

Machine Intelligence

**Project Report**

Energy prediction Machine Learning algorithms

**Team Members**

Maram Mostafa 19105606

Muhannad Yasser

Rana tolba 202000763

Gehadgamal 18200050

Adham waleed 202000857

**Submitted To:**

Dr. ghada korba

Contents

[Introduction 4](#_Toc14542)

[Dataset Overview 4](#_Toc14543)

[Literature Review 5](#_Toc14544)

[Methodology 8](#_Toc14545)

[Optimizing Performance 8](#_Toc14546)

[Data Description 9](#_Toc14547)

[Code: 9](#_Toc14548)

[Output: 10](#_Toc14549)

[Data Merging 11](#_Toc14550)

[Code and Output 11](#_Toc14551)

[Histogram Plots 12](#_Toc14552)

[Year Built 12](#_Toc14553)

[Floor Count 12](#_Toc14554)

[Cloud Coverage 12](#_Toc14555)

[Wind Speed 13](#_Toc14556)

[Dew Temperature 13](#_Toc14557)

[Air Temperature 13](#_Toc14558)

[Additional Plots 14](#_Toc14559)

[Scatter Plot 14](#_Toc14560)

[Box Plot 14](#_Toc14561)

[Correlation Heatmap 15](#_Toc14562)

[Data Cleaning 15](#_Toc14563)

[Code 16](#_Toc14564)

[Results 16](#_Toc14565)

[Linear Regression Model 17](#_Toc14566)

[Data Manipulation 17](#_Toc14567)

[Linear Regression Object creation 18](#_Toc14568)

[Linear Regression fitting on test data 18](#_Toc14569)

[Ridge 18](#_Toc14570)

[Decision Tree 19](#_Toc14571)

[Decision tree regression 19](#_Toc14572)

[Generating the predicted values and creating data frame 20](#_Toc14573)

[Calculating errors & accuracy 20](#_Toc14574)

[At depth = 2 20](#_Toc14575)

[At depth =5 20](#_Toc14576)

[Random Forest 20](#_Toc14577)

[Random Forest Regression Object creation 21](#_Toc14578)

[Random Forest Regression data fitting 21](#_Toc14579)

[Random Forest Regression with 10 estimators 21](#_Toc14580)

[Results 22](#_Toc14581)

[Table of Comparison 22](#_Toc14582)

[Conclusion 23](#_Toc14583)

[References 24](#_Toc14584)

# Introduction

Machine learning has been very trending in the past few decades. Accordingly, there have been new advances lately that made life easier and more luxurious such as autonomous vehicles, which made driving easier and safer. However, everything has its own pros and cons. In order to better understand and acknowledge machine learning and how to develop it. This paper aims to apply three machine learning techniques for Reducing the overall energy consumption and associated greenhouse gas emissions in the construction sector which will sooner be more sustained due to the application of the machine learning techniques intended. One promising technology adoption is an energy metering infrastructure, which has been widely deployed around the world. This metrics include different sites, air temperature, air pressure and more data for the prediction to be made. In our paper we applied three machine learning algorithms Linear regression, Decision tree regression & Random Forest Regression.

# Dataset Overview

The data comes from over 1,000 buildings over a three-year timeframe. With better estimates of these energy-saving investments, large scale investors and financial institutions will be more inclined to invest in this area to enable progress in building efficiencies.

As per ASHRAE Assessing the value of energy efficiency improvements can be challenging as there's no way to truly know how much energy a building would have used without the improvements. The best we can do is to build counterfactual models. Once a building is overhauled the new (lower) energy consumption is compared against modeled values for the original building to calculate the savings from the retrofit. More accurate models could support better market incentives and enable lower cost financing.

This competition challenges you to build these models across four energy types, based on historic usage rates and observed weather. **The dataset includes three years of hourly meter readings from over one thousand buildings at several different sites around the world.**  **Meter reading**, is the target variable, AKA Energy consumption, in kWh (or equivalent)

# Literature Review

The competition was launched online on October 15, 2019, on the Kaggle website. During the competition, a total of 4,370 participants took part, comprising a total of 3,614 teams. There was a total of 80 countries represented among these participants.

Upon launch of the competition, the contestants were made aware of the financial rewards of winning the competition. The monetary prizes were supported by ASHRAE and included the following breakdown in US dollars:

\_ 1st place - $10,000

\_ 2nd place - $7,000

\_ 3rd place - $5,000

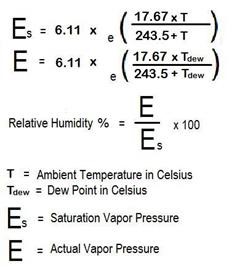
\_ 4th place - $2,000

\_ 5th place - $1,000

Vaishali Dhar worked on the dataset and published a paper on Jun 21, 2021

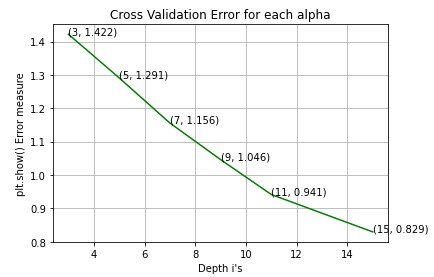
After reviewing the dataset, she dropped building 1099 site 13 because it was an outlier, she also dropped year\_built and floor\_count features because more than 50% of these values were missing. She filled in all the missing values from the weather data in the other columns. The weather changes significantly from one location to another, as well as from month to month and day to night. She imputed the missing values based on this data. She discovered the average site-wise temperatures for each day of the month and used this to fill in the missing values. She also imputed the rest of the values with the median in cloud\_coverage, precip\_depth\_1\_hr and sea\_level\_pressure because some of the slices were completely NAN. Then she added now features like holiday because holidays can affect the daily energy consumption, and other features like season feature isDaytime feature, and Relative Humidity feature. She used both the air temperature and the dew temperature features in the dataset to add the relative humidity information, which influences how a person feels at the time and, as a result, affects energy usage.

Using this formla.

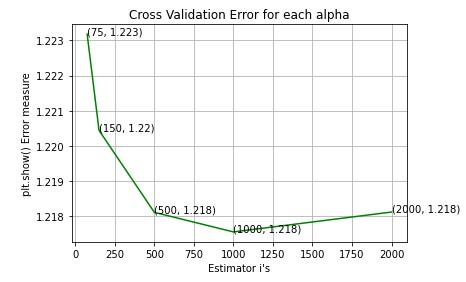


She used Baseline Model, Decision Tree Regressors, LGB Random Forest, LGB GBDT, Catboost GBDT, MLP Regressor models on the dataset and the results were as following: Baseline Model which Root Mean Square Error (RMSE) for train data is 1.500695067537 and for the cv data is 1.5341502380

Decision Tree Regressors was a pretty decent score

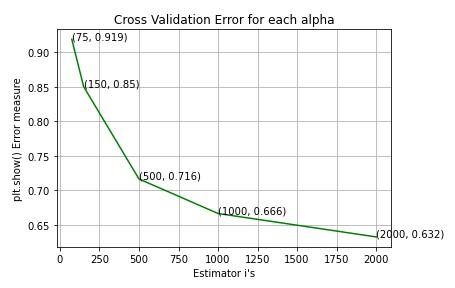


LGB Random Forest did not provide any improvisation over decision tree model

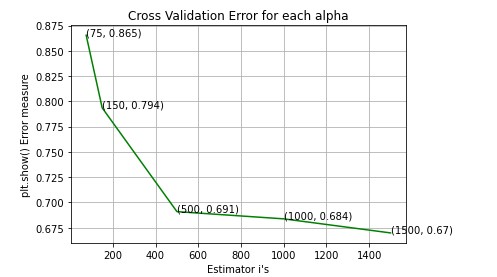


LGB GBDT

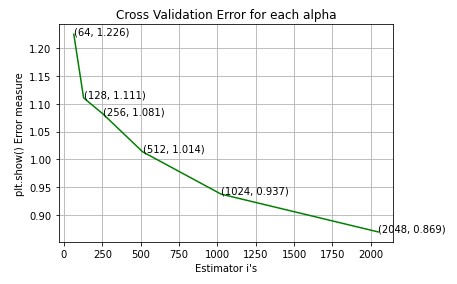
was the best score



Catboost GBDT



MLP Regressor



With her best model, she got a public score of 1.292, which is fairly good considering the obstacles this dataset presents, such as its large size, which makes cross validation and hyperparameter tweaking difficult.

# Methodology

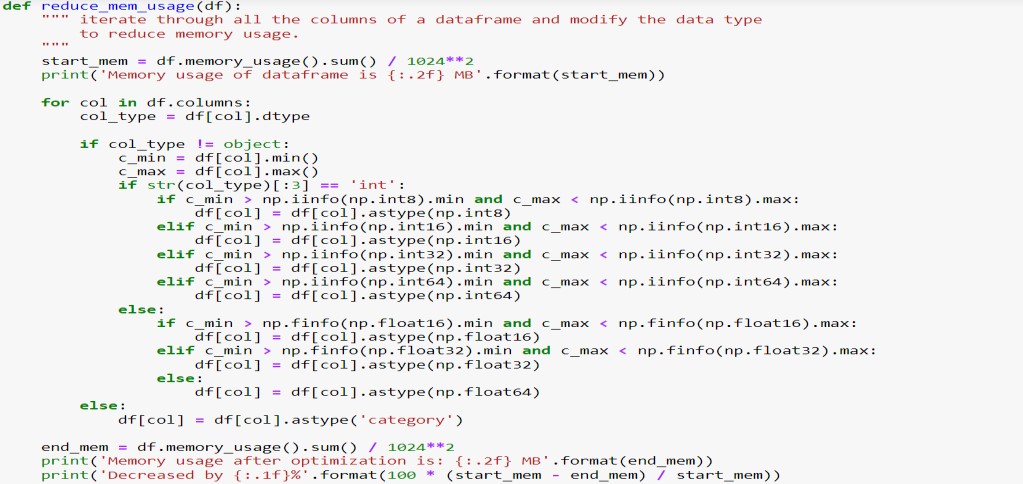
In this paper, we decided to try several machine learning algorithms on a certain dataset, to help solve a major problem proposed by ASHRAE, to help estimate and decrease energy consumption in high skyscrapers. There are several features used for prediction and they are: 1. Site id

1. Timestamp
2. Air temperature
3. Cloud coverage
4. Dew temperature
5. Precipitation depth
6. Sea level pressure
7. Wind direction
8. Wind speed
9. Building ID
10. Primary use
11. Square feet
12. Year built
13. Floor count

The data used were divided into training data & testing data to avoid cheating and be accurate as much as possible

# Optimizing Performance

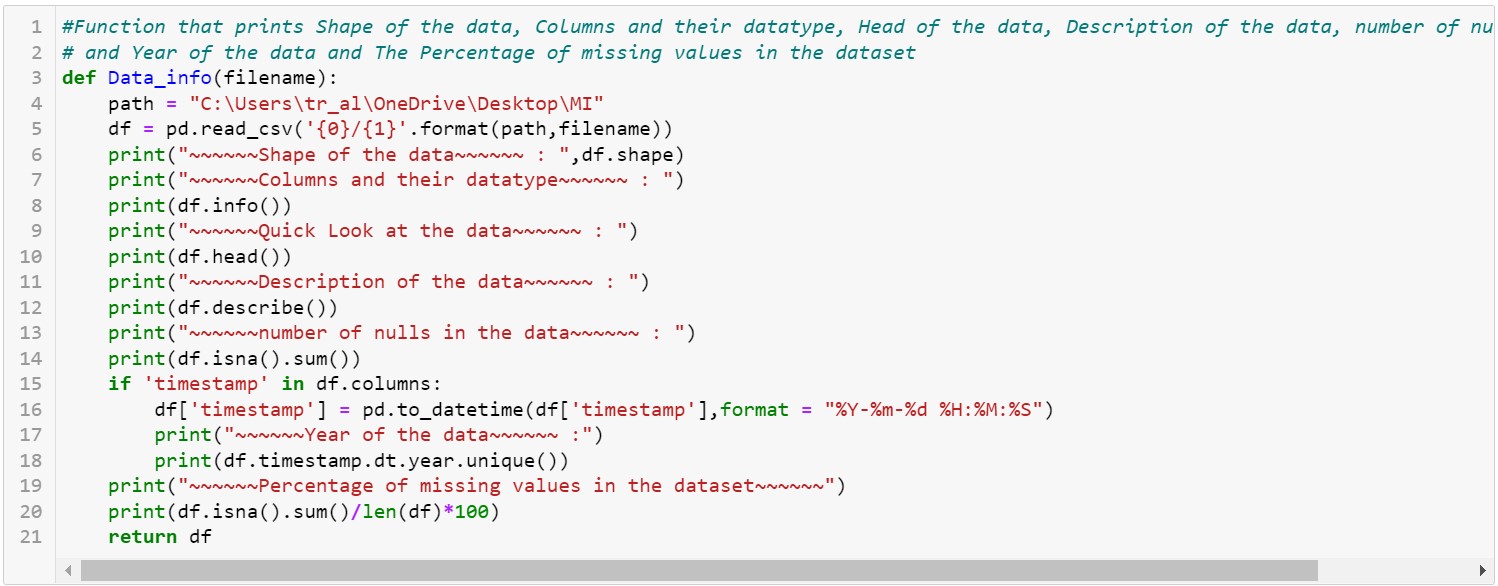
The data used were passed to a memory usage optimizer in order to decrease the overall size of the dataset, for easy and flexible computations. This function can decrease the size of data by changing the data type for a more suitable one and decrease the waste in memory due to bigger datatype being reserved for no purpose.



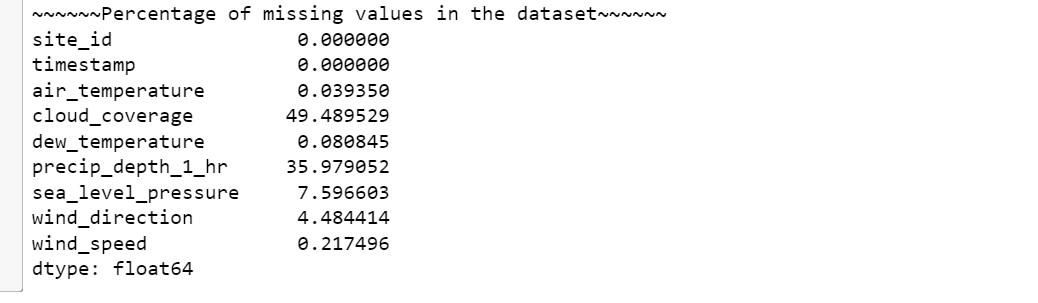
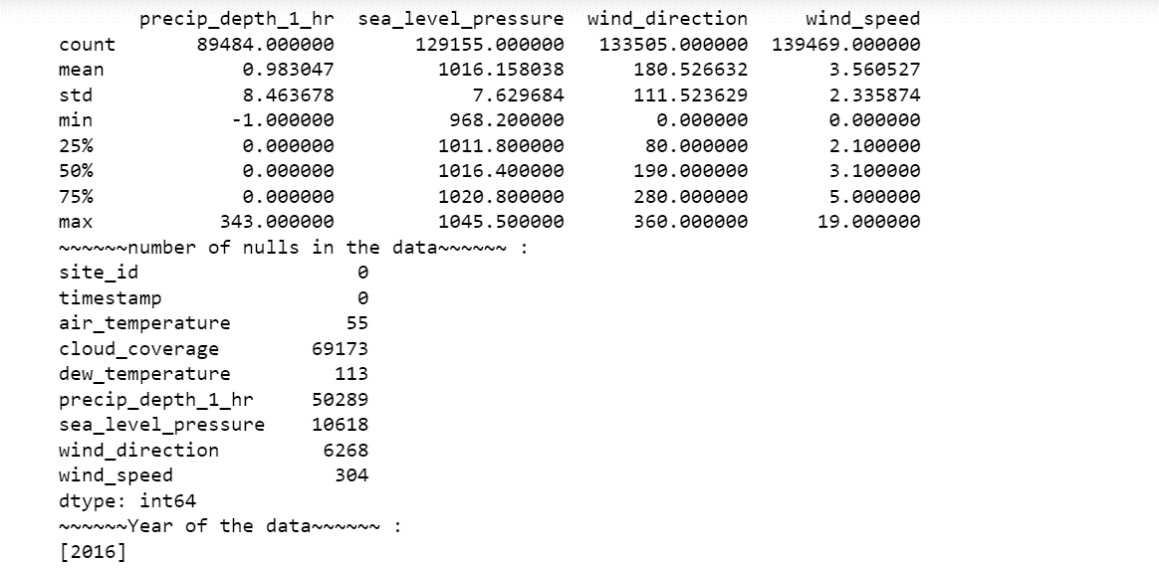
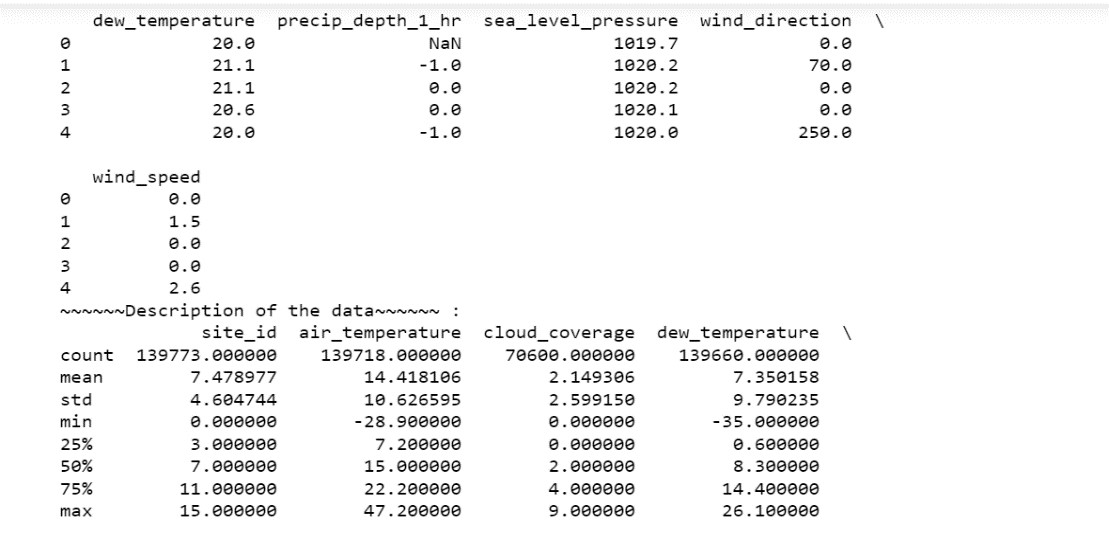
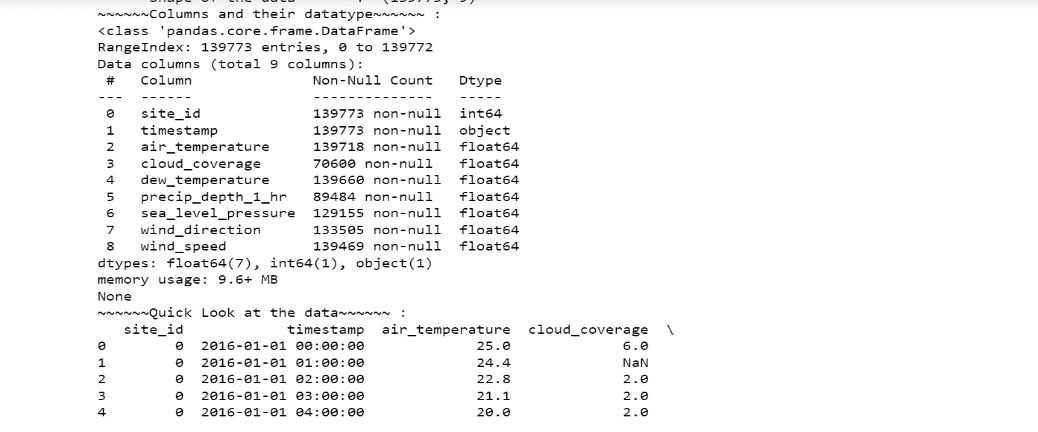
## Data Description

In this part we described data, Columns and their datatype, Head of the data, Description of the data, number of nulls in the data and percentage of missing data values in the dataset.

### Code:



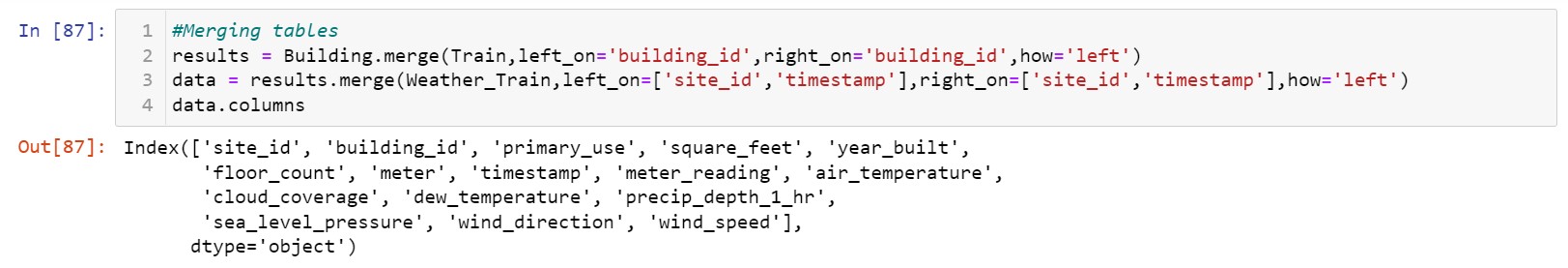
### Output:



## Data Merging

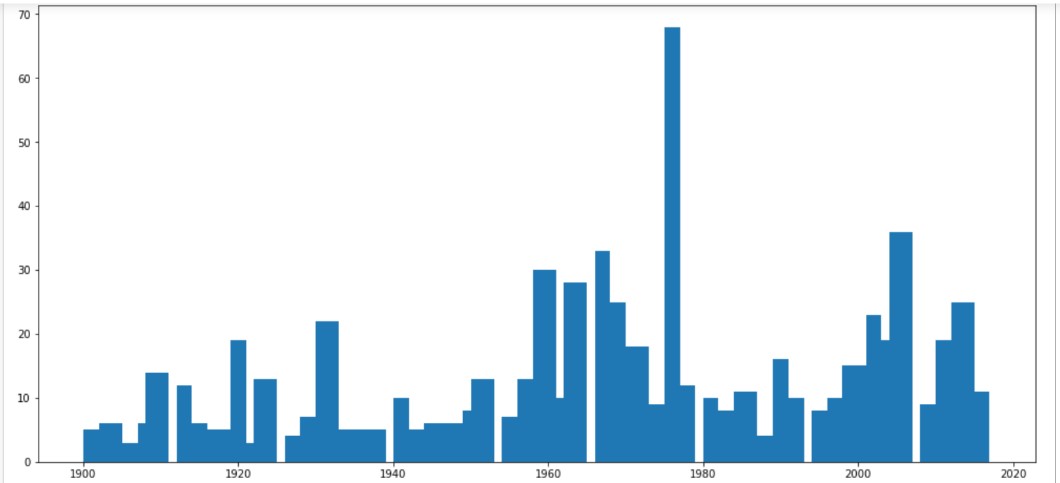
the data is from each table is merged so that it all map to one main table where tests are held and plots. Data from building table (Contains building details) and data from weather table (contains different weather conditions) are merged with a common column between both.

### Code and Output

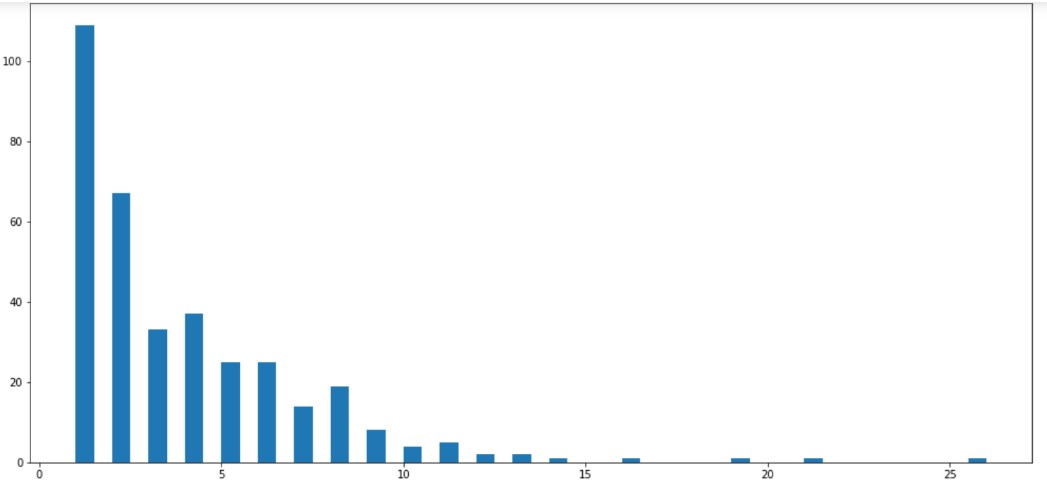


## Histogram Plots

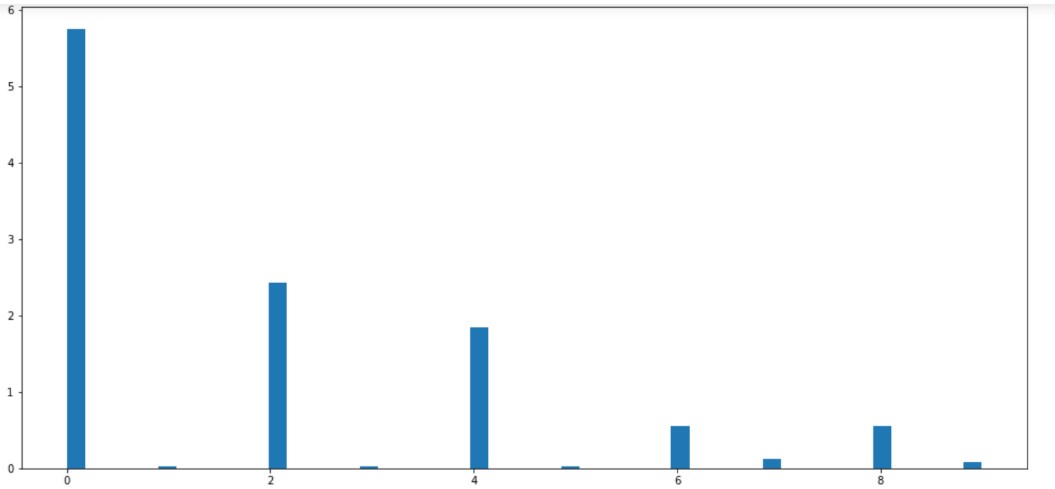
### Year Built



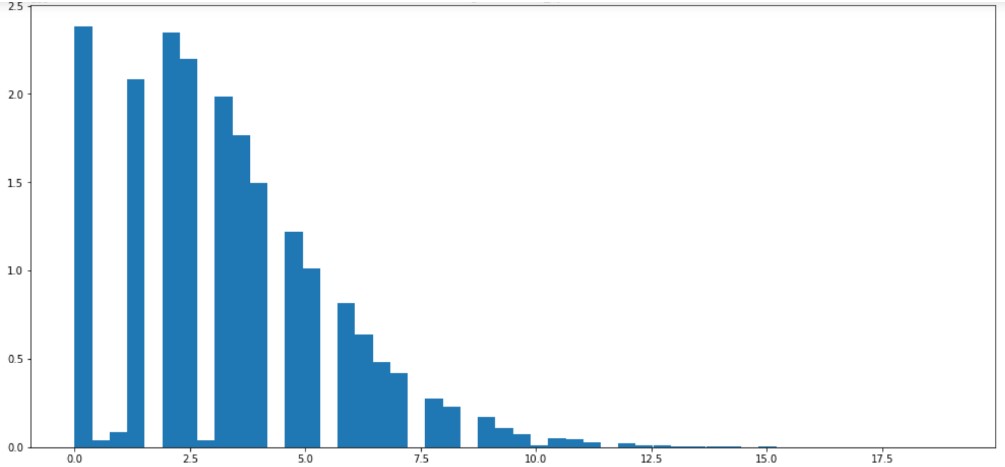
### Floor Count



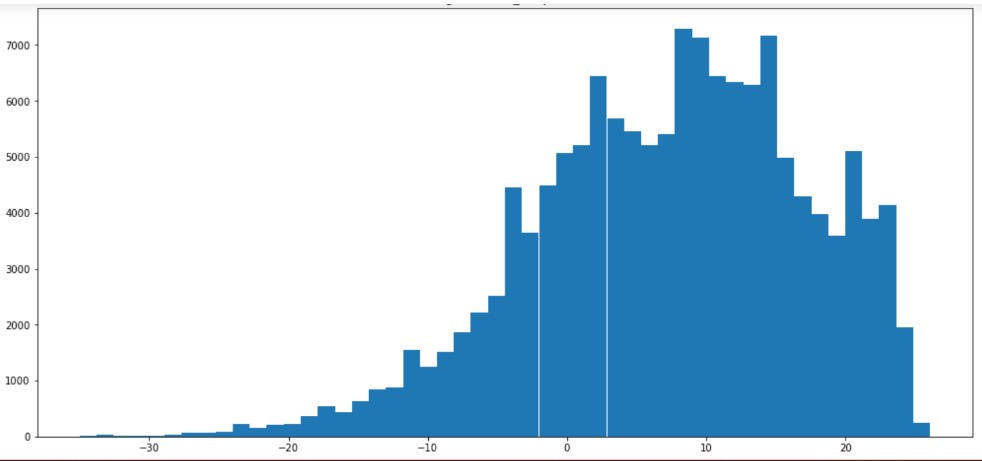
### Cloud Coverage



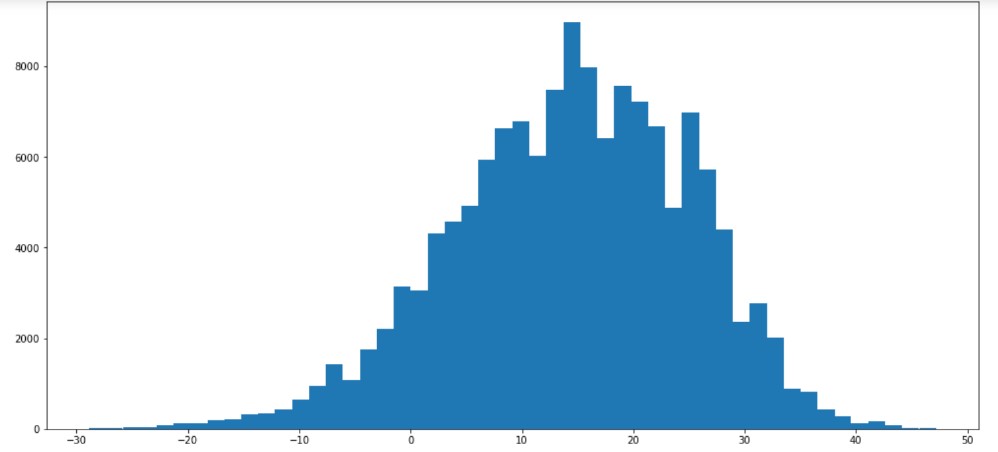
### Wind Speed



### Dew Temperature

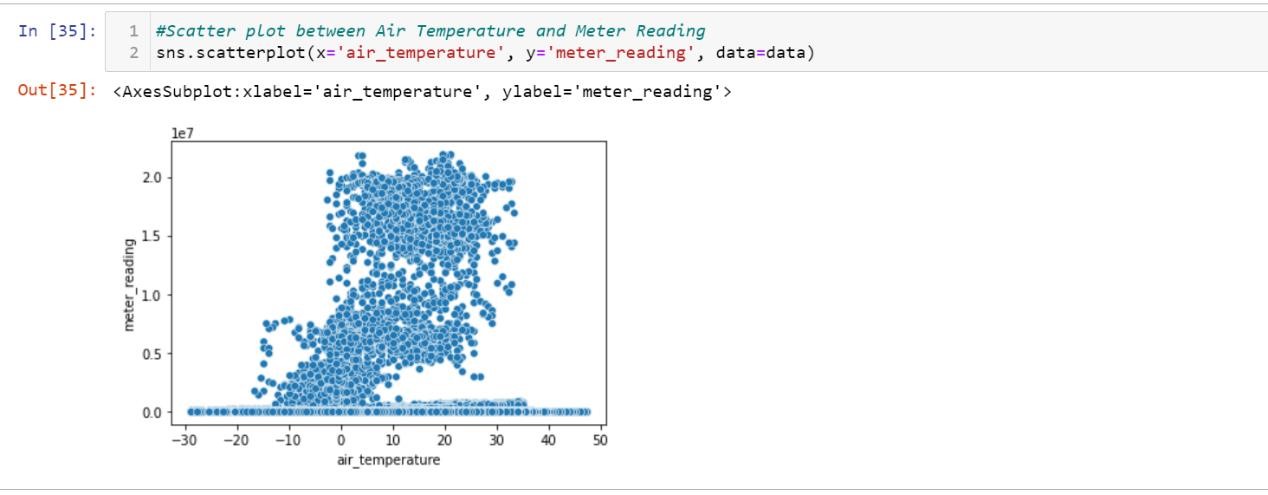


### Air Temperature

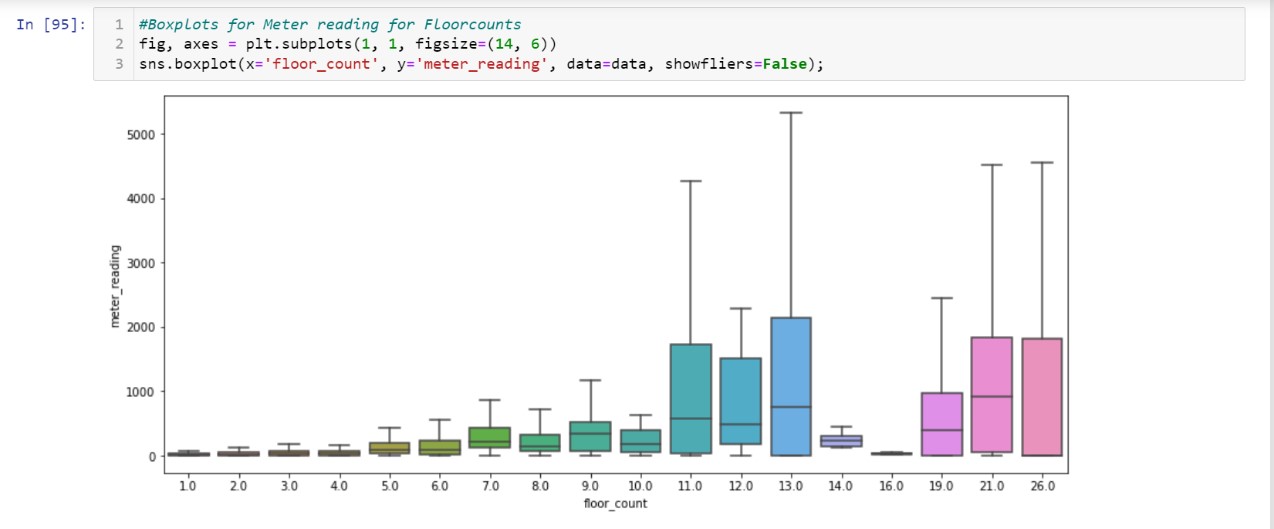


## Additional Plots

### Scatter Plot



### Box Plot



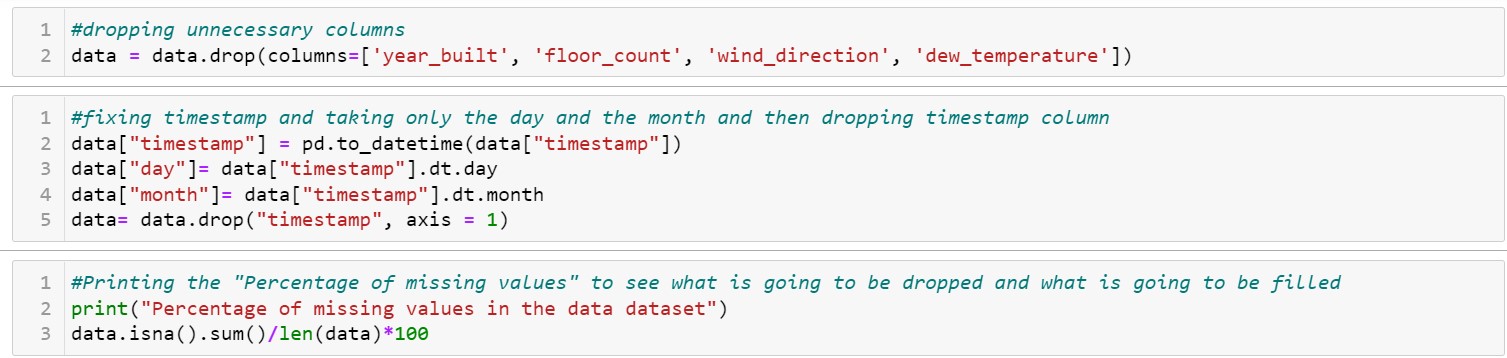
### Correlation Heatmap



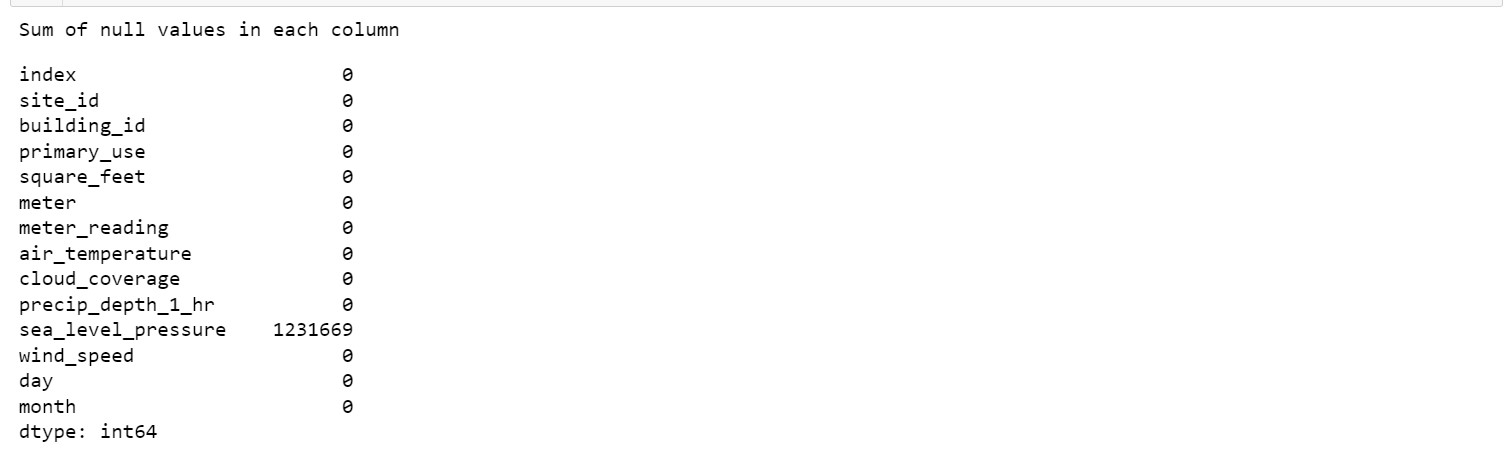
## Data Cleaning

In this process we applied the following codes for data cleaning process such as dropping unnecessary columns, fixing timestamp, and taking only the day and the month and then dropping timestamp column & finding the percentage of missing values to see what is going to be dropped and what is going to be filled.

### Code



### Results



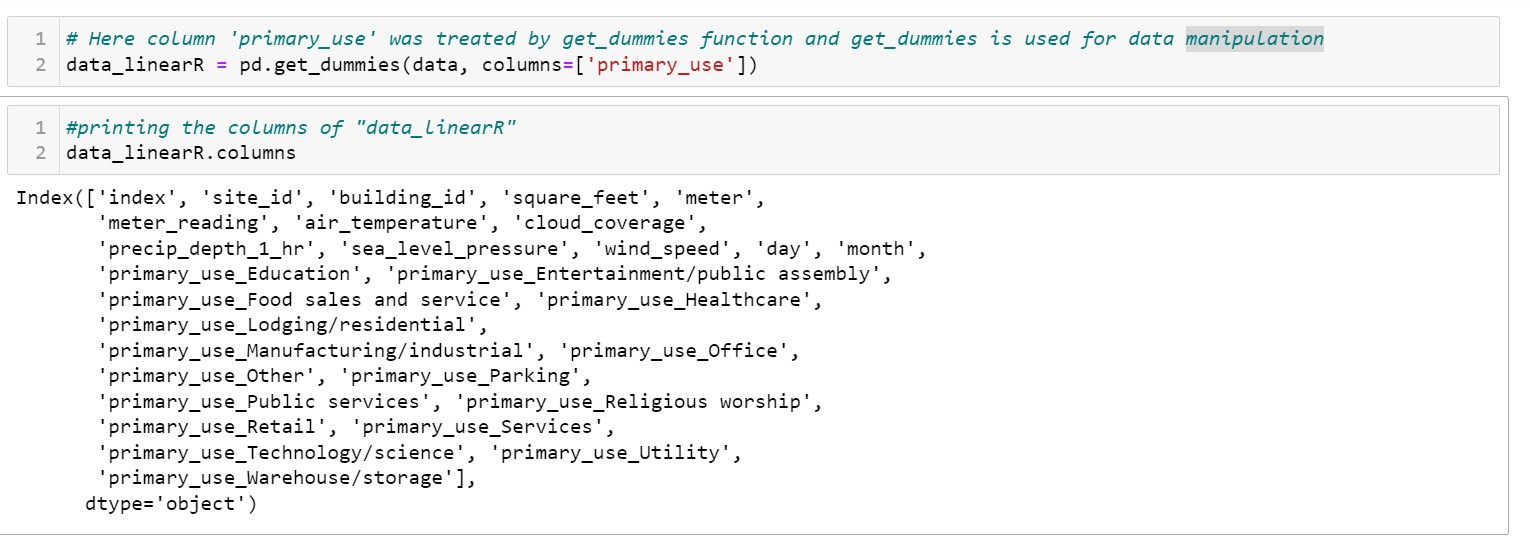
# Linear Regression Model

Linear regression is a **linear model**, e.g., a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

In our project we first manipulated the data that we will extract the features from, to be more eligible for our prediction Linear model.

## Data Manipulation

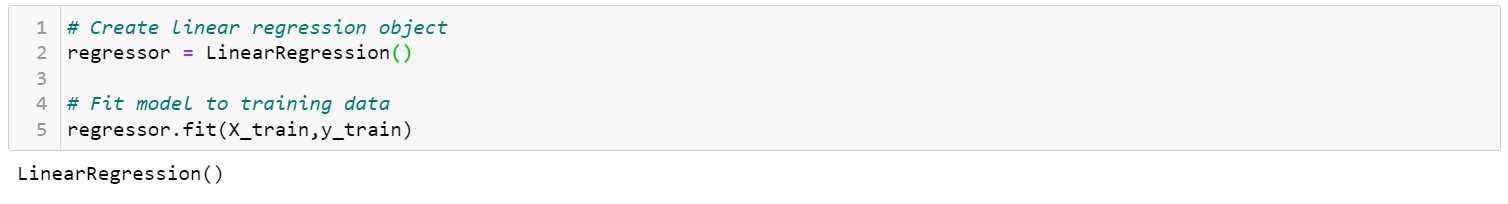
In this part we used get\_dummies function for the “Primary Use” column that contain the important features that will be used for prediction. The use of this function is mainly for data manipulation by converting categorial data into dummies or indicator variables.



In this part the important features are assigned as the X & Y of the dataset so that they can be used as train & test for the linear model.

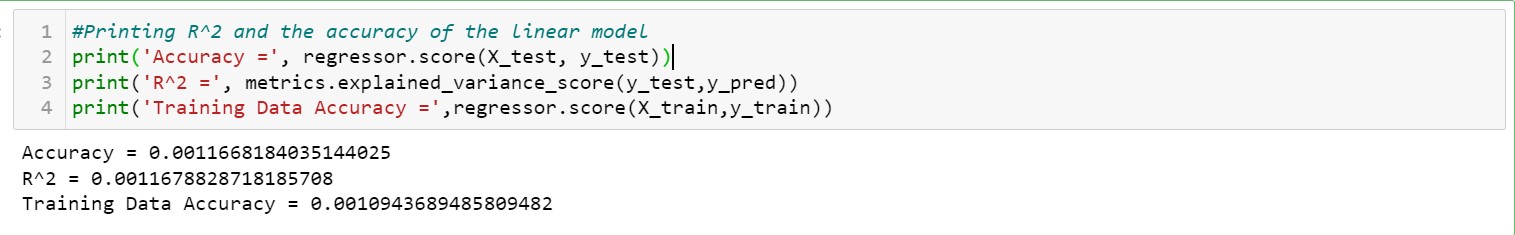


## Linear Regression Object creation



## Linear Regression fitting on test data

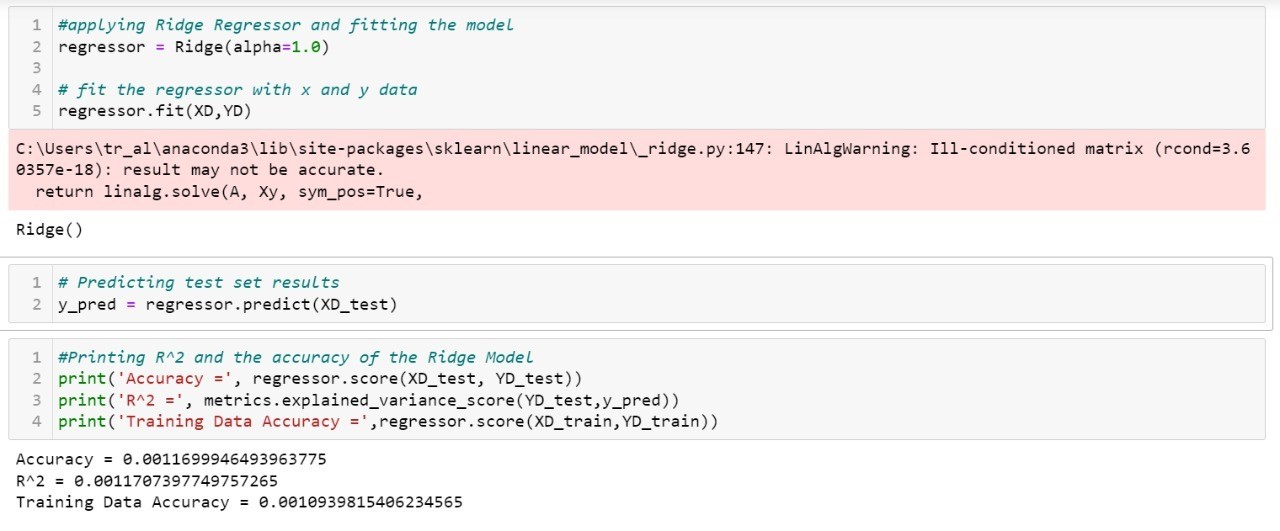
Data were fitted to the linear regression model to get the accuracy of the algorithm the R^2 (coefficient of determination)



# Ridge

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

The ridge implementation on this dataset was very simple, although this model was very poor, it was slightly better than linear regression. Below is the implementation in code and the output accuracy.

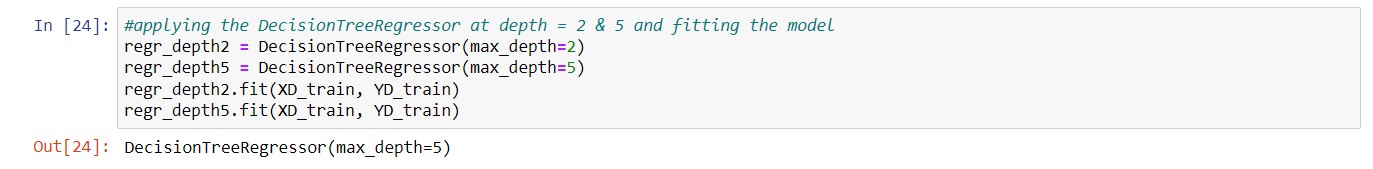


# Decision Tree

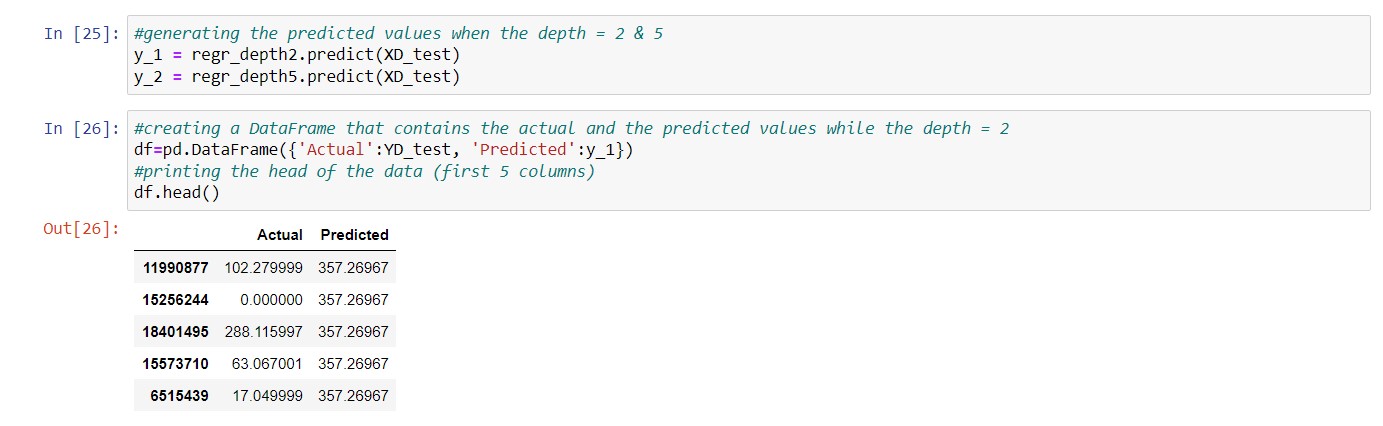
Decision Trees are a type of supervised machine learning in which the data is continuously split according to a parameter (by describing what the input is and what the related output is in the training data). Two entities, decision nodes and leaves, can be used to explain the tree. The decisions or ultimate outcomes are represented by the leaves. And the data is separated at the decision nodes.

## Decision tree regression

Decision tree regression analyses an object's attributes and trains a model in the shape of a tree to forecast future data and create meaningful continuous output. The output/result is not discrete, in the sense that it is not represented solely by a discrete, known set of numbers or values.

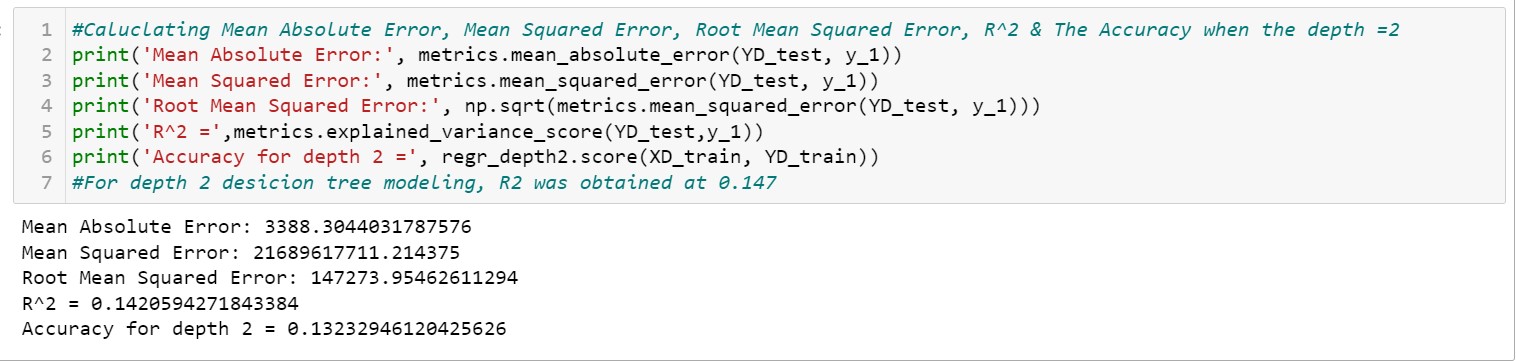


## Generating the predicted values and creating data frame

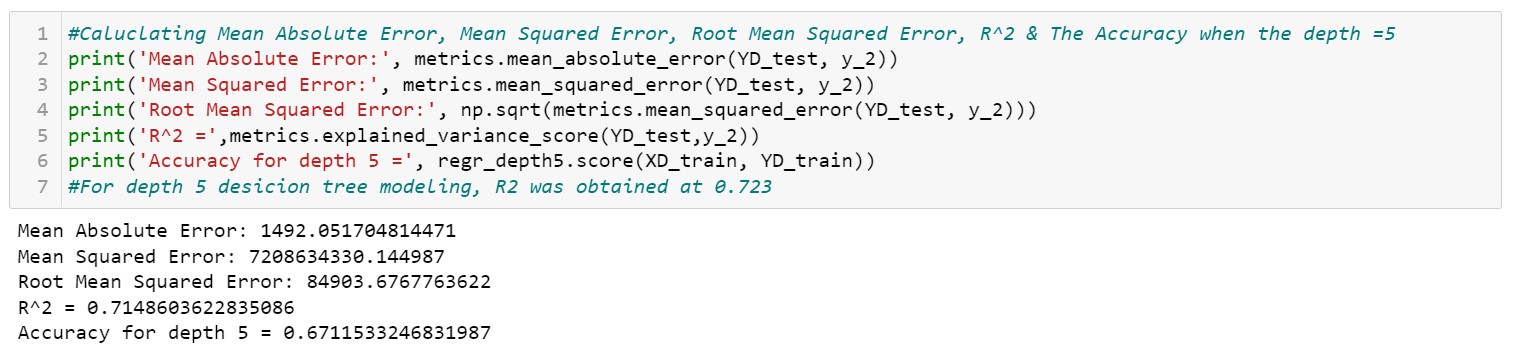


## Calculating errors & accuracy

### At depth = 2



### At depth =5



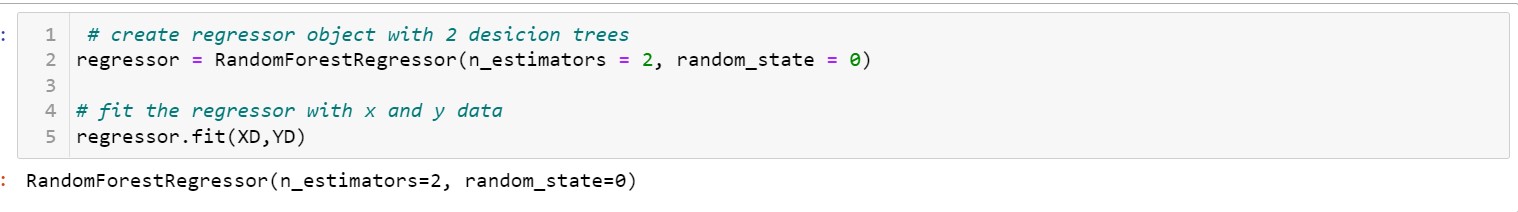
# Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

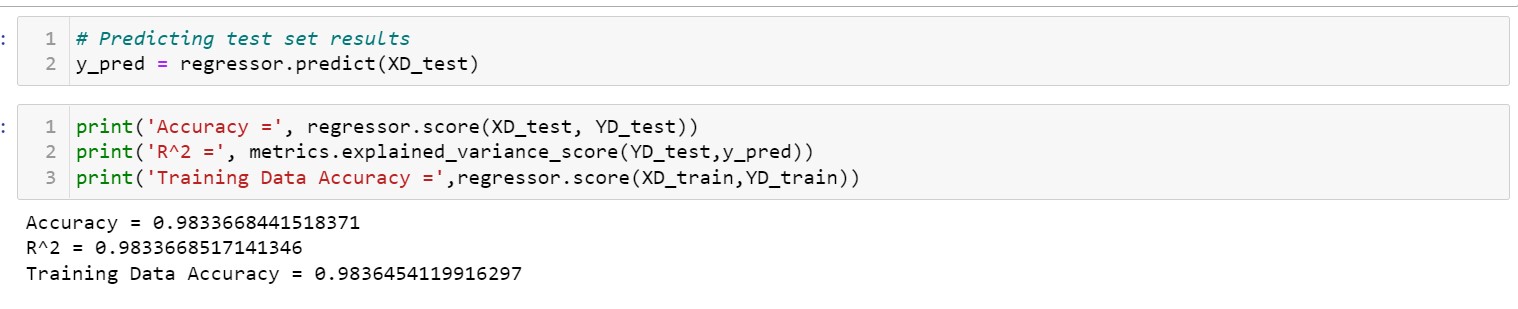
One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

The random forest algorithm relies on having more than one class of different decision trees that vote for the classification. In our implementation we used 2 decision trees to get higher accuracy than a single decision tree implementation.

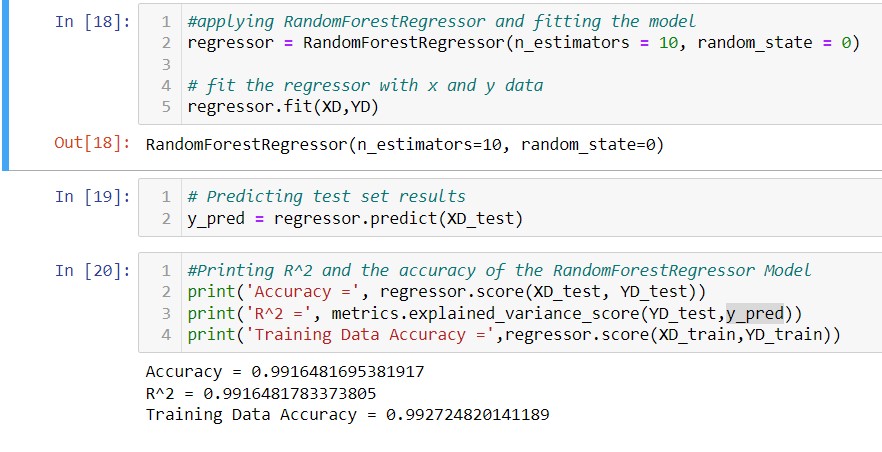
## Random Forest Regression Object creation



## Random Forest Regression data fitting



## Random Forest Regression with 10 estimators



# Results

After applying three different machine learning techniques we were able to get different accuracy scores for each of them that can give us the best results on application when predicting results.

The first approach used which is the **Linear Regression** model gave very low accuracy score R^2 which make it unreliable technique compared to other ones. It is only higher than the ridge regression technique but with a slightly different which make it still an unreliable technique. The second approach used was the **Decision Tree** algorithm at different levels of depth so that we can compare the results according to the depth of the tree and see how it affects the accuracy score. The tests were carried out depth 2 & 5, at depth 2 the results showed a way better accuracy score than the linear regression model making it a more efficient technique for prediction. Further improvement of the technique is using a higher depth of the tree which is 5 that gave a higher accuracy score.

The third approach used was the **Random Forest** algorithm which gave the highest accuracy score compared to the previous algorithm as it uses n-number of decision trees and claim voting from each one for classification. We used 2 decision trees as an estimator to examine the score and it gave a very high accuracy score.

## Table of Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Linear Regression | Decision Trees D=5 | Random Forest E=2 |
| Accuracy Score | 0.0011668184035144 | 0.6711533246831987 | 0.9833668441518371 |
| R^2 Score | 0.0011678828718185 | 0.7148603622835086 | 0.9833668517141346 |

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Ridge Regression | Decision Trees D=2 | Random Forest E=10 |
| Accuracy Score | 0.0011699946493964 | 0.1323294612042562 | 0.9916481695381917 |
| R^2 Score | 0.0011707397749758 | 0.1420594271843384 | 0.9916481783373805 |

# Conclusion

In conclusion the different machine learning techniques used were all aiming to bring about the highest accuracy which indeed gives more accurate and reliable predictions, and as it was discussed throughout the paper that the Highest accuracy score was given by the Random Forest algorithm with higher number of estimators and this is due to the accurate derivation and overfitting avoidance it does so it is the most reliable algorithm among the others.

# References

1. Brownlee, J. (2022). How to Develop Ridge Regression Models in Python. Retrieved 11 June 2022, from <https://machinelearningmastery.com/ridge-regression-with-python/>
2. Brownlee, J. (2022). Linear Regression for Machine Learning. Retrieved 11 June 2022, from <https://machinelearningmastery.com/linear-regression-for-machine-learning/>
3. Chauhan, N. (2022). Decision Tree Algorithm, Explained - KDnuggets. Retrieved 11 June 2022, from <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>
4. Feature Transformation for Multiple Linear Regression in Python. (2022). Retrieved 11 June 2022, from [https://towardsdatascience.com/feature-transformation-for-multiplelinear-regression-in-python-8648ddf070b8](https://towardsdatascience.com/feature-transformation-for-multiple-linear-regression-in-python-8648ddf070b8)
5. Li, H., Li, H., Lamm, M., Rafferty, B., & Coop, A. (2022). Which machine learning algorithm should I use?. Retrieved 11 June 2022, from

[https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learningalgorithm-use/](https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/)

1. Miller, C., Picchetti, B., Fu, C., & Pantelic, J. (2022). Limitations of machine learning for building energy prediction: ASHRAE Great Energy Predictor III Kaggle competition error analysis. Retrieved 11 June 2022, from <https://arxiv.org/abs/2106.13475>
2. Random Forest | Introduction to Random Forest Algorithm. (2022). Retrieved 11 June

2022, from <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>

1. What Is R-Squared?. (2022). Retrieved 11 June 2022, from https://www.investopedia.com/terms/r/r-squared.asp