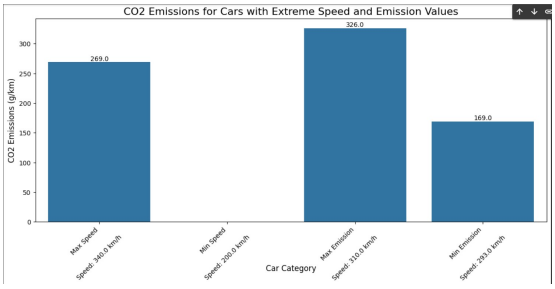

Fig. 9. Graph of speed and emission ratio

3. If a car have better specification than the other cars does it goes more fast? Here we checks the specifications of each Porsche car and determine that the car with better specification will also go more fast. For this evaluation we have used the columns Horsepower and Maximum-speed for predicting the values as well as to mark the name of the car we have used generation column also. From the analysis it is found that the car with highest horsepower is also have the highest top-speed. The name model is 911. The details can be seen in the graph below.

4. Is there any change in fuel consumption over generations? By analyzing the history of Porsche car's generation, we will find out is there any increase in fuel-consumption in newer generation models as compared to older generation models. For this evaluation we have used the columns fuel consumption combined and CO2-emissions for predicting the values as well as to name the model we have used generation column also. From the analysis it is clear that from the coming the consumption of fuel have been increased but the production of emission have been decreased. The car with highest fuel consumption is Porsche 911 Coupe and the car with highest carbon emission is Porsche 911.

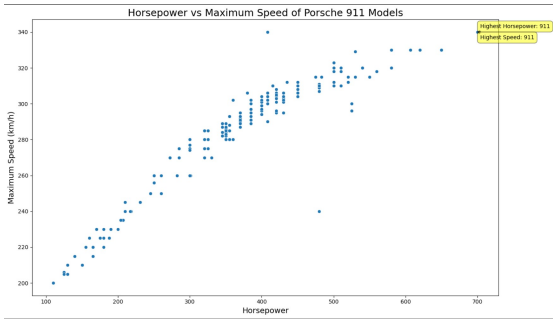

Fig. 10. Horse Power and Speed Ratio

5. The car that consumes the most fuel also produces the most CO2 emissions? By analyzing the data we will also predict and check the Porsche car that consumes the most fuel will also produces the most CO2 emission. For this analysis we have used the columns fuel-consumption and CO2-emissions and also used the generation column for mentioning the name of the models. It is clear from the graph that the model that consume the highest fuel is not producing the a lot of carbon emission as compared to other Porsche models. The detailed graph is shown below.

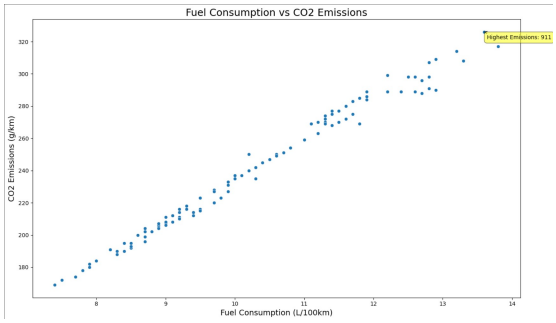

Fig. 11. Fuel Consumption and Carbon Emission ratio

We have also plotted a graph to represent the normalized form of all Porsche models and to compare all the models in all generations. The is shown below.

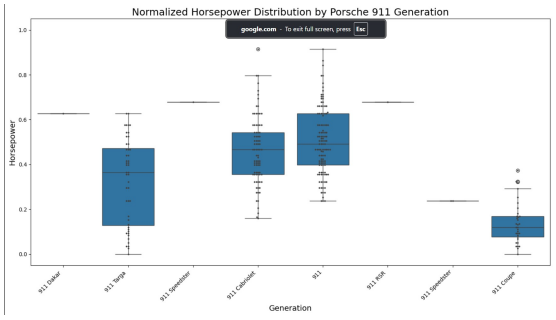

Fig. 12. Normalized Graph

A. Limitations

In this project, the speed of Porsche cars is predicted using emission standards and this carries some limitations like, The emission can also vary due to external factors like temperature, humidity as well as road conditions so this may sometime results in faulty results. The cars that uses modern technologies like catalytic converter can alter emission charteristics and thus the speed predicting model will shows different results. it is not applicable with the latest fully electric or hybrid vehicles models as it does not emits carbon dioxide or other substances.

B. Future Directions

Calibrate this model to individual cars, by considering the vehicle age, and engine type, then the accuracy of this model can be further improved. Also, create standardized protocols for emission data in order to ensure consistency and comparability across different types of vehicles. Develop methods to identify anomalies or outliers in emissions data such as those caused by road grade, non-uniform acceleration, or equipment malfunction. Also, build one model that analysis the driver behaviors and identify patterns that significantly influence emissions and speed relationship

VI. Conclusion

To conclude this, the research meets a crucial need in the Automotive specter by delivering the application of machine learning algorithms in knowing the status of a car's maximum speed, carbon emission, aerodynamics, and more. With the help of models such as Random Forest, Gradient Boosting, KNN, data cleaning, and feature engineering we are able to predict the accuracy. The final product is a predictive model where it helps the automotive designers car manufacturers and other stakeholders related to the automotive industries as a helpful tool to produce an efficient car for the future. This machine learning model helps not only the industries to make better profits by creating or discovering new technologies that would help to produce good cars with low carbon emission but also helps to reduce the consumption of raw materials and this would indirectly make a better for the environment with less pollution and clean air.

References

[1] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[2] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[3] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[4] Y. Song, J. Huang, D. Zhou, H. Zha, and C. L. Giles, "Iknn: Informative k-nearest neighbor pattern classification," in *European conference on principles of data mining and knowledge discovery*. Springer, 2007, pp. 248–264.

[5] M. A. Kabir, "Vehicle speed prediction based on road status using machine learning," *Advanced Research in Energy and Engineering*, vol. 2, no. 1, 2020.

[6] B. Cantor, P. Grant, and C. Johnston, *Automotive engineering: lightweight, functional, and novel materials*. CRC press, 2008.

[7] A. S. Mohammed, A. S. Mohammed, and S. W. Kareem, "Deep learning and neural network-based wind speed prediction model," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 30, no. 03, pp. 403–425, 2022.

[8] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *Ieee transactions on intelligent transportation systems*, vol. 16, no. 2, pp. 865–873, 2014.

[9] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.
[10] G. Fennessy, "Autonomous vehicle end-to-end reinforcement learning model and the effects of image segmentation on model quality," Ph.D. dissertation, 2019.