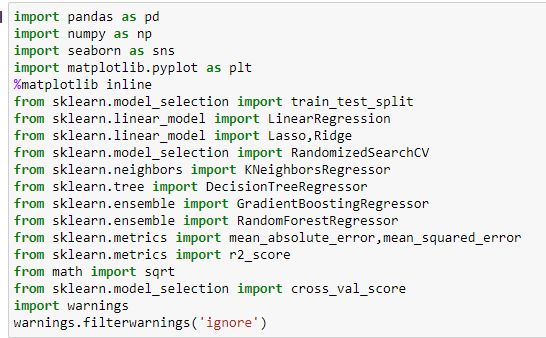
**Problem Statement**

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

**Attribute Information**

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict.   
  
Name / Data Type / Measurement Unit / Description  
-----------------------------  
Sex / nominal / -- / M, F, and I (infant)  
Length / continuous / mm / Longest shell measurement  
Diameter / continuous / mm / perpendicular to length  
Height / continuous / mm / with meat in shell  
Whole weight / continuous / grams / whole abalone  
Shucked weight / continuous / grams / weight of meat  
Viscera weight / continuous / grams / gut weight (after bleeding)  
Shell weight / continuous / grams / after being dried  
Rings / integer / -- / +1.5 gives the age in years.

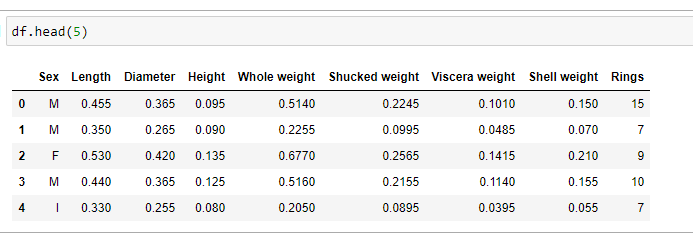
You have to predict the rings of each abalone which will lead us to the age of that abalone.

As our first step we need to import the libraries in our jupyter notebook, for abalone we are imported different libraries

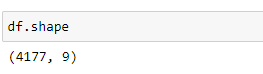
We import pandas, numpy, seaborn, matplotlib, model selection and some machine learning algorithms and some metrics to know accuracy and finally we imported warnings (to ignore warnings)

After we import, we need to upload our data into our file by using pandas reading file. Here we using csv data file

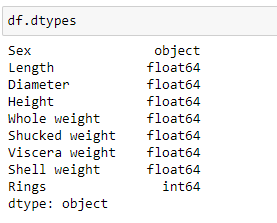
By using head function, we can see our data



Shape function gives us dimension of dataframe. In this csv file we have 4177 entries with 9 columns

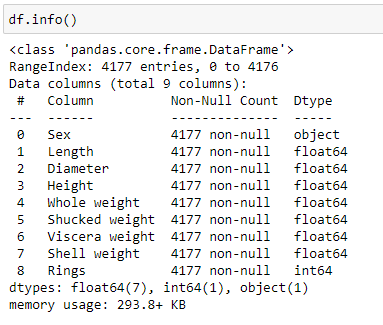


dtypes is to know about the datatypes in each column

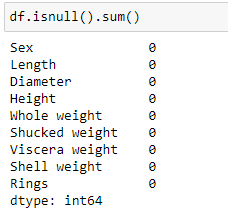


In this we have all columns with numerical data except Sex column, Sex column is object type

Info function is shows that dataframe details in short form all dtypes and count of dtypes and non- null values presented in each column and all the column name like that it gives different details about the dataframe

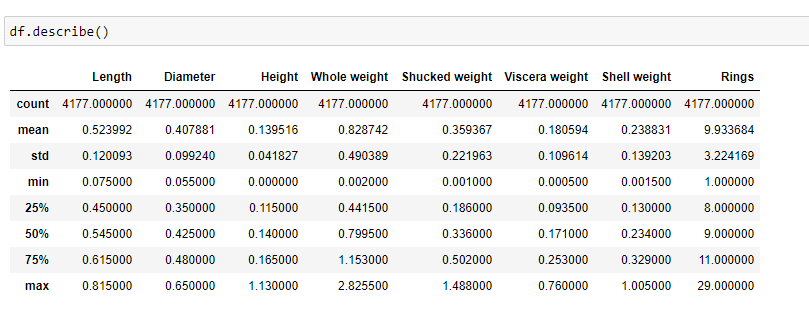


Isnull function is gives that is there any null value are not in the dataframe



As above picture we can see that there is no null value in the dataframe

**Describe** is gives the information of count of variables that used to calculate if there is null values that shows less number of rows than given in dimension, and function also it gives the information of mean, standard deviation, minimum value,25th percentile, 50th percentile(which is median),75th percentile and in last it gives maximum value of each column



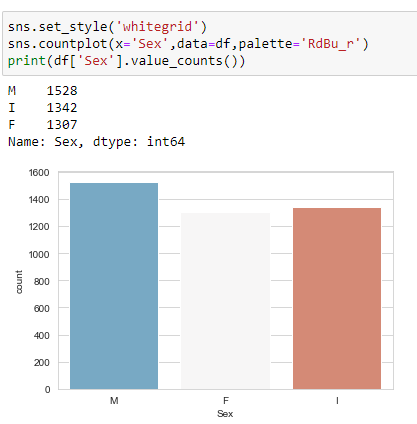
As usual we will compare the mean value and 50th percentile(median) if there is mean is more than median or not and next, we have to concentrate on the difference between 75th percentile and max value if we have more difference or not, this all indicates it has some outliers or skewness are present in the column or not

**Data visualization**

Data visualization is a technique that uses an array of static and interactive visuals within a specific context to help people understand and make sense of large amounts of data. The data is often displayed in a story format that visualizes patterns, trends and correlations that may otherwise go unnoticed.

we are using two libraries

1.matplotlib.pyplot 2.seaborn



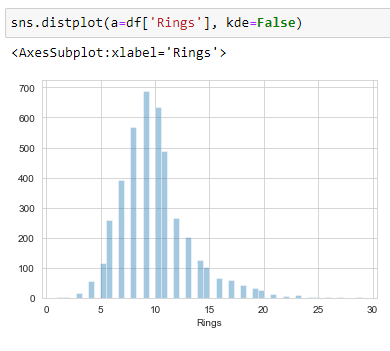
Here I use count plot to show the Sex column that how many categorise and counting categories. Here we have 3 different categorise Male as M, Female as F, Infant as I

Here we have

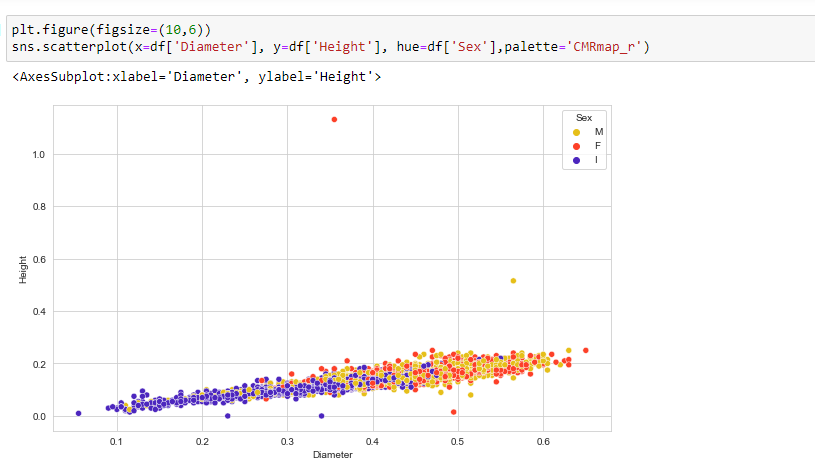
M 1528

I 1342

F 1307

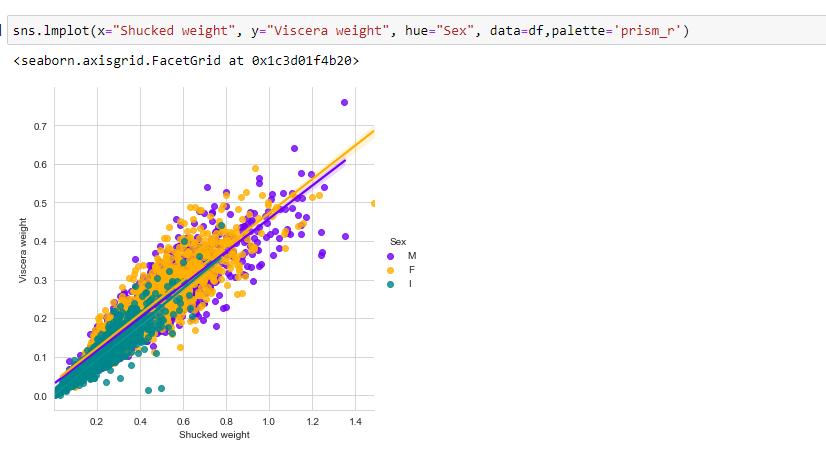
Here we use distplot from seaborn, we plot rings in this plot. For this projects ring is target variable here rings are indicating age of shells

As we can see that the most of the shells are in 5 to 20 rings



**Scatter plots relationships between two numeric variables, the dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole, identification of correlation relationship are common with scatter plots**

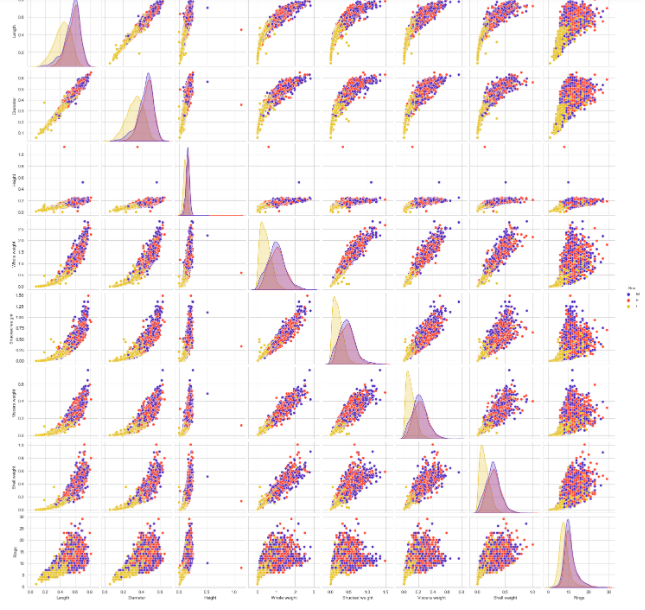
Here we can see the scatter plot of comparing Diameter and Height and here we use to hue to plot, hue is Sex column. We can see the good relation in the both columns and we can see outliers are present in the column

 **lmplot is a 2D scatterplot with an optional overlaid regression line. This is useful for comparing numeric variables**

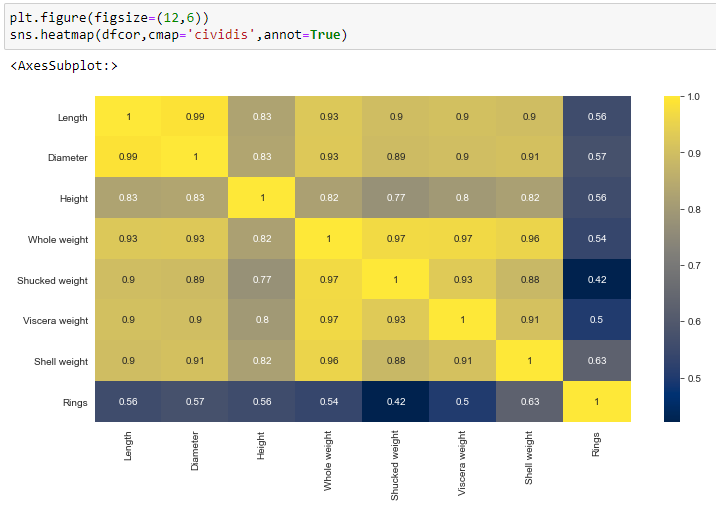
Here we can see the plot that good relation in both columns data and overlaid of regression line

Here the plot is pair plot

**Pair Plots are a really simple (one-line-of-code simple!) way to visualize relationships between each variable**

****

**Checking correlation**



Above plot is shows correlation

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation.

A positive correlation is a relationship between two variables in which both variables either increase or decrease at the same time. An example would be height and weight. Taller people tend to be heavier.

A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).

A zero correlation exists when there is no relationship between two variables. For example, there is no relationship between the amount of tea drunk and level of intelligence.

### Strength of correlation

prefect +1,-1

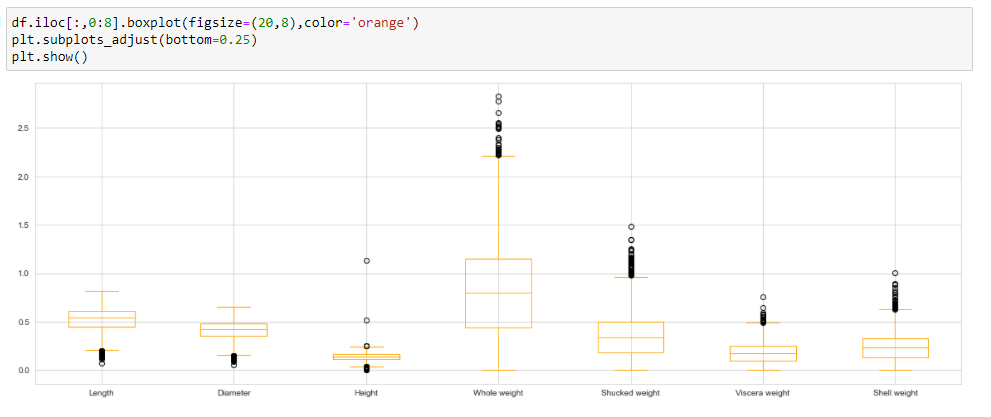
strong +(0.9 to 0.7) & (-0.9 to -0.7)

moderate +(0.6 to 0.4) & (-0.6 to -0.4)

week +(0.3 to 0.1) & (-0.3 to -0.1)

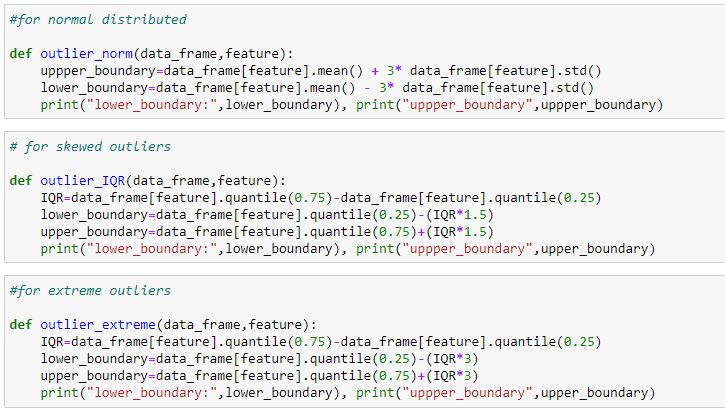
Zero 0

**Checking outliers**



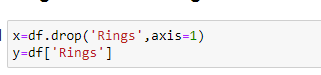
Above box plot shows that we have outliers in the dataframe

We can detail this with different models. Here I use the followed methods



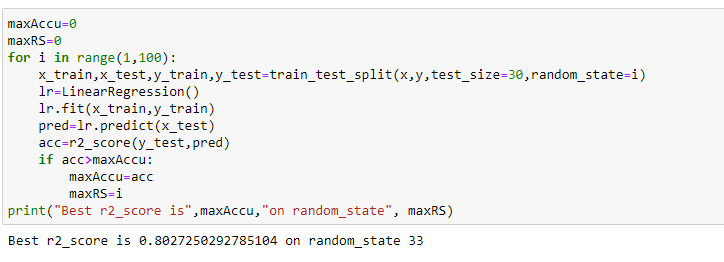
After all these we got the cleaned data and we can go to next step

**Splitting dependent variable and independent variables**

****in the dataframe, Rings is target variables and remaining all the columns are independent features

Next

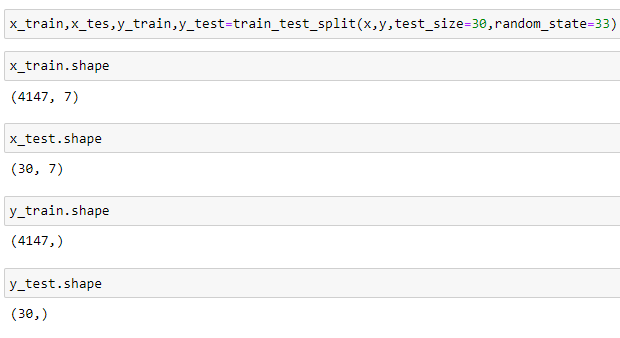
Here I try to find the best random state for good accuracy

****

After applying the above code, we got 33 as the best random state and got good accuracy

Next step is train\_test\_split

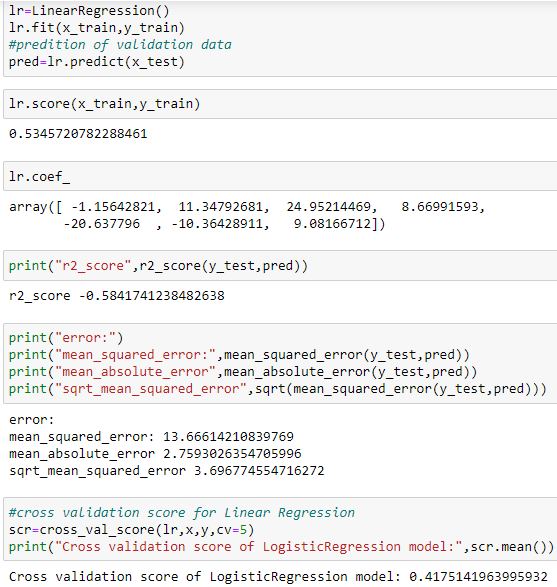
Here we train the data and test with some of it data only for this we need to split the by using sklearn.model\_selction



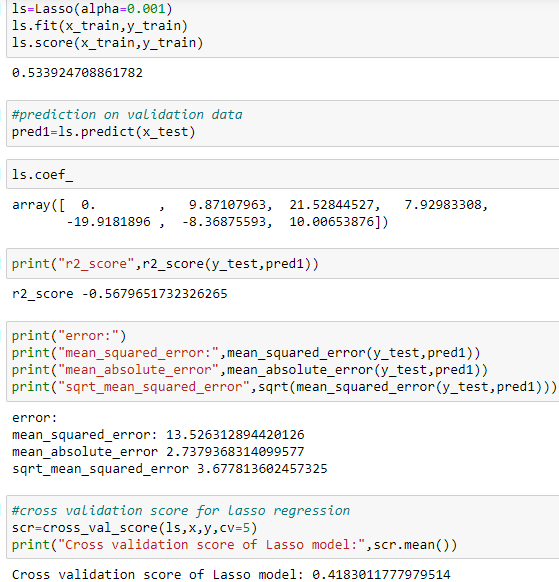
**Building Machine Learning Models**

Here I used different machine learning algorithms

* Linear Regression
* Lasso
* Ridge
* KNeighborsRegressor
* DecissionTreeRegressor,
* GradientBoostingRegressor

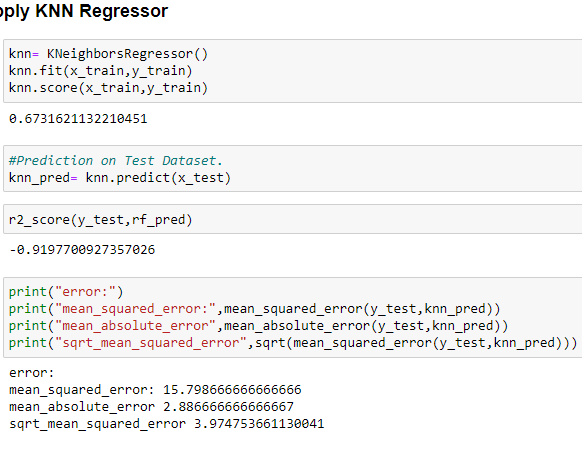
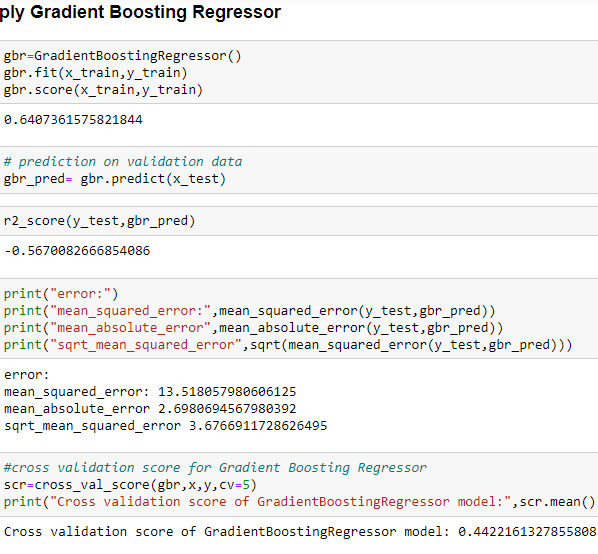


**Linear Regression**

I started by using linear regression in order to model the relationship between the features and the target variable. MAE, MSE and RMSE results were in terms of MPa (unit for the compressive strength) and they tell us that, the difference between the observed and predicted results are 2.76(MAE), 13.66(MSE) and 3.69(RMSE). R2-score of -0.5841 

**Lasso**

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.

**Ridge**

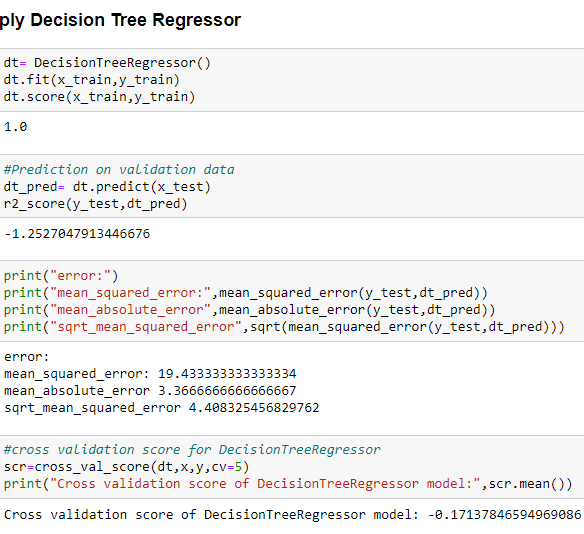
Ridge Regression is defined as a model tuning method that is use to analyse any data that suffers from multicollinearity. Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.

**Gradient Boosting Regressor**

Gradient Boosting Regressor builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

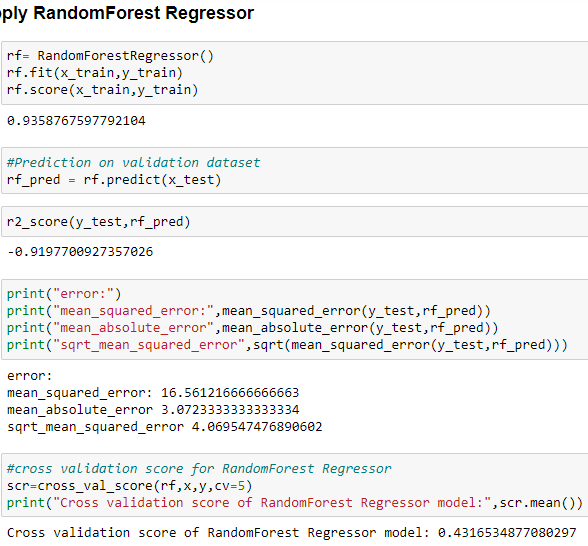
**KNN regressor**

The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set.

**Decision Tree Regressor**

The [decision trees](https://scikit-learn.org/stable/modules/tree.html#tree) is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve.

We can see that if the maximum depth of the tree (controlled by the max\_depth parameter) is set too high, the decision trees learn too fine details of the training data and learn from the noise, i.e. they overfit.

****

**Random Forest Regressor**

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.

**Conclusion**

After analysing data and cleaning and modelling the data at the end we need to conclude that which is the best suitable mode for the project. We can’t take only r2\_score value we have to check the cross-validation score also, at the end here my best solution is Gradient Boost Regressor