**Carbon Nanotubes Data Set**

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of dataset** | **Dataset Characteristics** | **Number of attributes** | **Number of instances** |
| Regression | Univariate | 8 | 10721 |
|  |  |  |  |

**Introduction:**

The dataset used here is Air Foil Self Noise dataset which is been taken from UCI Repository https://archive.ics.uci.edu/ml/datasets/Carbon+Nanotubes

**Data Set Description:**

CASTEP can simulate a wide range of properties of materials proprieties using density functional theory (DFT). DFT is the most successful method calculates atomic coordinates faster than other mathematical approaches, and it also reaches more accurate results.

The dataset is generated with CASTEP using CNT geometry optimization. Many CNTs are simulated in CASTEP, then geometry optimizations are calculated. Initial coordinates of all carbon atoms are generated randomly. Different chiral vectors are used for each CNT simulation. The atom type is selected as carbon, bond length is used as 1.42 AÂ° (default value). CNT calculation parameters are used as default parameters.

To finalize the computation, CASTEP uses a parameter named as elec\_energy\_tol (electrical energy tolerance) (default 1x10-5 eV) which represents that the change in the total energy from one iteration to the next remains below some tolerance value per atom for a few self-consistent field steps. Initial atomic coordinates (u, v, w), chiral vector (n, m) and calculated atomic coordinates (uâ€™, vâ€™, wâ€™) are obtained from the output files.

**Attribute Information:**

This problem has the following inputs:(X – independent variable)

1. Chiral indice n: n parameter of the selected chiral vector.  
 2. Chiral indice m: n parameter of the selected chiral vector.  
 3. Initial atomic coordinate u: Randomly generated u parameter of the initial atomic coordinates of all carbon atoms.  
 4. Initial atomic coordinate v: Randomly generated v parameter of the initial atomic coordinates of all carbon atoms.  
 5. Initial atomic coordinate w: Randomly generated w parameter of the initial atomic coordinates of all carbon atoms.

The output is: (y – Target variable)

1. Calculated atomic coordinate uâ€™: Calculated uâ€™ parameter of the atomic coordinates of all carbon atoms.  
 2. Calculated atomic coordinate vâ€™: Calculated vâ€™ parameter of the atomic coordinates of all carbon atoms.  
 3. Calculated atomic coordinate wâ€™: Calculated wâ€™ parameter of the atomic coordinates of all carbon atoms.

**Machine Learning Procedure:**

**Supervised Machine Learning:**

Supervised Learning, also known as Supervised Machine Learning is defined by its use of labelled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately.

This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting.

Supervised Learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in Supervised Learning include Neural Networks, Naive-Bayes, Linear Regression, Logistics Regression, Random Forest, Support Vector Machine(SVM) and more.

**Regression Algorithm:**

In Machine Learning, we use various kinds of algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns or rules identified from the dataset.

Regression is a supervised machine learning technique which is used to predict continuous values. Predicting prices of a house given the features of house like size, price etc is one of the common examples of Regression.

The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data. The three main metrics that are used for evaluating the trained regression model are variance, bias and error .

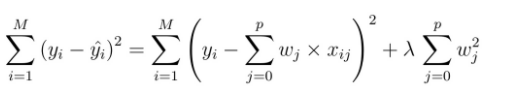
Types of algorithms used for this dataset are Linear, Ridge, Lasso, ElasticNet, Random Forest Regression Algorithm

**Linear Regression Algorithm:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

**Ridge Algorithm:**

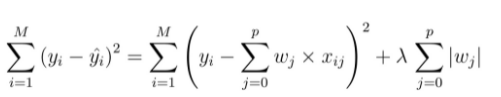
In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.



**Lasso Algorithm:**

The only difference in Lasso over Ridge Algorithm is instead of taking the square of the coefficients, magnitudes are taken into account.

In [statistics](https://en.wikipedia.org/wiki/Statistics) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning), lasso (**least absolute shrinkage and selection operator**; also Lasso or LASSO) is a [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) method that performs both [variable selection](https://en.wikipedia.org/wiki/Variable_selection) and [regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)) in order to enhance the prediction accuracy and interpretability of the resulting [statistical model](https://en.wikipedia.org/wiki/Statistical_model).Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).



**Elastic Net Algorithm:**

Elastic net is a popular type of regularized linear regressionthat combines two popular penalties, specifically the L1 and L2 penalty functions. Elastic Net is an extension of linear regression that adds regularization penalties to the loss function during training.

In [statistics](https://en.wikipedia.org/wiki/Statistics) and, in particular, in the fitting of [linear](https://en.wikipedia.org/wiki/Linear_regression) or [logistic regression](https://en.wikipedia.org/wiki/Logistic_regression) models, the elastic net is a [regularized](https://en.wikipedia.org/wiki/Regularization_(mathematics)) regression method that [linearly combines](https://en.wikipedia.org/wiki/Linear_combination) the [L1](https://en.wikipedia.org/wiki/Taxicab_geometry) and [L2](https://en.wikipedia.org/wiki/Norm_(mathematics)" \l "Euclidean_norm) penalties of the lasso and [ridge](https://en.wikipedia.org/wiki/Tikhonov_regularization) methods. The elastic net method overcomes the limitations of the [LASSO](https://en.wikipedia.org/wiki/Lasso_(statistics)) (least absolute shrinkage and selection operator) method.

**Random Forest Regression:**

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

1. Pick at random k data points from the training set.

2. Build a decision tree associated to these k data points.

3. Choose the number N of trees you want to build and repeat steps 1 and 2.

4. For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

**Data Preprocessing:**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project and while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

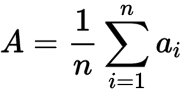
There are many ways to handle missing data, the one we used in Carbon Nanotubes dataset is **Mean**:

We used Scikit-learn library in our code, which contains various libraries for building machine learning models. SimpleImputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. It replaces the NaN values with a specified place holder. It is implemented by the use of the SimpleImputer() method.

The goal is to find out which is a better measure of the central tendency of data and use that value for replacing missing values appropriately.

We will calculate the **Mean** of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data and helps in increasing the efficiency.

To calculate mean, add together all of the numbers in a set and then divide the sum by the total count of numbers. The formula for mean is:



**A = Arithmetic Mean**

**n = Number of values**

**ai = data sets values**

**Outliers and Missing Treatment:**

One of the biggest downfalls for any model performance is the outliers present in the data. Outliers ideally are the extreme values for the specific column which affects the generalization of the data and model. Outliers mostly affect the regression models as it changes the equation drastically. It is an observation that is numerically distant from the rest of the data.

When reviewing a box plot, an outlier is defined as the data points that is located outside the whiskers of the box plot. IQR is used to measure variability by dividing a data set into quartiles. The data is sorted in ascending order and split into 4 equal parts. Q1, Q2, Q3 called first, second and third quartiles are the values which separate the 4 equal parts.

IQR is the range between the first and the third quartiles namely Q1 and Q3:**IQR = Q3 – Q1.** The data points which fall **below Q1 – 1.5 IQR** or **above Q3 + 1.5 IQR** are outliers.

The intention of Outlier detection and treatment is to ensure us to get the best model out of the data considering the fact that your data is qualified to work with the algorithm.

**Algorithm for for IQR:**

1. Sort the data in increasing order.

2. Calculate the first quartile(q1) and third quantile(q3).

3. Find Inter Quantile Range(q3-q1).

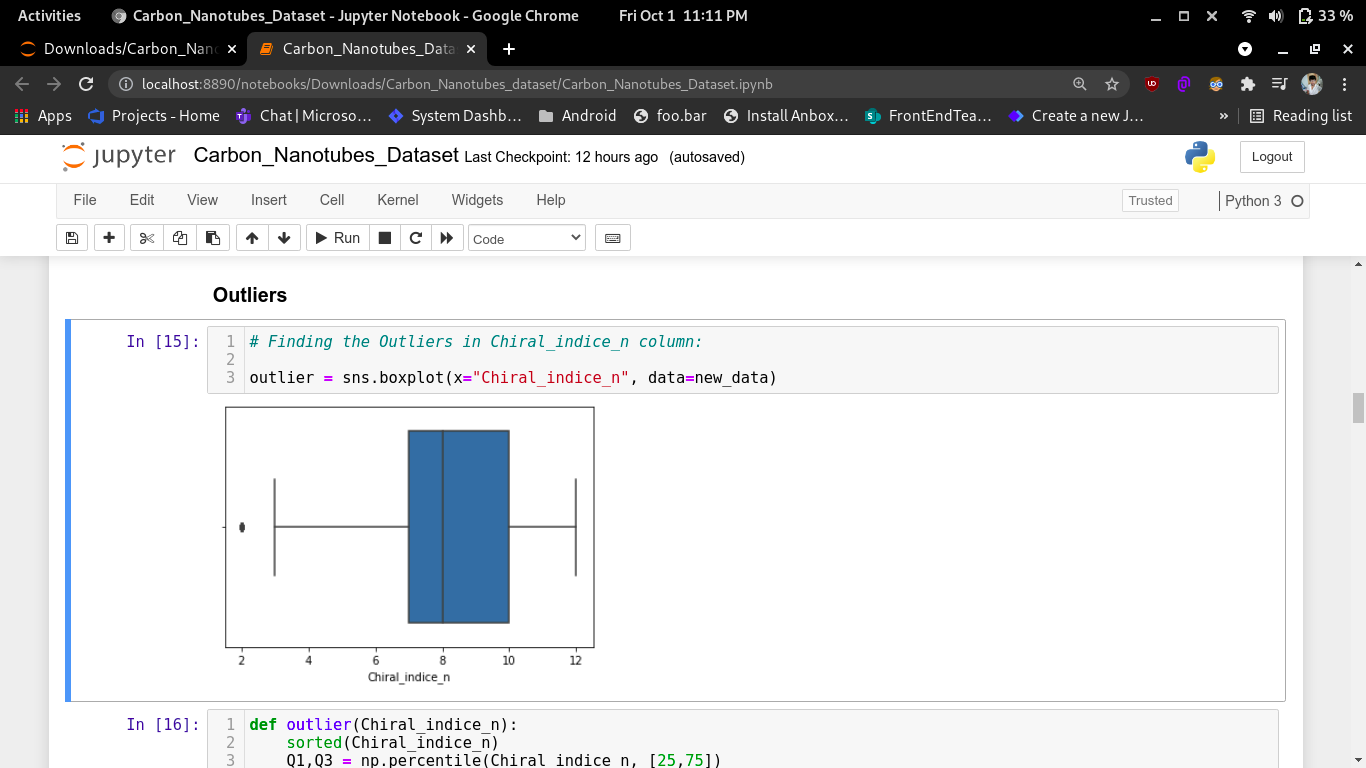
4. Find the lower\_bound = q1 – 1.5\*IQR

5. Find the upper\_bound = q3 + 1.5\*IQR

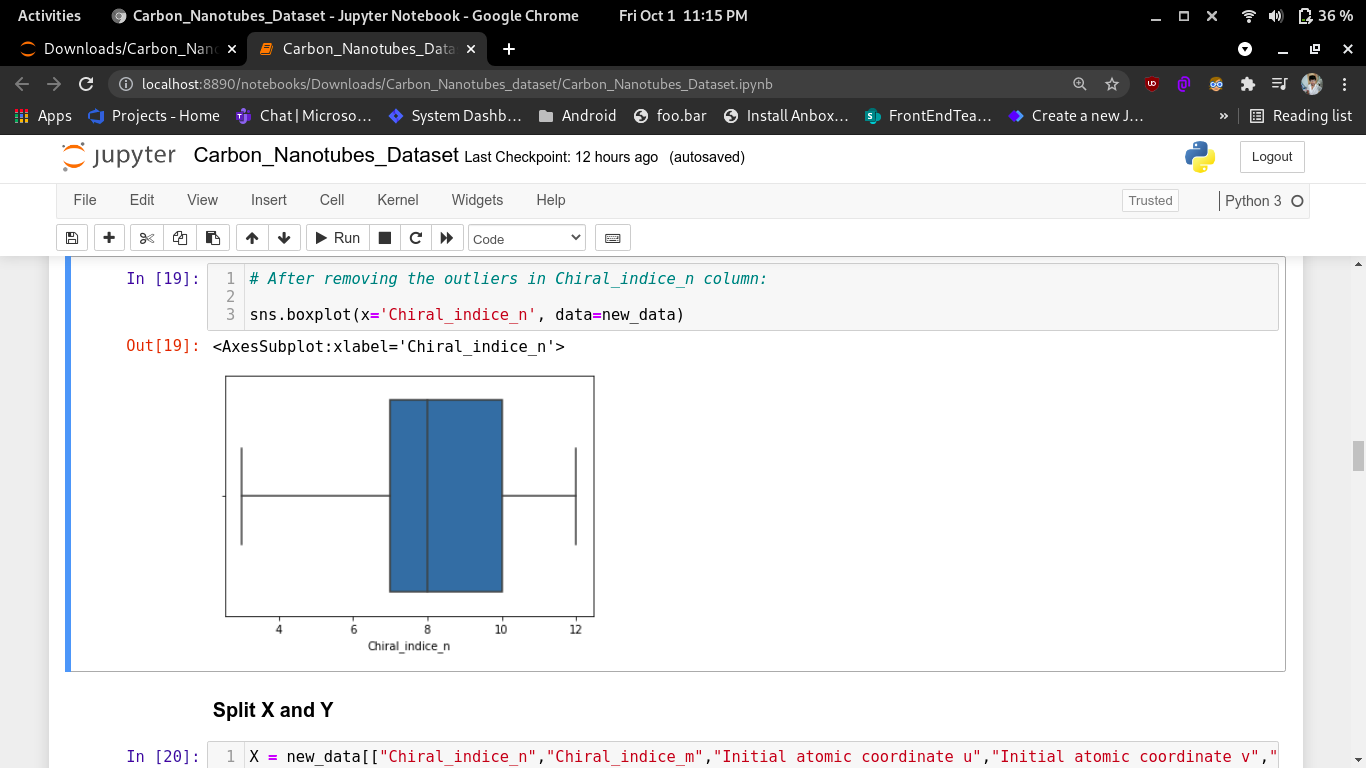
6. Points that lies above or below the IQR is an Outlier.

Box plots are useful as they provide a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set and signs of skewness.

Here , is the screenshot of outliers that are present in Carbon\_Nano tubes Dataset:



Here, is the screenshot of after removing the outliers in Carbon\_Nano tubes Dataset:



**Significance of Outliers:**

• Outliers badly affect mean and standard deviation of the dataset. These may statistically give erroneous results.

• Most machine learning algorithms do not work well in the presence of outlier. So it is desirable to detect and remove outliers.

• Outliers are highly useful in anomaly detection like fraud detection where the fraud transactions are very different from normal transactions.

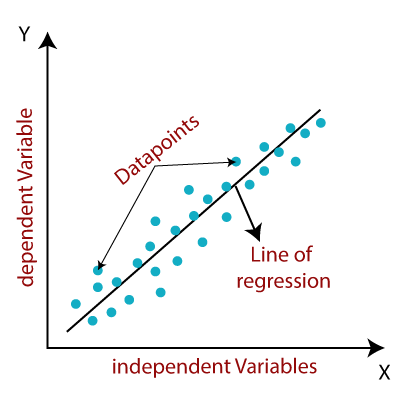
**Machine Learning Model:**

**Linear Regression Algorithm:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. If there is only one input variable (x), then such linear regression is called simple linear regression. And if there is more than one input variable, then such linear regression is called multiple linear regression. The relationship between variables in the linear regression model can be explained using the below image.



Formula: **y = a 0 +a 1 x+ ε**

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error. The different values for weights or the coefficient of lines (a 0 , a 1 ) gives a different line of regression, so we need to calculate the best values for a 0 and a 1 to find the best fit line, so to calculate this we use cost function.

**Cost function:**

o The different values for weights or coefficient of lines (a 0 , a 1 ) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

o Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

o We can use the cost function to find the accuracy of the mapping function, which maps the input variable to the output variable. This mapping function is also known as Hypothesis function.

**Best Fit Algorithm:**

The algorithm that gives maximum accuracy is Random Forest Regression and hence it is chosen as best fit algorithm for Carbon Nanotubes Dataset.

**Analysis of the dataset:**

By using the various Regression Algorithms we have analysed the accuracy as R2 Score for each algorithm and they are tabulated below:

|  |  |
| --- | --- |
| **Algorithm** | **R2 Score** |
| Linear Regression | 0.5381938826894261 |
| Ridge Algorithm | 0.5381938941968674 |
| Lasso Algorithm | 0.538193921571357 |
| Elastic Net Regression | 0.5382320488797316 |
| Random Forest Regression | 0.5919286233054354 |

This table shows the accuracy for each algorithm for Carbon Nano-tubes Dataset. Here, Random Forest Regression Algorithm gives the maximum accuracy of 59%.

**Conclusion:**

From all the above discussions, we have concluded that the performance will be more accurate only after all the process have done. In this Carbon Nano-tubes dataset the Calculated atomic coordinate uâ€™, Calculated atomic coordinate vâ€™ and Calculated atomic coordinate wâ€™ accuracies has been predicted. Hence, for this Carbon Nano-tubes dataset **Random Forest Regression Algorithm** gives the maximum accuracy of **59%.**