**LINEAR REGRESSION AND RANDOM FOREST CLASSIFICATION USING PYTHON**

**Table of Contents**

[Introduction 3](#_Toc101956993)

[Data 3](#_Toc101956994)

[Method 3](#_Toc101956995)

[Results 5](#_Toc101956996)

[Discussion 6](#_Toc101956997)

[Conclusion 7](#_Toc101956998)

[References 8](#_Toc101956999)

# Introduction

Linear regression is the machine learning algorithm, which is a part of each and every data scientist. This model is a simple model but for this model, everyone needs to master this model because this model lays the foundation for another type of machine learning algorithm. This article is the basis of the linear regression model and the implementation of the python programming language. Linear regression is the statistical method that is based on the dependent variable and the independent variable. The variable linear regression studies the linear relationship between the dependent variable and also the independent variable. The Y is the dependent variable and X is the independent variable. The model linear regression describes the dependent variable within a single straight line which is based on the equation y= a+b \* x. Random forest classification is also used in the implementation part using the python language. The accuracy score and prediction value is the part of random forest classification.

# Data

The dataset that has been used here is named "car price assignment", where the dependent variable and the independent variable are used in a proper way. The car ID is used here as the dependent variable where the linear regression process has taken the part. The data set that is used here is named the linear regression process where the information of the dataset is shown here with different types of generated parts. As per the view of Zha *et al.*(2020), there are rows and columns in the dataset that have been shown specifically step by step. The car height, car length, and fuel type are also generated in the dataset in which the linear regression model with a proper machine learning algorithm has taken the part. The data set is included categorical values and also the integer value. The categorical value has been dropped from the dataset to generate the regression procedure and the integer value is the process where the value has been part of the machine learning algorithm.

The next dataset is named glass where the random forest classification has been used. The dataset consists different columns of glass types called Al. Si, Mg etc where the integer variables is the most important number with different classification.

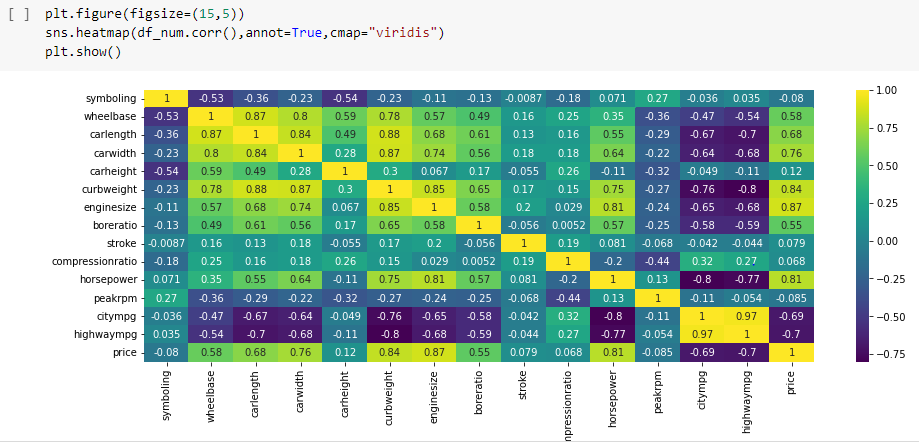
The latest dataset is named social networking ad where the columns are age, estimated and purchased. The 3 columns are consisting numerical variable. On this numerical values, the random forst classification has been used in the implementation part.

# Method

The particular column price is the main target variable here and the rest of the columns are the independent variable. After plotting the heatmap the important part is to visualize the multiple correlation method of the numerical variable. The target variable has been used here as price and the correlation of the numerical variable is also used here with the different portions. The price portion is related to the wheelbase, carbs, car width, horsepower, and the engine size method step by step. The particular price is negatively correlated with the city mpg and highway mpg. The method and the data suggest the cars that have the higher mileage maybe fall into the economic cars category portion and the particular price is lower than the lower part. There are many independent variables that are highly correlated with the car which are based on the car length, and engine size which all measure that the cars are positively correlated with each other. There are no missing values in the dataset s there data-cleaning part is not needed in the dataset for the similar portion and for the similar category. There are two types of linear regression that have been used in the dataset and which are simple linear regression and multiple linear regression. In any case, the single independent variable is mainly used for the prediction of a numerical variable, then the linear algorithm is mainly called the simple linear regression model. Multiple linear regression is used because of the linear regression algorithm which is based on the prediction of the different independent and dependent variables. In any case, a dependent variable increase the Y-axis and the indepndentt variable always increases the x-axis, then it is called the positive linear relationship of the variable. As per the view of Zin *et al.*(2019), On the other hand, the dependent variable decreases the Y-axis and the independent variable increases on the x axis, then, the relationship is called the negative linear relationship method.

The r square method is used in the linear regression which determines the goodness of fit. The method always measures the length of the relationship between the dependent and also the independent variable with a scale of 0-100%. The higher value of R square always determines the less difference part of the predicted values and the predicted values then it always represents the good model and good linear relationship. As per the view of Samal *et al.*(2019), the coefficient determination, the coefficient multiple determination has been used herein the linear relationship and linear value for the multiple processes. The r squared linear relationship can be used here with the explained variation and the total variation. The linear relationship assumes the regression between the dependent and also independent variable. There is a high correlation between dependent and independent variables hs been used in the linear relationship model. There are multiple ways to true relationship the predictions and the target variables has been used with the different variables. The predictor part is used as the target variable, which affects the linear regression and linear relationship. It has been assumed that multiple collinearity features are used in the independent and dependent variable models. The linear regression model assumes that the error term should be followed by the normalization of the term that can be distributed. The intervals can become easy to predict therefore the linear relationship. The q-q plot has been described in the linear relationship. The q-q plot shows the straight line for the linear relationship that are normally distributed part-wise.

There is no autocorrelation model that has been used in the linear relationship because of the prediction of the accuracy model and the autocorrelation terms. The autocorrelation usually occurs if there is a dependency between the residual errors.

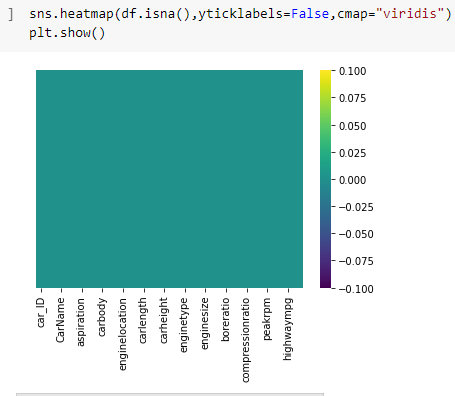


**Figure 1: sns heatmap and prediction**

(Source: Acquired from Jupyter notebook)

# Results

The technical part is completed with the help of a jupyter notebook regarding python language. With the help of the jupyter notebook, some heatmap has been shown in the implementation portion. The prediction part has been done with the help of accuracy scope and linear regression method. The description part has been shown in the dataset with the help of the "object" variable.



**Figure 2: sns heatmap and prediction 2**

(Source: Acquired from Jupyter notebook)

After that the car id has been dropped for the prediction and the sum has been dependent on the different variables. The box plot part is shown with the help of sns in the different portions and the boxplot is different on each other variable. As per the view of Zu *et al.*(2018), The x train test and y train test are shown in the model fit portion. The train test split is used for the dependency of the linear regression and linear relationship method partwise.



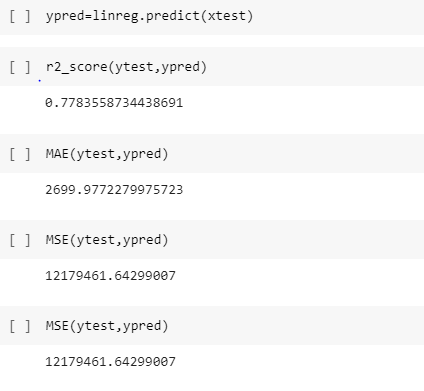
**Figure : plot with numerical variables**

(Source: Acquired from Jupyter notebook)

The sklearn model is used for the prediction of the x train test and the y train test model issued. There are different solutions for the model fit portion and the accuracy score that is used for the jupyter notebook portion.

# Discussion

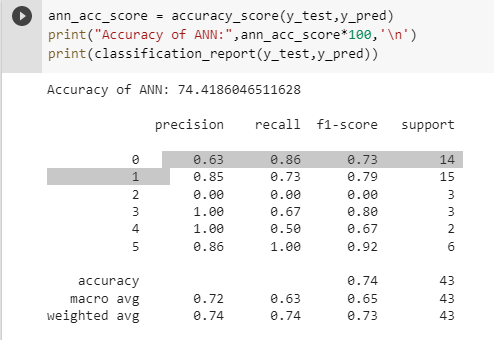
The study is based on the car price prediction with the help of linear regression of a machine learning algorithm. There are different types of variables are used in the car precise prediction. In the software part, the important library are the initial part of the discussion of the whole method.



**Figure 3: Accuracy scores after prediction**

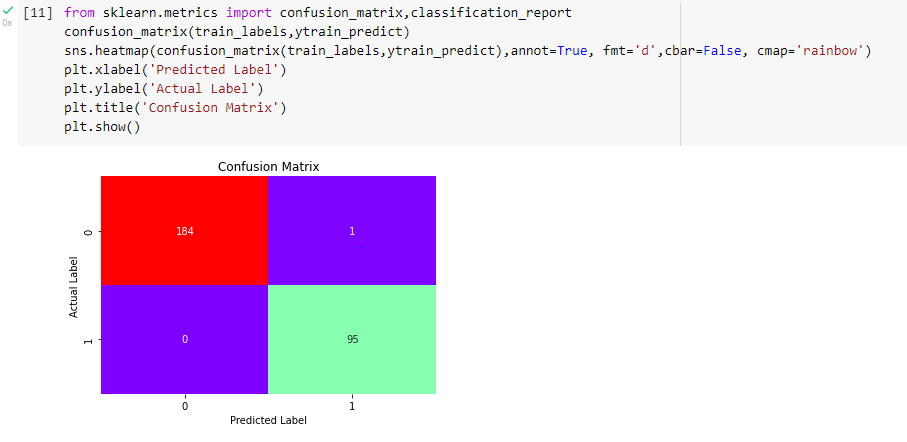
(Source: Acquired from Jupyter notebook)

After that, all the method is discussed as the train test split and the model fit with the accuracy score. There are different box plots and the heatmaps are shown in the figure where the different linear regression is used here with the algorithm portion. As per the view of Ghosal *et al.*(2019), In the box plots, there are different kind of algorithm with different solution has been used where all the dependent and independent variable has been used in the data model portion. In the data model portion, all the single and multiple linear regression is used partwise.



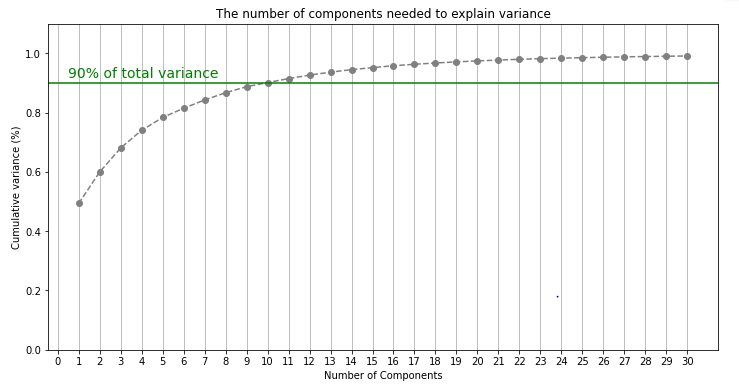
**Figure 4: Accuracy score using random forest classification**

(Source: Acquired from Jupyter notebook)



**Figure 5 : Confusion matrix using random forest**

(Source: Acquired from Jupyter notebook)



**Figure 6: graph regarding components**

(Source: Acquired from Jupyter notebook)

# Conclusion

This study ends with a brief description of a linear regression model where the car price prediction is used step by step. In this topic, all the regression model is mainly used here as the model fit portion and the graphs and box plots are shown on the jupyter notebook model. The jupyter notebook is used here with different solutions of the algorithm of x train test and y train test. All the values that have been predicted in the algorithm have been shown part-wise in the software work. The random forest classification is used here for the implementation part.

# 

# References

Fu, Z., Jiang, J., Gao, Y., Krienke, B., Wang, M., Zhong, K., Cao, Q., Tian, Y., Zhu, Y., Cao, W. and Liu, X., 2020. Wheat growth monitoring and yield estimation based on multi-rotor unmanned aerial vehicles. *Remote Sensing*, *12*(3), p.508.

Ghosal, S., Sengupta, S., Majumder, M., and Sinha, B., 2020. Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases-March 14th 2020). *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(4), pp.311-315.

Hsieh, I.Y.L., Pan, M.S., Chiang, Y.M. and Green, W.H., 2019. Learning only buys you so much: Practical limits on battery price reduction. *Applied Energy*, *239*, pp.218-224.

Jin, X.B., Gong, W.T., Kong, J.L., Bai, Y.T. and Su, T.L., 2022. PFVAE: a planar flow-based variational auto-encoder prediction model for time series data. *Mathematics*, *10*(4), p.610.

Pai, P.F. and Liu, C.H., 2018. Predicting vehicle sales by sentiment analysis of Twitter data and stock market values. *IEEE Access*, *6*, pp.57655-57662.

Samal, K.K.R., Babu, K.S., Das, S.K. and Acharaya, A., 2019, August. Time series-based air pollution forecasting using SARIMA and prophet model. In *Proceedings of the 2019 international conference on information technology and computer communications* (pp. 80-85).

Xiao, Z., Li, P., Havyarimana, V., Hassana, G.M., Wang, D. and Li, K., 2018. GOI: A novel design for vehicle positioning and trajectory prediction under urban environments. *IEEE Sensors Journal*, *18*(13), pp.5586-5594.

Zha, H., Miao, Y., Wang, T., Li, Y., Zhang, J., Sun, W., Feng, Z. and Kusnierek, K., 2020. Improving unmanned aerial vehicle remote sensing-based rice nitrogen nutrition index prediction with machine learning. *Remote Sensing*, *12*(2), p.215.