**Abstract:**

Abalone Case Study

Abalone is a marine snail found in the cold coastal regions. Age is a vital characteristic that is used to determine its worth. Currently, the only viable solution to determine the age of abalone is through very detailed steps in a laboratory. Now we will use various Machine Learning algorithms for abalone age prediction.

**Introduction:**

Abalones are types of single-shelled marine snails found in the cold coastal waters worldwide, majorly found along the coastal regions of some countries such as Australia, Western North America, South Africa, New Zealand, and Japan. The age of the abalone is highly correlated to its prices as it is the sole factor used to determine its worth however, determining the age of abalone is a highly involved process that is usually carried out in a laboratory

Knowing the correct price of the abalone is important to both the farmers and consumers. In addition, knowing the correct age is also crucial to environmentalists who seek to protect this endangered species. Due to the inherent inaccuracy in the manual method of counting the rings and thus calculating the age, researchers have tried to employ physical characteristics of the aba-lone such as sex, weight, height, and length to determine its age.

Thus, by applying machine learning on a dataset containing a large number of training samples of physical measurements of abalone, its age can be predicted quickly and more accurately. Machine learning algorithms are data-driven approaches that can effectively recognize certain patterns. Over the last decade, machine learning techniques have been successfully applied across various domains such as for Unicode symbols identification, classification of large data sets, ordinal classification, etc. Among machine learning approaches, the most successful and widely used techniques include Artificial neural networks (ANN), KNN, random forest, Gauss Naïve Bayes and SVM.

**Problem Definition:**

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**

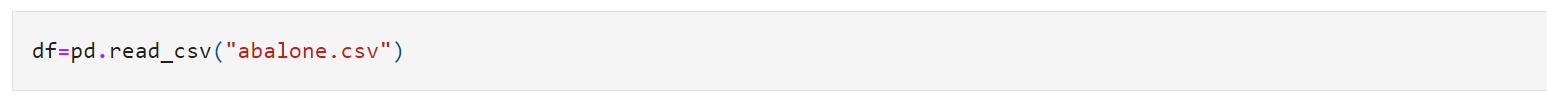
Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict.   
  
Name / Data Type / Measurement Unit / Description  
Sex / nominal / -- / M, F, and I (infant)  
Length / continuous / mm / Longest shell measurement  
Diameter / continuous / mm / perpendicular to length

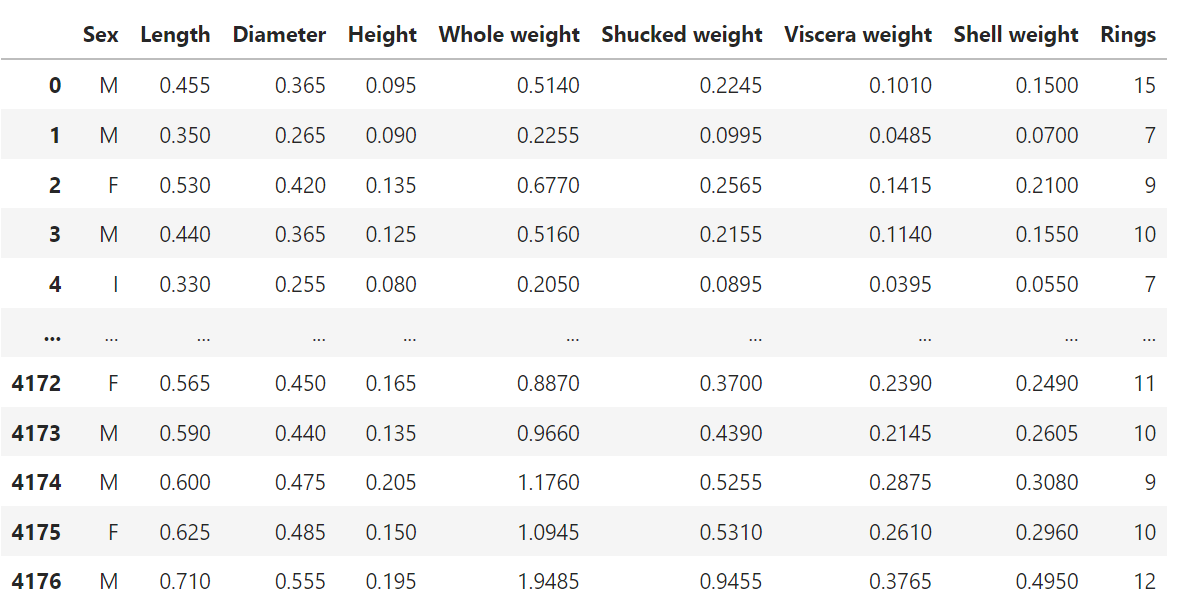
Height / continuous / mm / with meat in shell  
Whole weight / continuous / grams / whole abalone  
Shucked weight / continuous / grams / weight of meat  
Viscera weight / continuous / grams / gut weight (after bleeding)  
Shell weight / continuous / grams / after being dried  
Rings / integer / -- / +1.5 gives the age in years.

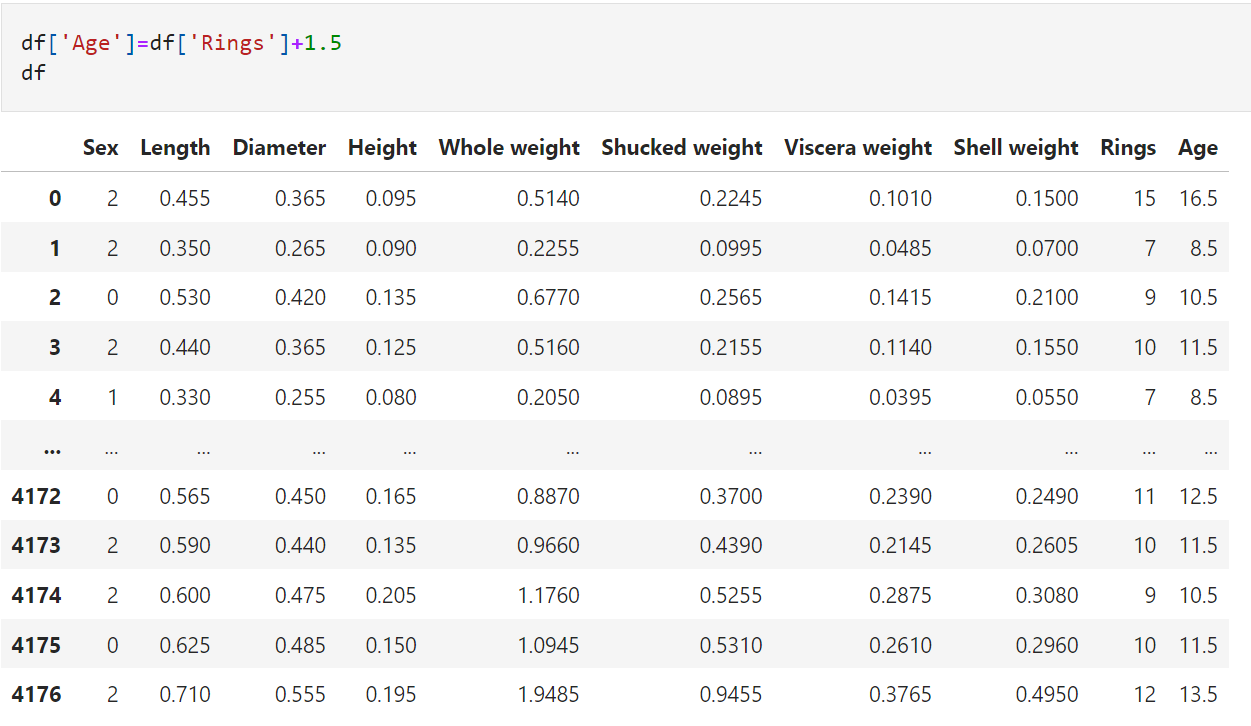
You have to predict the rings of each abalone which will lead us to the age of that abalone.

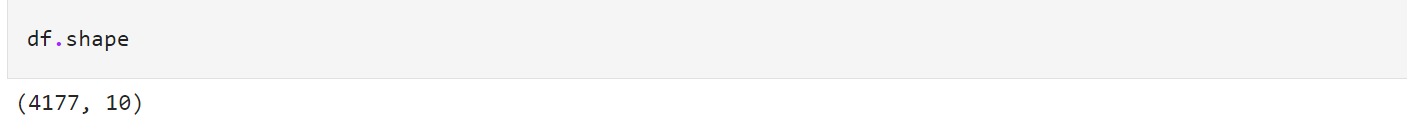
**Data Analysis:**

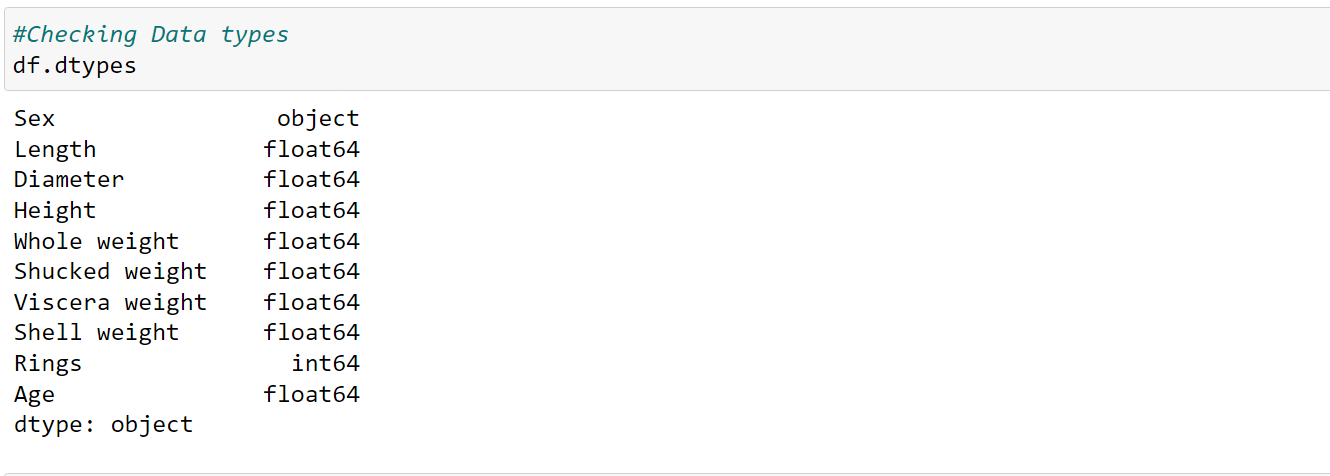
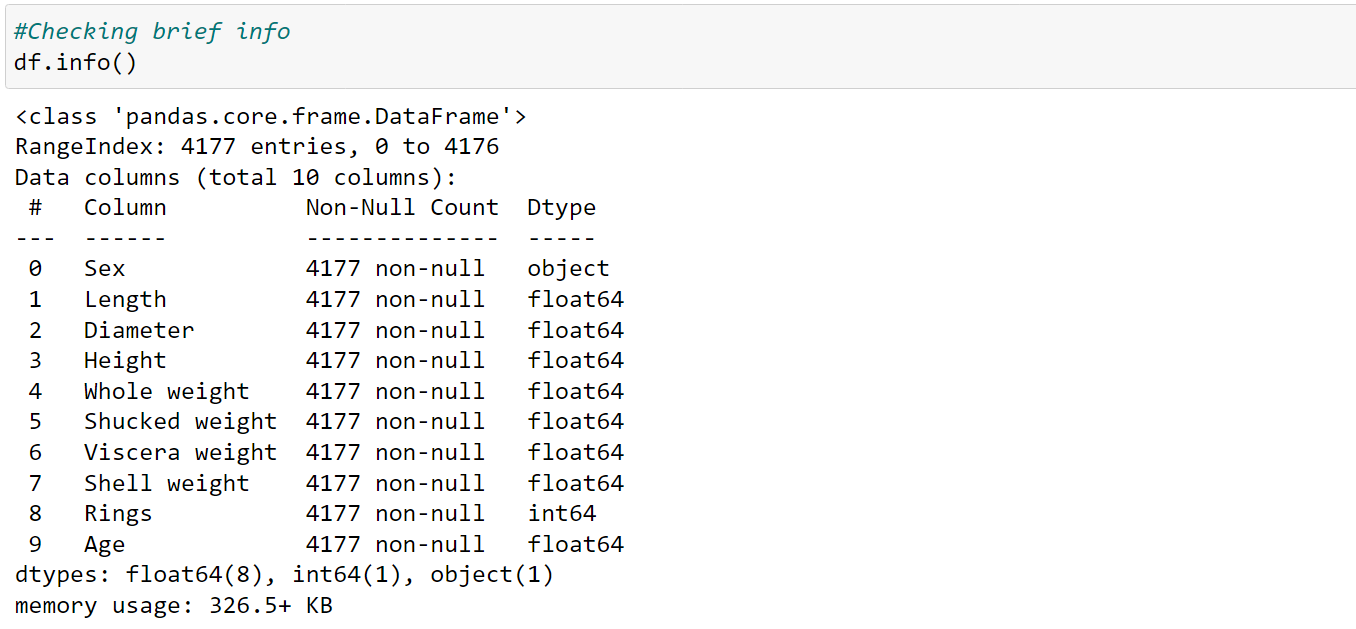
Importing the dataset which is in CSV format using pandas.

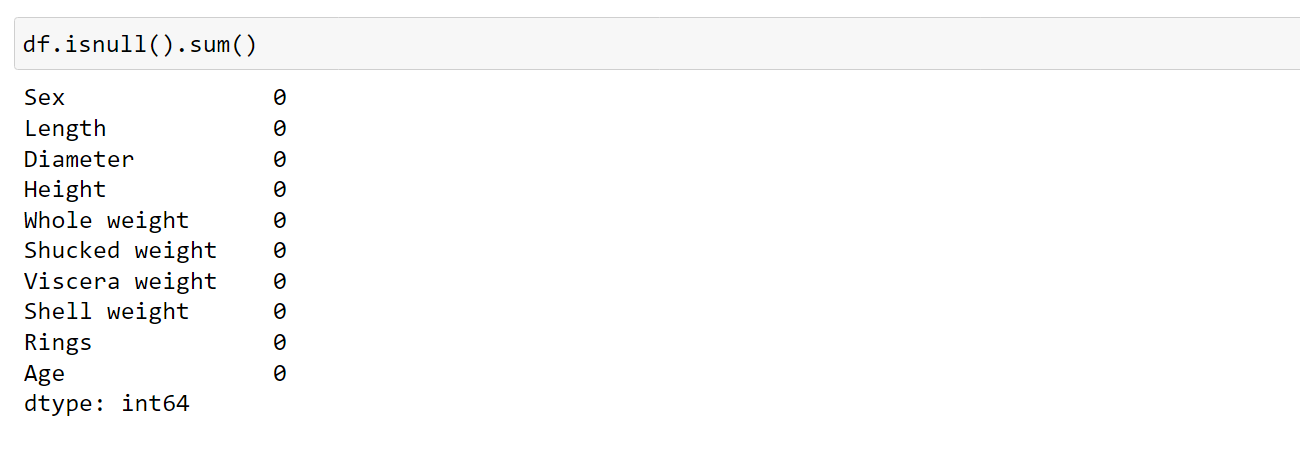
****

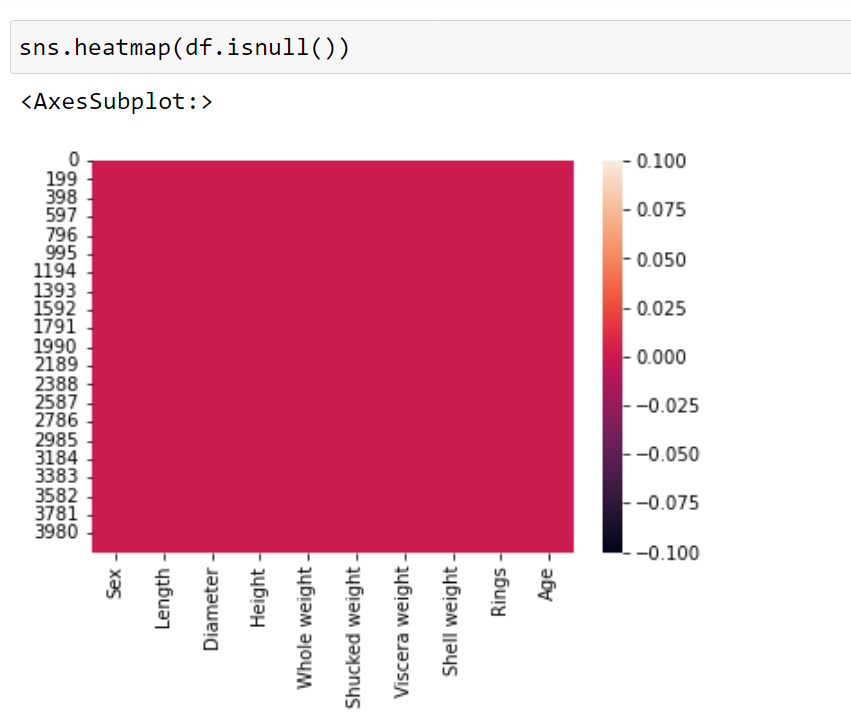
****

Getting access to the rings of an abalone involves cutting the shell. After polishing and staining, a lab technician examines a shell sample under a microscope and counts the rings. Because some rings are hard to make out using this method, 1.5 is traditionally added to the ring count as a reasonable approximation of the age of the abalone. Knowing the correct price of the abalone is important to both the farmers and consumers****

****

The abalone dataset is a dataset that contains measurements of physical characteristics of different abalones. It has 4177 instances.**** ****

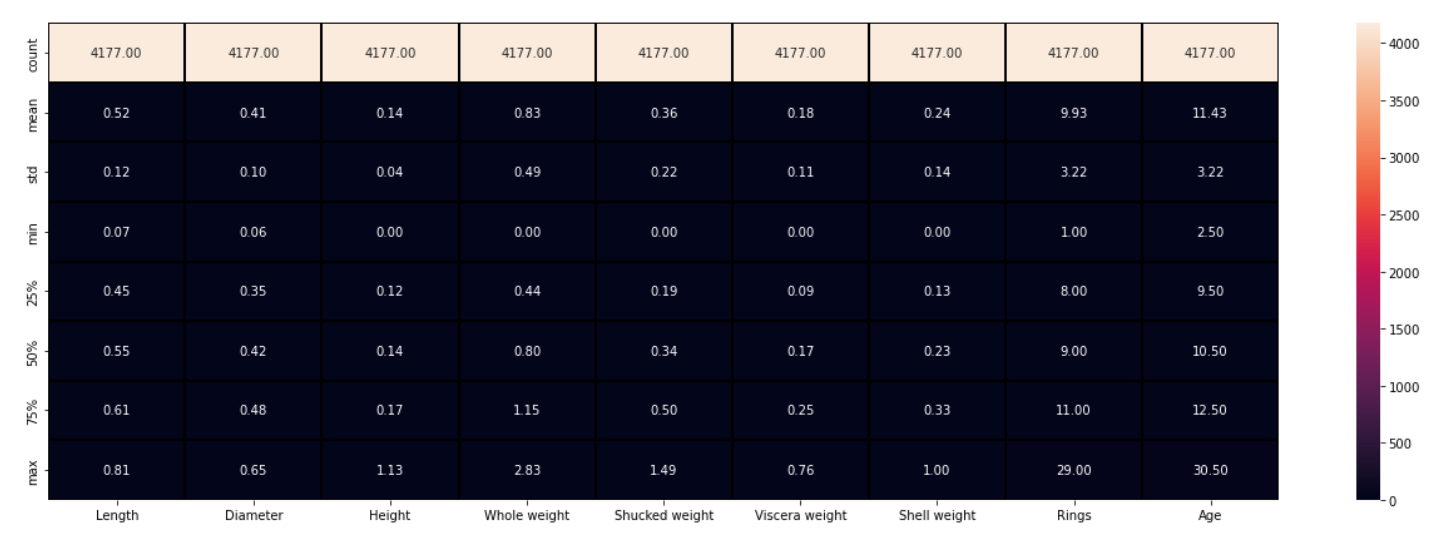


****

It is clearly visible that our Dataset is free from the null values.

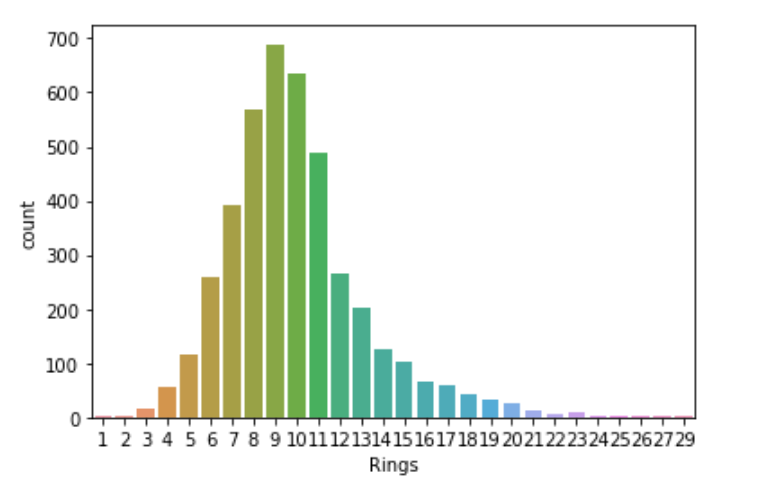
**Brief Statistical Summary of the Dataset:**



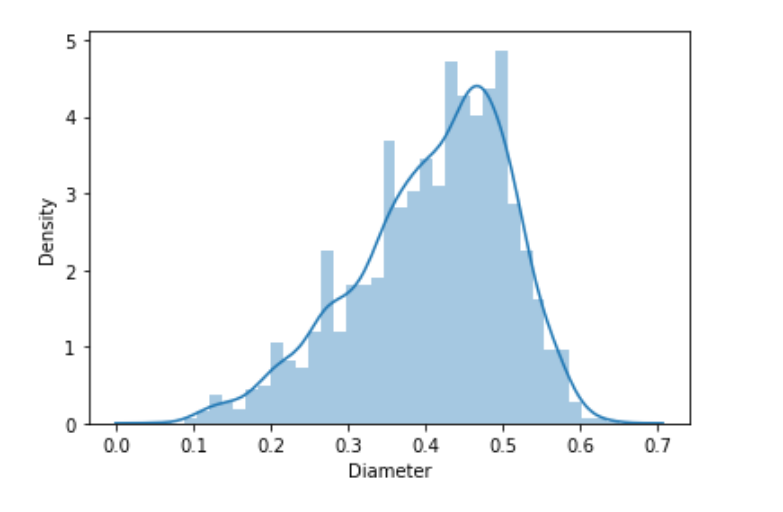
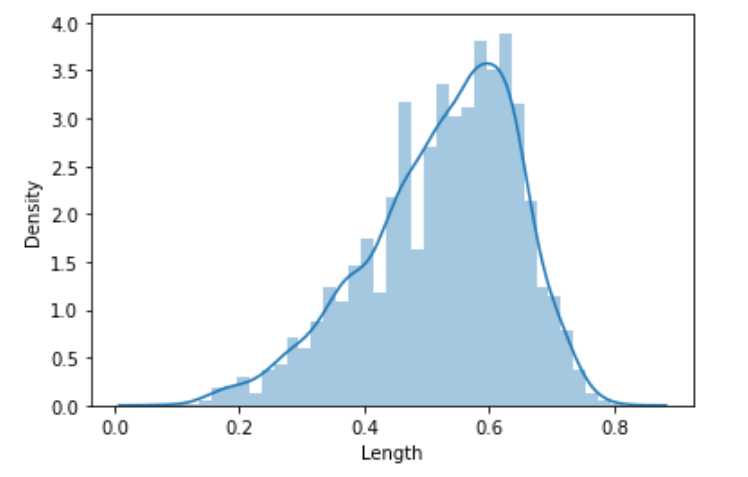


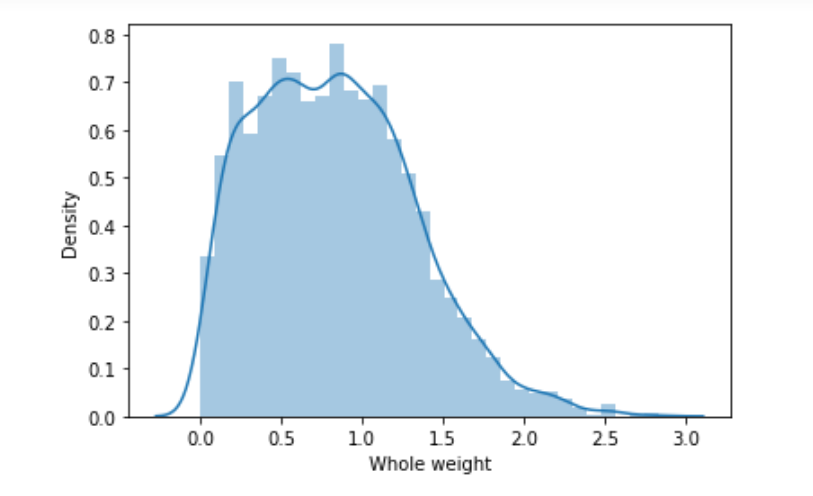
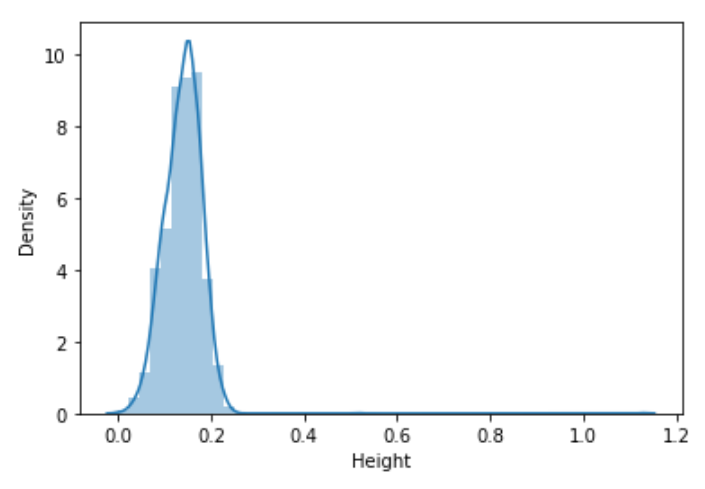
Based on above table and visualisation,we have a clear view of the whole data.Since the dataset is clean,we will use this for machine learning process.

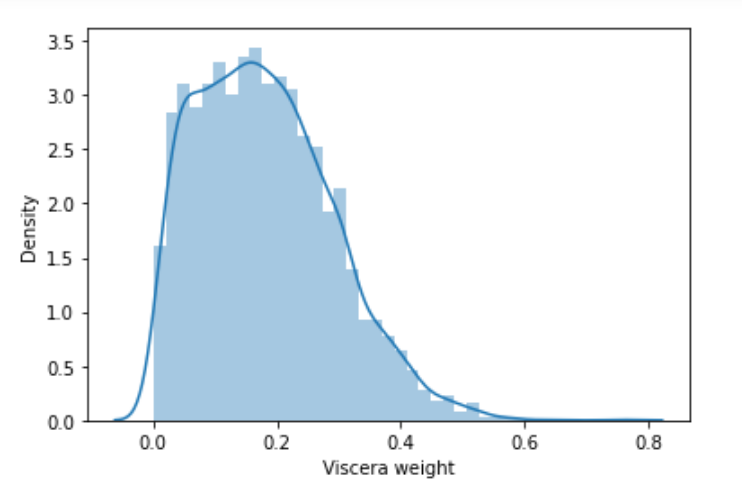
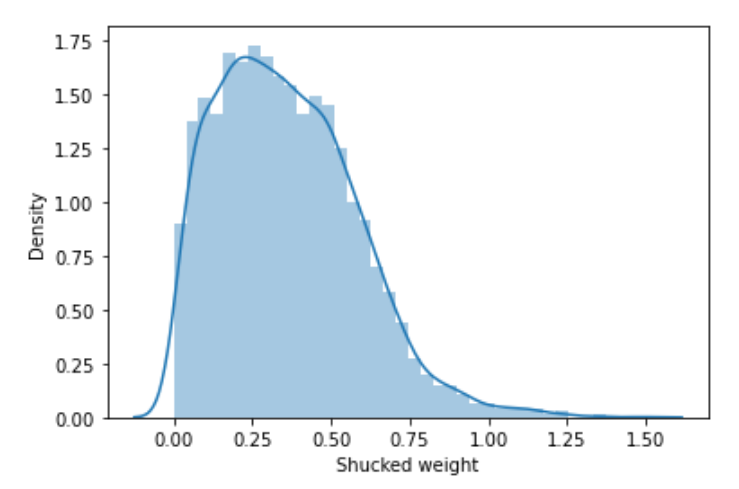
**Data Visualisation:**

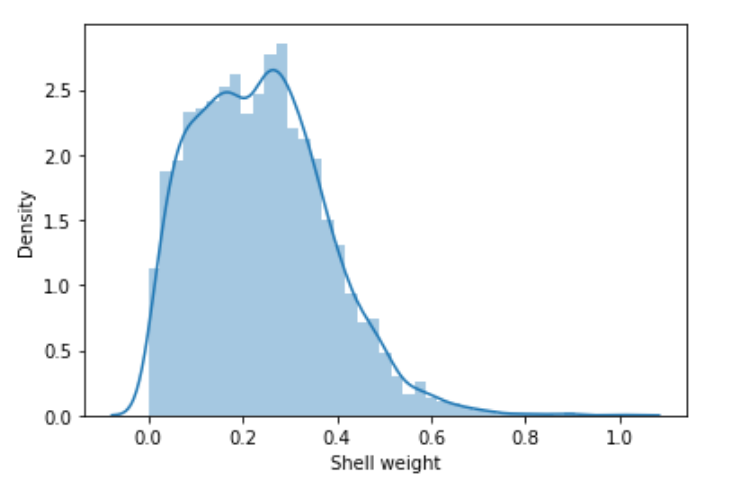
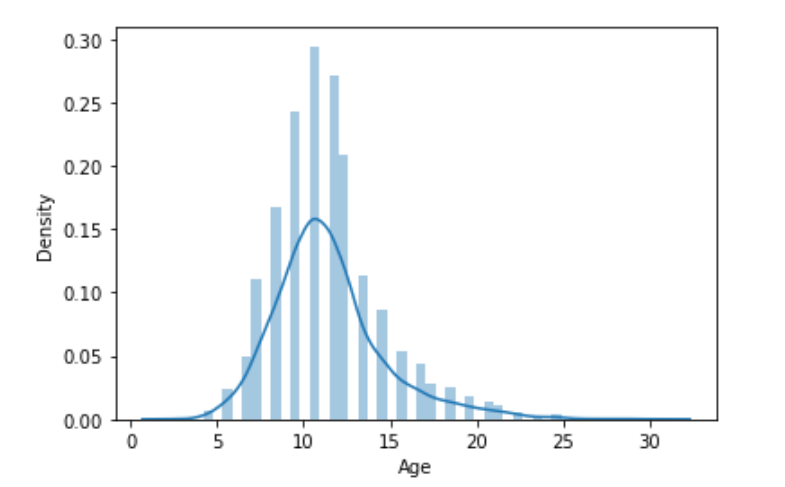
****

The above figure shows the count of rings in the abalone dataset. It can be seen that the dataset is having maximum rings in the range of 7-14 with very few examples having rings above 20.

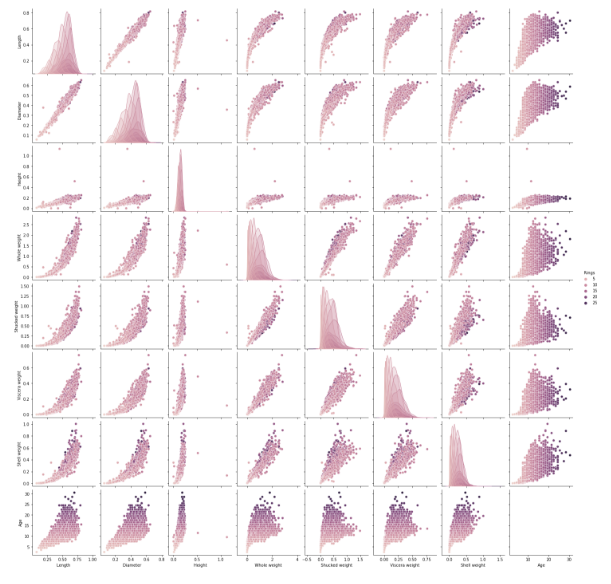
****

****

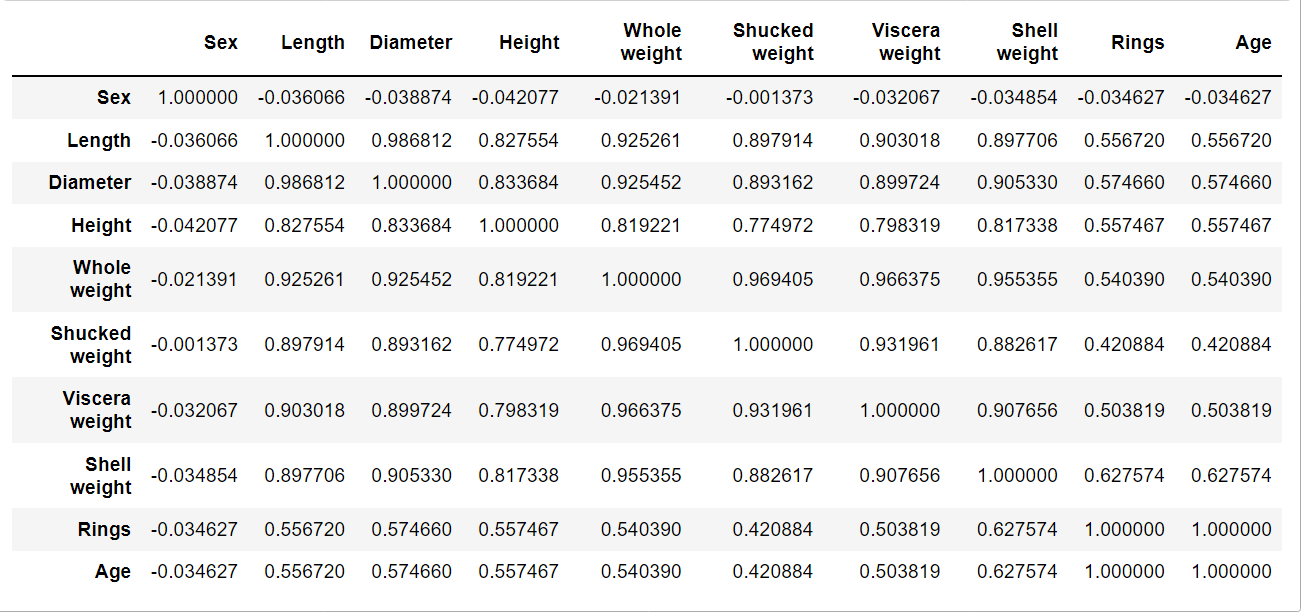
****

The above figures are the distributipn plots,shows the distribution of Length,Diameter,Height,Whole weight,Shucked weight,Viscera weight, Shell weight and Age in the abalone dataset. It can be seen that the distribution of the length and diameter features was left-skewed, while the whole weight, viscera weight, shucked weight, and shell weight appeared to have a right-skewed distribution. The height feature did not have a clear skewness to the distribution.



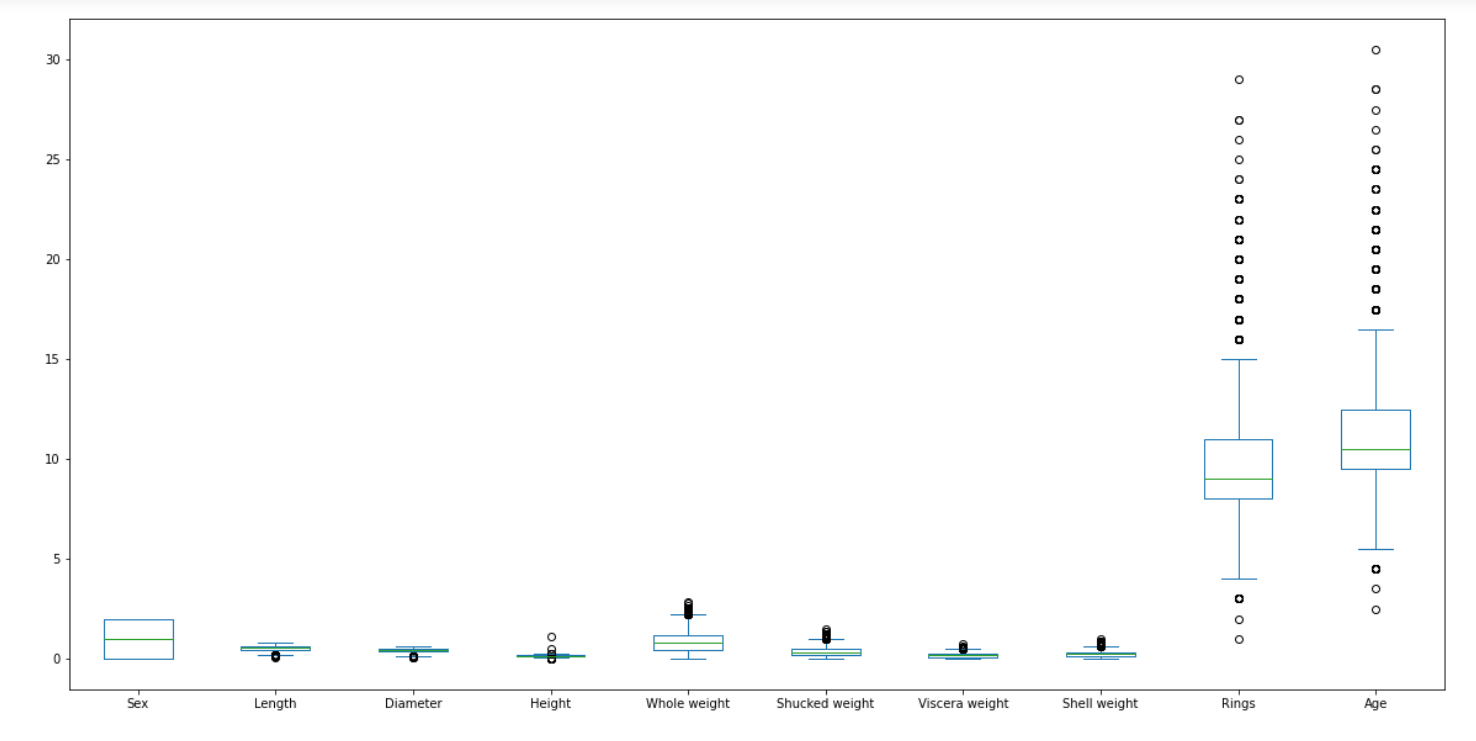
The pair plot plot is shown in the image above. It shows pairwise relationships in a dataset.



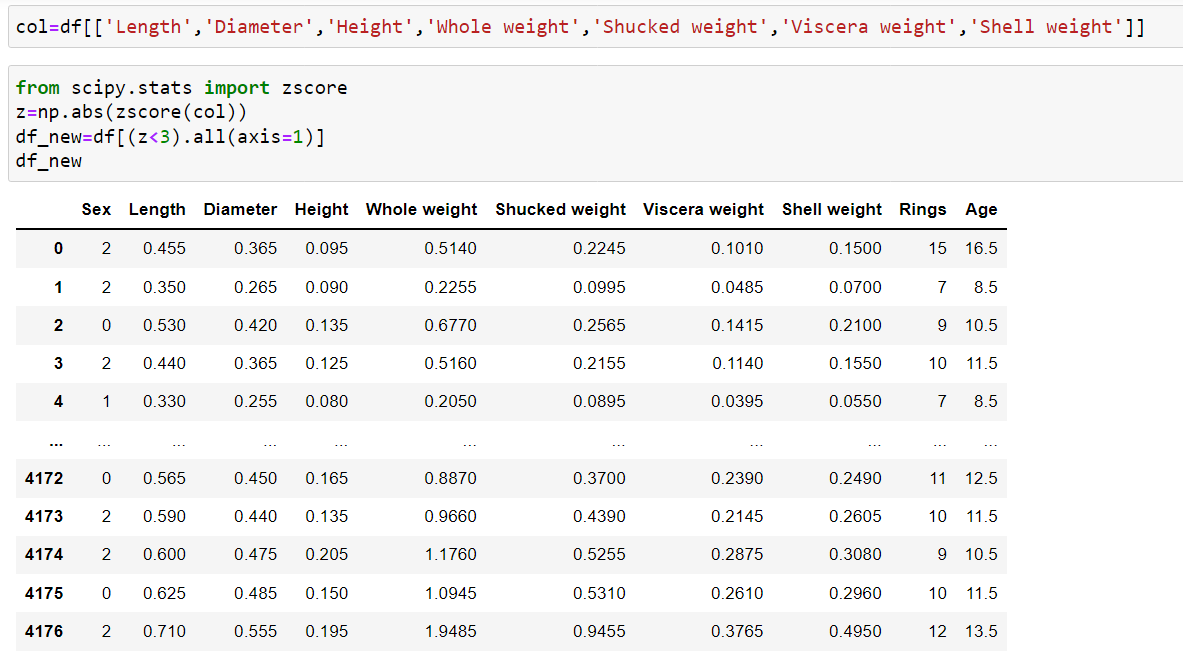


Features are highly correlated with each other and moderately correlated with rings, which is a proxy for the age of the abalone. Since shell weight and Rings have a moderately high correlation, the shell weight might be an important feature for predicting the age of the abalone. In the context of abalone, this could mean that older abalone require a heavier shell, whereas younger abalone may only need a lighter shell. The diameter of an abalone is highly correlated with the length of the abalone.

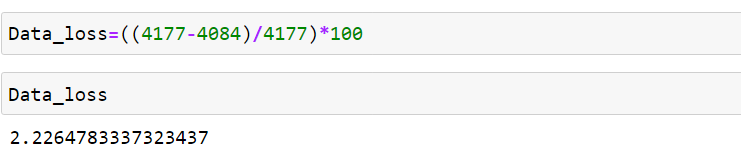
**Identifying the Outliers using Boxplot:**

****

We can clearly see that the outliers are present in all columns except Sex column. Except our Target let’s remove outliers in all other columns.

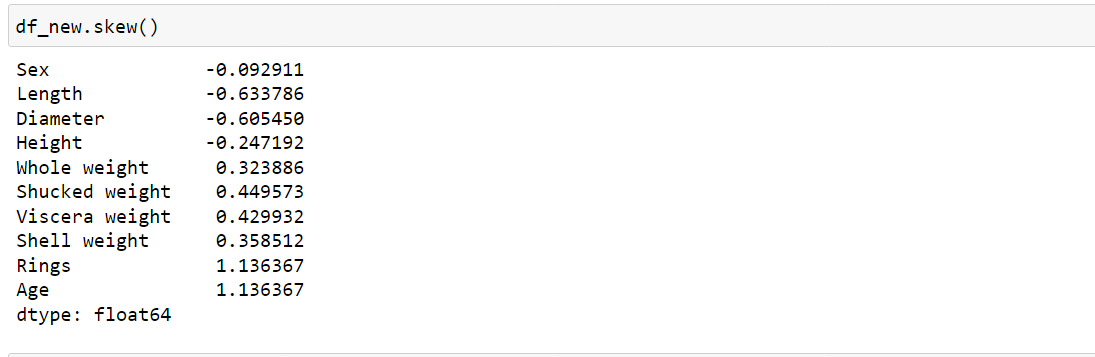


**Percentage of Data loss after removing the Outliers:**

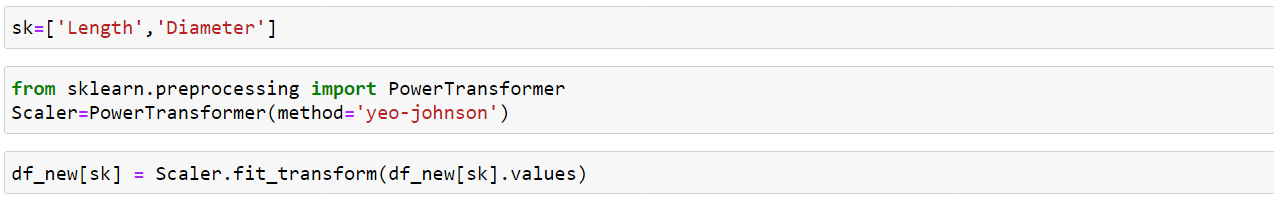


We lost the data around 2% which is acceptable.

**Skewness:**

****

Skewness is identified in ‘Length’ and ‘Diameter’ columns. It can removed by using yeo-johnson transformation.

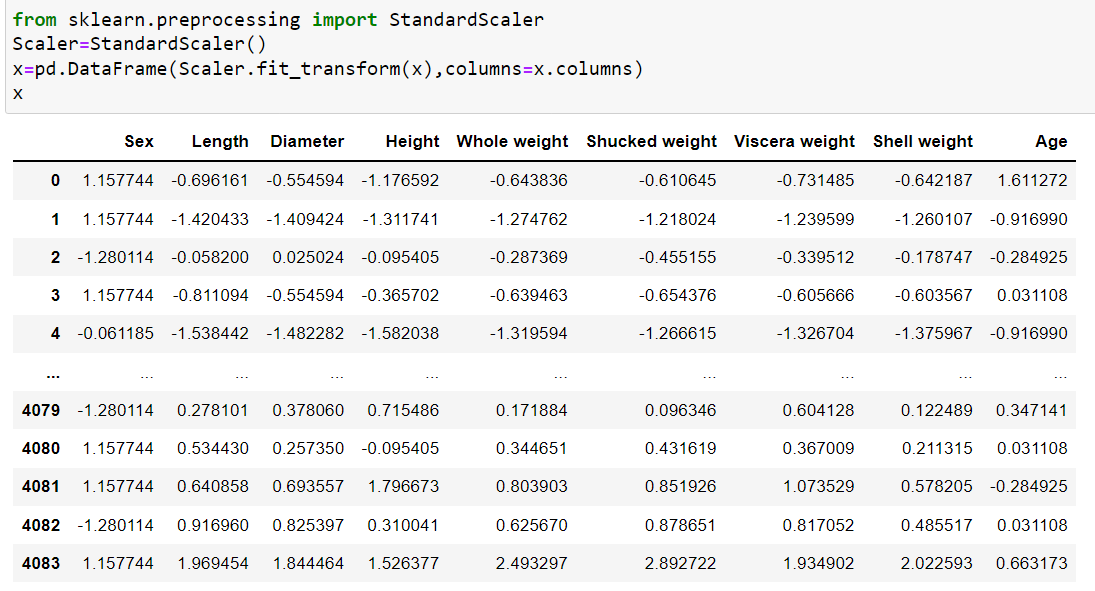


**Pre-Processing Pipeline:**

Seperating Features and Target into x and y



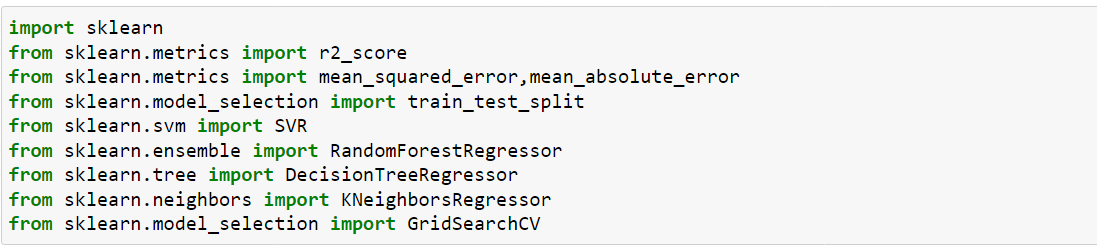
**Scaling:**



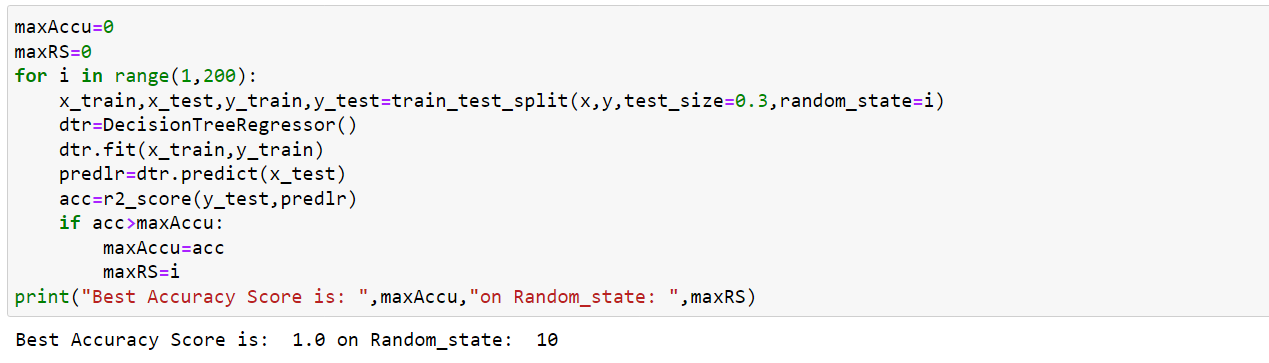
Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance). It standardizes features by subtracting the mean value from the feature and then dividing the result by feature standard deviation.

**5.Building Machine Learning Models:**

Importing Required Libraries for building Machine Learning Models



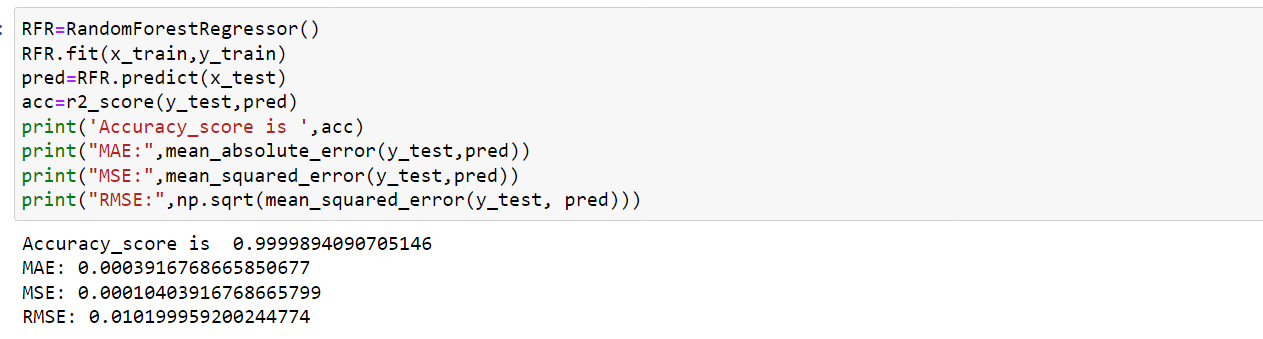
The first step towards applying Decision Tree Regressor to predict the age of the abalone.





**Random Forest Regressor:**

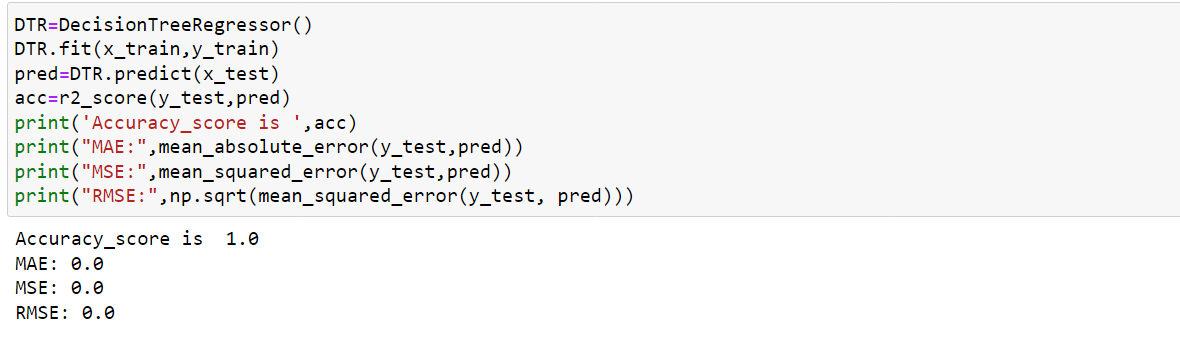
Random forests are ensemble techniques that are widely used for both and classification and regression. They employed multiple decision trees for training and testing.

****

This model has good accuracy of 99% and very low value of MAE,MSE and RMSE.

**Decision Tree Regressor:**

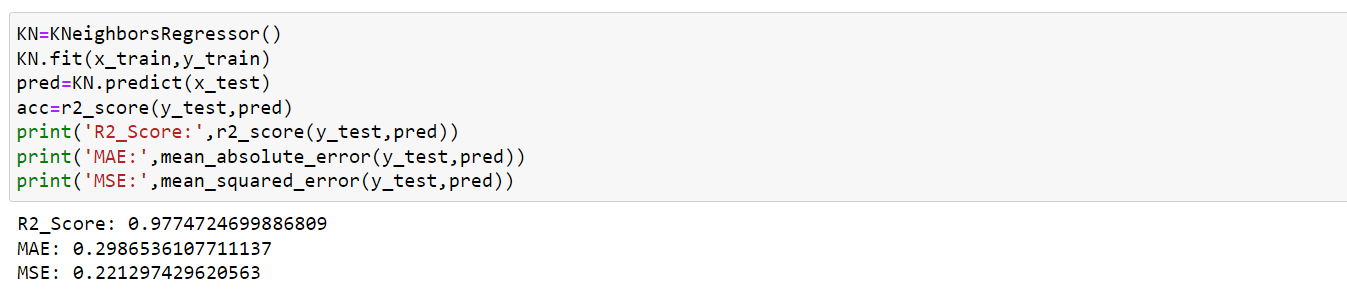
A Decision tree is a simple supervised learning algorithm that can be employed for both classification and regression tasks.It continuously split the data into smaller subset based on some criteria.

****

Decision Tree Regressor is performing well with 100% accuracy.

**K-Neighbors Regressor:**

K-Neighbors is one of the oldest,simple and efficient supervised machine learning algorithms.It has been applied in various applications, including problems of classification and regression.

****

This model’s accuracy is considerably low among other models.

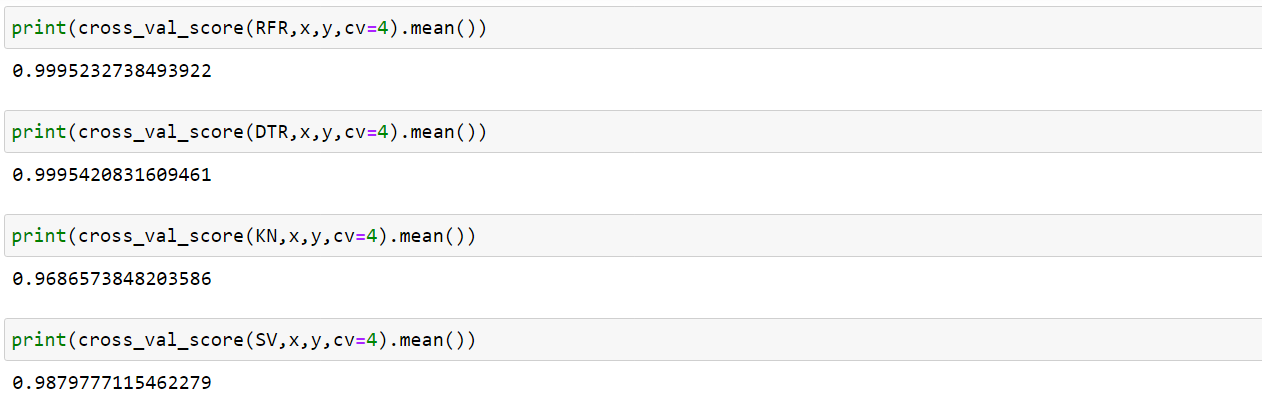
**Support Vector Regressor:**

It constructs a hyperplane in an infinite-dimensional space which is used for classification or regression.

****

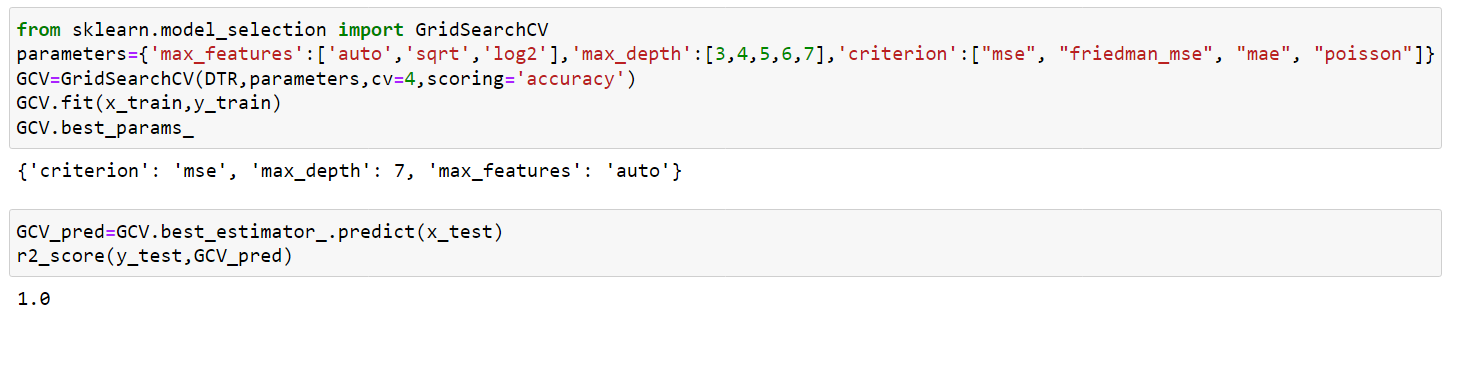
It has 99% accuracy with low mean squared and mean absolute errors.

**Cross Validation:**

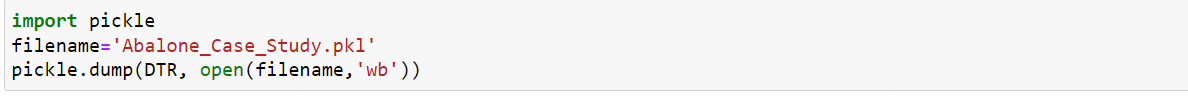
****

By considering accuracy score and cross validation score,we can conclude that DecisionTreeRegressor is the best model.Even though our model is 100% accurate let’s do Hyperparameter Tuning.

**Hyperparameter Tuning:**

****

**Saving the Model:**

****

**Concluding Remarks:**

In this report, five classification algorithms and two regression algorithems were used to predict the target

feature. For each method, cross validation method were used. We use grid search in hyperparameter tuning

and get the best performance of each model based on their accuracy. In the end, we present and discuss the

results in forms of accuracy scores, confusion matrix, classification report (recall, precision, F1-score). Based

on the above analysis, RF model have the best accuracy as well as best recall and f1-score among all

classification models.

In this report, five classification algorithms and two regression algorithems were used to predict the target

feature. For each method, cross validation method were used. We use grid search in hyperparameter tuning

and get the best performance of each model based on their accuracy. In the end, we present and discuss the

results in forms of accuracy scores, confusion matrix, classification report (recall, precision, F1-score). Based

on the above analysis, RF model have the best accuracy as well as best recall and f1-score among all

classification models.

Here, we focused on abalone age prediction using machine learning techniques. In this report, four regression algorithms were used to predict the target feature. For each method, cross validation method is used. We used grid search in hyperparameter tuning. In the end, we present and discuss the results in forms of accuracy scores. Based on the above analysis, DecisionTreeRegressor model have the best accuracy among all regression models i.e 100%.