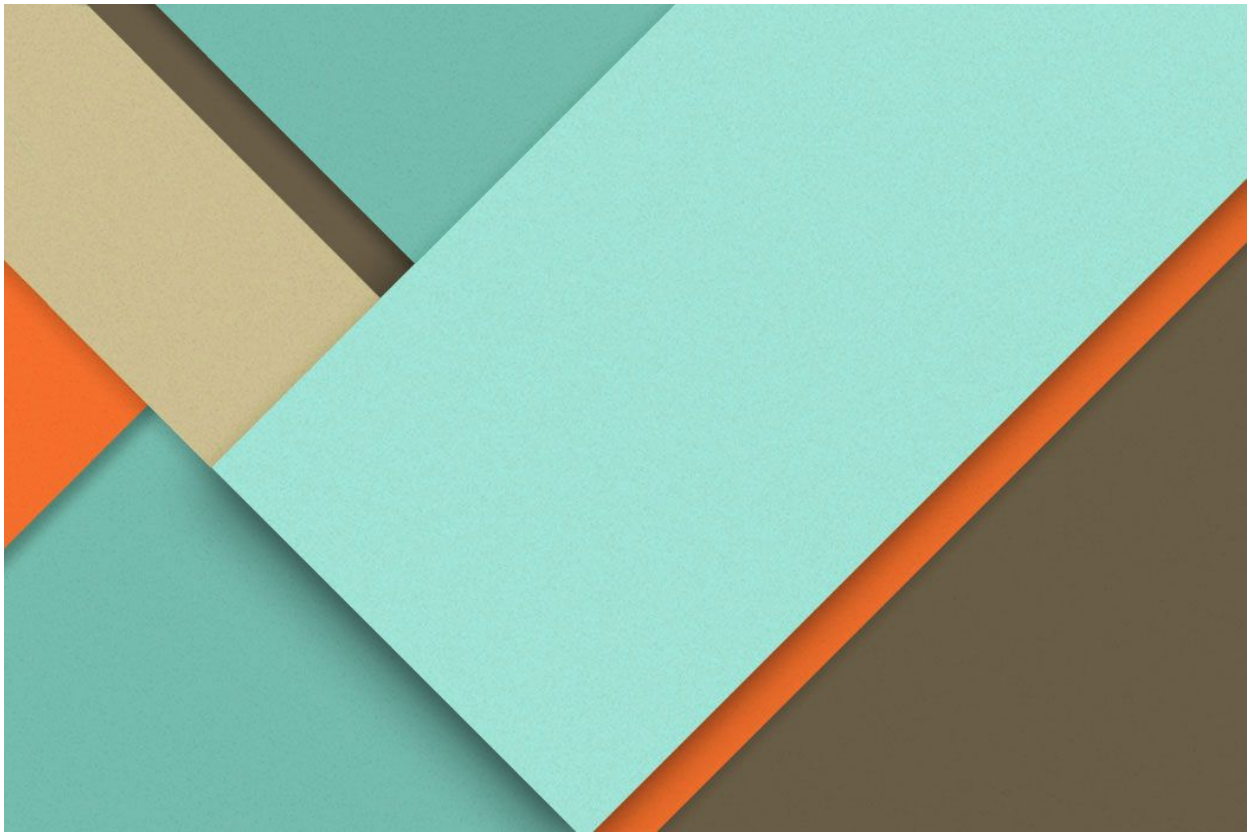


Abalone Age Prediction

21.06.2019

Team Name: A Team Has No Name

Team Members:

Avaneesh Pathak (Team Leader)

Abhinav Gupta

Abhishek Singh

PROJECT REPORT

1.1. Introduction

Python is an interpreted and high-level programming language. Python's design emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aims to help programmers write clear, logical code for small and large-scale projects.

Artificial Intelligence (AI) is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans. The term "artificial intelligence" is used to describe machines/computers that mimic "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".

1.2. Objectives of Research

Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. Other measurements, which are easier to obtain, are used to predict the age**.** We use machine learning algorithms in order to predict the age of a particular abalone by the help of the already provided dataset.

1.3 Problem Statement

To predict the age of an abalone using abalone dataset which comprises of various measurements of the body of the abalone.

2. Review of Literature

It is a multiple linear regression problem so by the use of it we have built a predictive model which can predict the age of abalones by fitting a model into the dataset. But it was then found out that the accuracy of the multiple linear regression model was very low of the order 53.4% so, another approach to obtain a higher accuracy value is to make the use of various

classification techniques in which the output is divided into several categories and then the model is required to predict the correct category.

3. Data collection

Source:

Data comes from an original (non-machine-learning) study:
Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994)
"The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait",
Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

Original Owners of Database:

Marine Resources Division
Marine Research Laboratories - Taroona
Department of Primary Industry and Fisheries, Tasmania
GPO Box 619F, Hobart, Tasmania 7001, Australia
(contact: Warwick Nash +61 02 277277, wnash **'@'** dpi.tas.gov.au)

Donor of Database:

Sam Waugh (Sam.Waugh **'@'** cs.utas.edu.au)
Department of Computer Science, University of Tasmania
GPO Box 252C, Hobart, Tasmania 7001, Australia

Data Set Information:

Predicting the age of abalone from physical measurements. 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.

From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).

4. Methodology

4.1 Exploratory data Analysis

Dataset:-

We obtained the following top 5 observations of our dataset using the '.head()' method

	Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
0	M	70	53	18	45.1	19.9	9.7	14.0	7
1	F	106	84	27	135.4	51.3	28.3	42.0	9
2	M	88	73	25	103.2	43.1	22.8	31.0	10
3	I	66	51	16	41.0	17.9	7.9	11.0	7
4	I	85	60	19	70.3	28.2	15.5	24.0	8

Description:

- From the original data examples with missing values were removed (the majority having the predicted value missing). For the purpose of this analysis, we will scale those variables back to its original form by multiplying by 200.
- Total number of observations in data-set: 4176
- Total number of variables in data-set : 8

Variable List

Name	Data Type	Measurement	Description
Sex	categorical (factor)		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	continuous		+1.5 gives the age in years

Categorical Variables

Sex	Number of observations
M	1527
F	1307
I	1342

Numeric Variables

	Length	Diameter	Height	Whole Weight	Shucked Weight	Viscera Weight	Shell Weight	Rings
Min	15.0	11.00	0.0000	0.4	0.20	0.10	0.30	1
Median	109.0	85.00	28.00	159.9	67.20	34.20	46.80	9
Mean	104.8	81.58	27.91	165.8	71.88	36.12	47.77	9.932
Max	163.0	130.00	226.00	565.1	297.60	152.00	201.00	29
Cor	0.557	0.5750	0.5581	0.5408	0.4212	0.5043	0.6280	1.000

- Looking at the dataset summary, we can see that data is quite evenly distributed between the three factor levels of Male, Female and Infant.
- Also from we see that there are four different measure of weight i.e. Whole_weight, Shell_weight and Shucked_weight. Whole_weight is linear function of other weight predictors with Unknown mass of water/blood lost from shucking process. Also we observed that min value of predictor Height is 0. Practically this is not possible, we will investigate these observations to look closely.

'.describe()' method is used to find the summary of the dataset.

	Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Age
count	4176.000000	4176.000000	4176.000000	4176.000000	4176.000000	4176.000000	4176.000000	4176.000000	4176.000000
mean	0.955699	104.801724	81.578305	27.905412	165.763506	71.879957	36.122534	47.770426	11.432471
std	0.827782	24.020509	19.849906	8.365278	98.084710	44.395943	21.924091	27.842510	3.223601
min	0.000000	15.000000	11.000000	0.000000	0.400000	0.200000	0.100000	0.300000	2.500000
25%	0.000000	90.000000	70.000000	23.000000	88.300000	37.200000	18.675000	26.000000	9.500000
50%	1.000000	109.000000	85.000000	28.000000	159.950000	67.200000	34.200000	46.800000	10.500000
75%	2.000000	123.000000	96.000000	33.000000	230.650000	100.400000	50.600000	65.800000	12.500000
max	2.000000	163.000000	130.000000	226.000000	565.100000	297.600000	152.000000	201.000000	30.500000

count - Count shows the number of non-null observations of a particular feature column.

mean - It calculates the mean value of the predictors by summing them up and dividing them by the total number of observations.

std - It calculates the standard deviation of particular column i.e it returns the value of deviation of each data point of the dataset from its mean value.

min - Shows the minimum value of each feature column.

max - Shows the maximum value of each feature column.

Quartiles - Quartiles are the values that divides a list of numbers into quarters.

There are 3 types of quartiles:-

25% - The first quartile (Q_1) is defined as the middle number between the smallest number and the median of the data set.

50% - The second quartile (Q_2) is the median of the data.

75% - The third quartile (Q_3) is the middle value between the median and the highest value of the data set.

Covariance

Covariance is found using '.cov()' method.

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Age
Sex	0.685277	-0.044614	-0.037646	-0.014471	-0.187240	-0.081018	-0.041256	-0.051342	-0.939017
Length	-0.044614	0.014422	0.011761	0.004157	0.054491	0.023935	0.011887	0.015007	0.215562
Diameter	-0.037646	0.011761	0.009849	0.003461	0.045038	0.019674	0.009787	0.012507	0.183872
Height	-0.014471	0.004157	0.003461	0.001750	0.016803	0.007195	0.003660	0.004759	0.075179
Whole weight	-0.187240	0.054491	0.045038	0.016803	0.240481	0.105518	0.051946	0.065216	0.854409
Shucked weight	-0.081018	0.023935	0.019674	0.007195	0.105518	0.049268	0.022675	0.027271	0.301204
Viscera weight	-0.041256	0.011887	0.009787	0.003660	0.051946	0.022675	0.012015	0.013850	0.178057
Shell weight	-0.051342	0.015007	0.012507	0.004759	0.065216	0.027271	0.013850	0.019377	0.281663
Age	-0.939017	0.215562	0.183872	0.075179	0.854409	0.301204	0.178057	0.281663	10.395266

Covariance indicates how two variables are related. A **positive covariance** means the variables are **directly** related, while a **negative covariance** means the variables are **inversely** related.

Correlation

Correlation is found using '.corr()' method.

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Age
Sex	1.000000	-0.448765	-0.458245	-0.417928	-0.461238	-0.440927	-0.454658	-0.445549	-0.351822
Length	-0.448765	1.000000	0.986812	0.827554	0.925261	0.897914	0.903018	0.897706	0.556720
Diameter	-0.458245	0.986812	1.000000	0.833684	0.925452	0.893162	0.899724	0.905330	0.574660
Height	-0.417928	0.827554	0.833684	1.000000	0.819221	0.774972	0.798319	0.817338	0.557467
Whole weight	-0.461238	0.925261	0.925452	0.819221	1.000000	0.969405	0.966375	0.955355	0.540390
Shucked weight	-0.440927	0.897914	0.893162	0.774972	0.969405	1.000000	0.931961	0.882617	0.420884
Viscera weight	-0.454658	0.903018	0.899724	0.798319	0.966375	0.931961	1.000000	0.907656	0.503819
Shell weight	-0.445549	0.897706	0.905330	0.817338	0.955355	0.882617	0.907656	1.000000	0.627574
Age	-0.351822	0.556720	0.574660	0.557467	0.540390	0.420884	0.503819	0.627574	1.000000

Correlation ranges from -1 to 1.

+1 indicates strong positive correlation

-1 indicates strong negative correlation

0 indicates no correlation

>+0.5 indicates good positive correlation

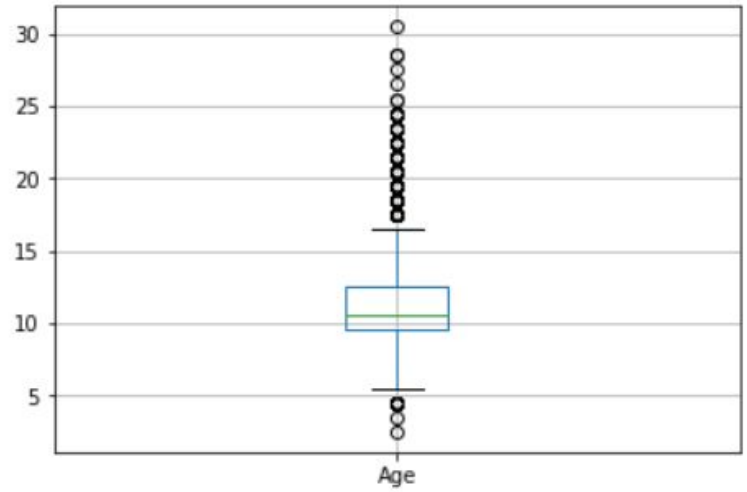
<+0.5 indicates weak positive correlation

>-0.5 indicates weak negative correlation

<-0.5 indicates good negative correlation

Boxplot (Age):-

It is a simple way of representing statistical data on a plot in which a rectangle is drawn to represent the second and third quartiles, usually with a vertical line inside to indicate the median value. The lower and upper quartiles are shown as horizontal lines either side of the rectangle.



Lower Whisker - It indicates the smallest sample value

Upper Whisker - It indicates the largest sample value

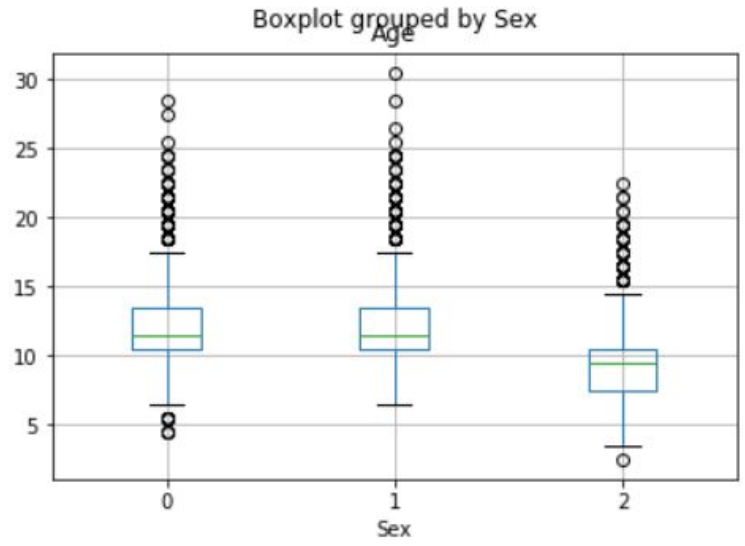
Lower Hinge - It indicates the 25% of the sample value

Upper Hinge - It indicates the 75% of the sample value

Median - It shows the 50% of the sample value

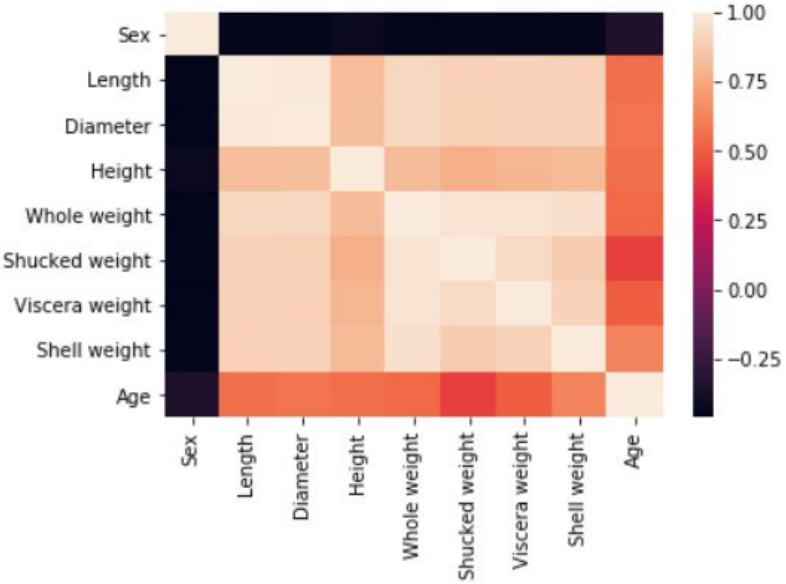
Outliers - These are too far away points, because they lie outside the range in which we expect them.

Below is the boxplot of ‘Age’ grouped by ‘Sex’ which in our dataset is divided into 3 categories i.e. Male - 0, Female - 1 and Infant - 2.



Heatmap :

A heat map (or heatmap) is a graphical representation of data where the individual values contained in a matrix are represented as colors.



Light colors represents strong correlation between 2 significant variables.

Dark colors represents weak correlation between 2 significant variables.

Data Modelling

OLS model for Simple linear regression

OLS Regression Results

Dep. Variable:	Age	**R-squared:**	0.945
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	7.120e+04
Date:	Fri, 21 Jun 2019	**Prob (F-statistic):**	0.00
Time:	19:04:33	**Log-Likelihood:**	-10222.
No. Observations:	4177	**AIC:**	2.045e+04
Df Residuals:	4176	**BIC:**	2.045e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Length	21.4774	0.080	266.824	0.000	21.320	21.635

Omnibus:	952.779	**Durbin-Watson:**	0.828
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2237.504
Skew:	1.270	**Prob(JB):**	0.00
Kurtosis:	5.532	**Cond. No.**	1.00

R^2 - It is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model

Adj. R^2 - It is a modified version of **R-squared** that has been **adjusted** for the number of predictors in the model. It shows the goodness of fit of the model.

Probability(P) - It shows the probability of test significance.

If P < alpha(0.05) - null hypothesis is rejected and alternate hypothesis is accepted

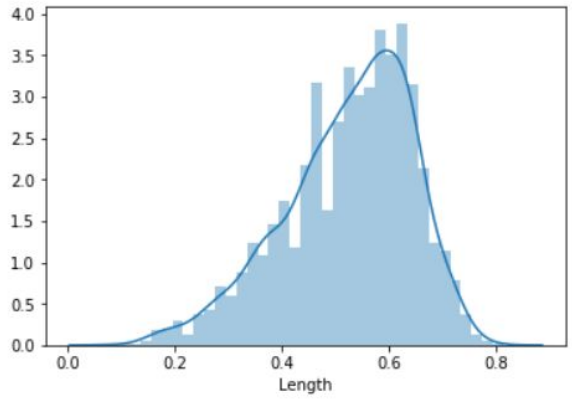
If P > alpha(0.05) - null hypothesis is accepted and alternate hypothesis is rejected

```
accuracy1 = model1.score(x_train1, y_train1)
accuracy1
```

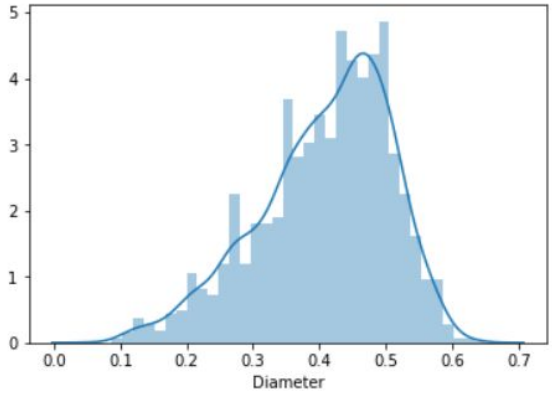
```
0.31412063293521053
```

Distribution plots:

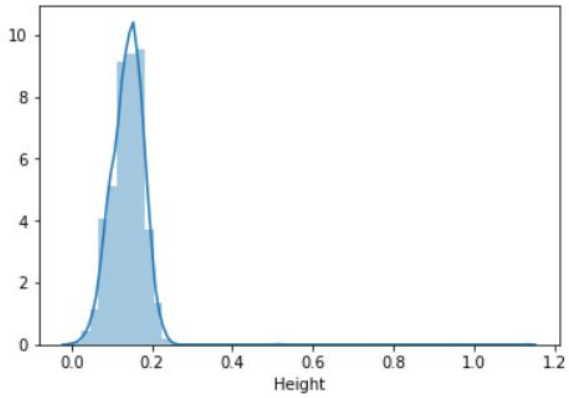
Length:



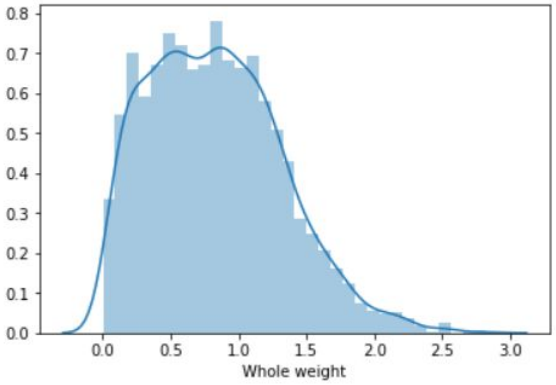
Diameter:



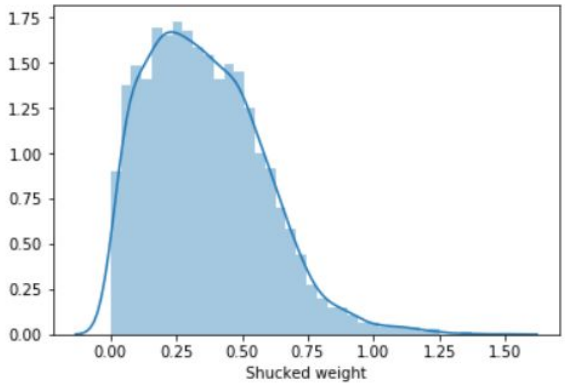
Height:



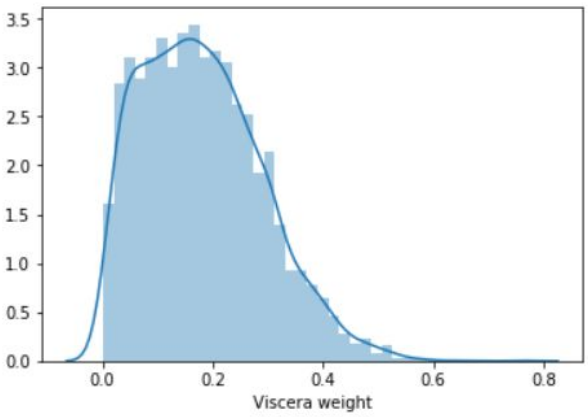
Whole_weight:



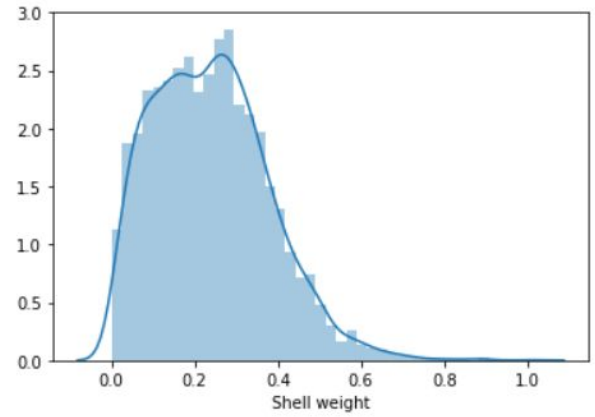
Shucked_weight:



Viscera_weight:



Shell weight:



It is observed in the above plots that all the predictors are normally distributed

Multiple Linear Regression:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

Using multiple linear regression we get the following OLS model

Dep. Variable:	Age	**R-squared:**	0.963
Model:	OLS	**Adj. R-squared:**	0.963
Method:	Least Squares	**F-statistic:**	1.539e+04
Date:	Fri, 21 Jun 2019	**Prob (F-statistic):**	0.00
Time:	19:04:39	**Log-Likelihood:**	-9393.4
No. Observations:	4177	**AIC:**	1.880e+04
Df Residuals:	4170	**BIC:**	1.885e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-0.1111	0.046	-2.418	0.016	-0.201	-0.021
x2	10.1509	1.765	5.752	0.000	6.691	13.611
x3	14.0344	2.307	6.083	0.000	9.511	18.558
x4	16.5017	1.577	10.465	0.000	13.410	19.593
x5	11.4576	0.443	25.845	0.000	10.589	12.327
x6	-24.1625	0.665	-36.328	0.000	-25.466	-22.859
x7	-14.9013	1.271	-11.725	0.000	-17.393	-12.410

Omnibus:	772.857	**Durbin-Watson:**	1.354
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2112.131
Skew:	0.986	**Prob(JB):**	0.00
Kurtosis:	5.872	**Cond. No.**	128.

Result:

Slope(m) and intercept(c) values are found

y = mx + c

```
Slope: [ -0.43566609   0.84819922  11.65450171   8.98841303  13.98231327
 -24.80732049 -13.16051502]
Intercept: 5.089058821903493


The model performance
--------------------------------------
RMSE is 2.3179333349841595
R2 score is 0.493436604396852
```

RMSE - **Root Mean Square Error** (**RMSE**) is the standard deviation of the prediction errors.

R^2 - It is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model

A best model should have low RMSE and high R^2

Accuracy:

```
accuracy = model.score(x_train, y_train)
print("Accuracy : " ,accuracy*100)
```

```
Accuracy :  53.93224669561214
```

Since, using the linear regression models, higher accuracy is not achieved therefore we classify the ‘Age’ into 2 groups i.e. 0 and 1,

0 ranges from 0 - 15 years of age

1 ranges from 15 - 31 years of age

Add a new column ‘Categ_Age’ which shows the age range

	Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Age	Categ_Age
0	0	70	53	18	45.1	19.9	9.7	14.0	8.5	0
1	1	106	84	27	135.4	51.3	28.3	42.0	10.5	0
2	0	88	73	25	103.2	43.1	22.8	31.0	11.5	0
3	2	66	51	16	41.0	17.9	7.9	11.0	8.5	0
4	2	85	60	19	70.3	28.2	15.5	24.0	9.5	0

Now we will check the accuracy through classification algorithms

KNN

An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors. If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

```
array([0.88133971, 0.88793103, 0.88398849])
```

The first value in the array represents the accuracy score for the train data,

Second represents the validation and third represents on test data.

```
Accuracy score: 90.325670
Recall score : 36.923077
ROC score : 67.422151

[[895  19]
 [ 82  48]]
```

Accuracy score - **Accuracy** is one metric for evaluating classification models. **Accuracy** is the fraction of predictions our model got right.

Accuracy=Number of correct predictions/ Total number of predictions

Recall score - **Recall** is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

ROC score - **ROC** curve or Receiver Operating Characteristic is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

Confusion Matrix - It is a specific table layout that allows visualization of the performance of an algorithm.

Model predicts that 895 out of 1044 samples are in the same category so does the scientists,

While 48 out of 1044 samples are predicted not to be in the same category.

19 out of 1044 samples are predicted by the model to be in the same category and scientists says that they are not,

82 out of 1044 samples are predicted by scientists to be in a category but the model predicts it to be in the other.

Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.In **regression** analysis, **logistic regression** is estimating the parameters of a **logistic model**.

```
array([0.88803828, 0.89272031, 0.8906999 ])
```

The first value in the array represents the accuracy score for the train data,

Second represents the validation and third represents on test data.

```
Accuracy score: 90.613027
Recall score : 36.923077
ROC score : 67.586265

[[898  16]
 [ 82  48]]
```

Model predicts that 898 out of 1044 samples are in the same category so does the scientists,

While 48 out of 1044 samples are predicted not to be in the same category.

16 out of 1044 samples are predicted by the model to be in the same category and scientists says that they are not,

82 out of 1044 samples are predicted by scientists to be in a category but the model predicts it to be in the other.

Naive Bayes Classification:

```
Accuracy score: 63.218391
Recall score : 65.384615
ROC score : 64.147450

[[575 339]
 [ 45  85]]
```

Model predicts that 575 out of 1044 samples are in the same category so does the scientists,

While 85 out of 1044 samples are predicted not to be in the same category.

339 out of 1044 samples are predicted by the model to be in the same category and scientists says that they are not,

45 out of 1044 samples are predicted by scientists to be in a category but the model predicts it to be in the other.

Support Vector Classification:

```
Accuracy score: 63.218391
Recall score : 65.384615
ROC score : 64.147450

[[575 339]
 [ 45  85]]
```

Model predicts that 575 out of 1044 samples are in the same category so does the scientists,

While 85 out of 1044 samples are predicted not to be in the same category.

339 out of 1044 samples are predicted by the model to be in the same category and scientists says that they are not,

45 out of 1044 samples are predicted by scientists to be in a category but the model predicts it to be in other.

Decision Tree Classifier:

```
array([0.8430622 , 0.85057471, 0.84755513])
```

The first value in the array represents the accuracy score for the train data,

Second represents the validation and third represents on test data.

```
Accuracy score: 84.865900
Recall score : 43.846154
ROC score : 67.273186

[[829  85]
 [ 73  57]]
```

Model predicts that 829 out of 1044 samples are in the same category so does the scientists,

While 57 out of 1044 samples are predicted not to be in the same category.

85 out of 1044 samples are predicted by the model to be in the same category and scientists says that they are not,

73 out of 1044 samples are predicted by scientists to be in a category but the model predicts it to be in other.

Random Forest Classifier:

```
array([0.88516746, 0.88505747, 0.88590604])
```

The first value in the array represents the accuracy score for the train data,

Second represents the validation and third represents on test data.

```
Accuracy score: 90.613027
Recall score : 36.923077
ROC score : 67.586265

[[898  16]
 [ 82  48]]
```

Model predicts that 829 out of 1044 samples are in the same category so does the scientists,

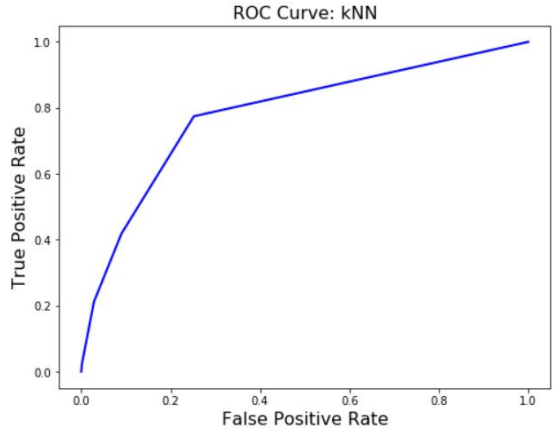
While 57 out of 1044 samples are predicted not to be in the same category.

85 out of 1044 samples are predicted by the model to be in the same category and scientists says that they are not,

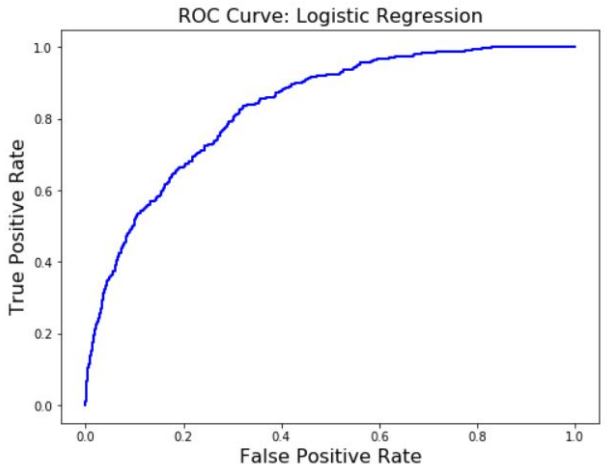
73 out of 1044 samples are predicted by scientists to be in a category but the model predicts it to be in other.

ROC Curves:

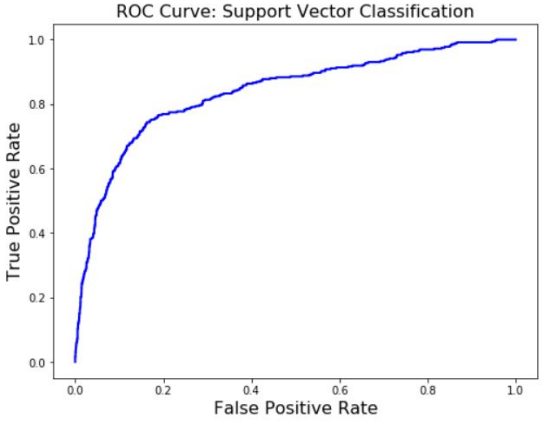
AUC Score (kNN): 0.78



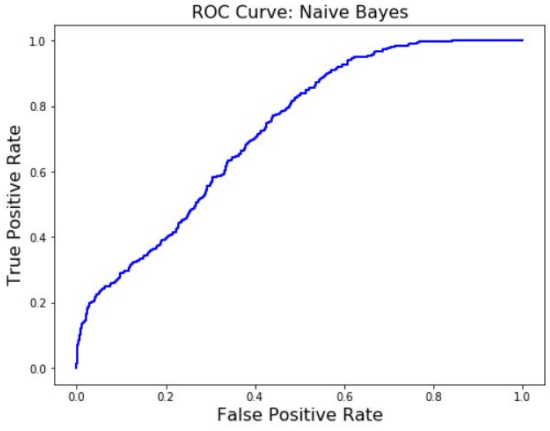
AUC Score (Logistic Regression): 0.83



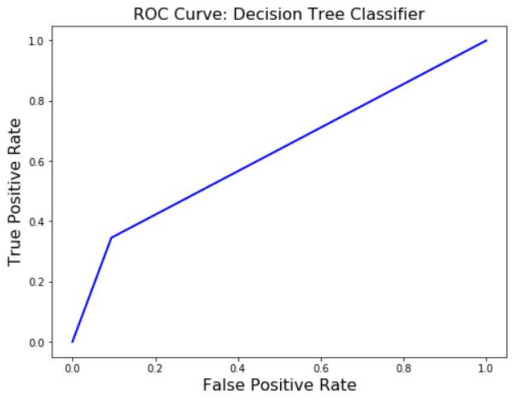
AUC Score (Support Vector Classification): 0.84



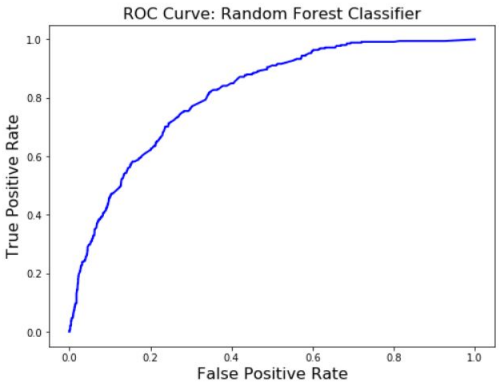
AUC Score (Naive Bayes): 0.73



AUC Score (Decision Tree Classifier): 0.63

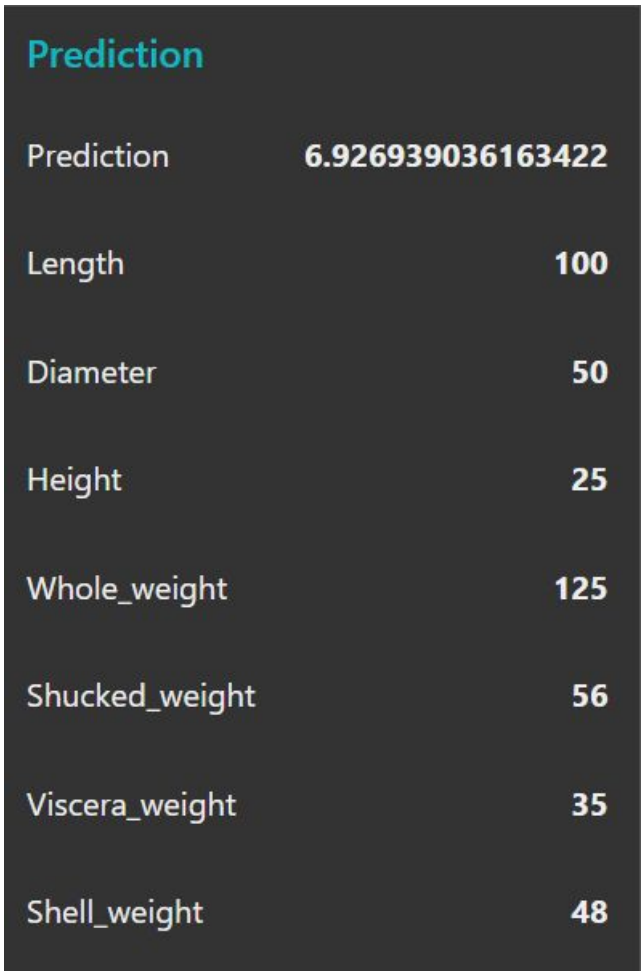


AUC Score (Random Forest Classifier): 0.81



UI Output:

We have used Watson Studio and Node-red to obtain the below prediction of our developed model.



Conclusion:

Since the output in the dataset is numerical, we use Multiple Linear Regression for our dataset to get higher accuracy upto 53%. Since it is not high enough, we classify the output variable into two categories and use classification algorithms to check for higher accuracy.

Performing KNN, we get accuracy up to 90%. Using Logistic Regression we get accuracy up to 91%.

Using Naive Bayes and SVC we get accuracy approx. 63%.

Using Decision Tree we get accuracy up to 84%.

Performing Random Forest Classifier we get accuracy up to 91%.

Therefore, we conclude that the highest accuracy can be obtained using Logistic Regression and Random Forest Classifier if we divide the output into categories.

But for Abalone Age Prediction we use Multiple Linear Regression for the highest accuracy since it contains numerical output.

Bibliography:

1. Abalone. http://en.wikipedia.org/wiki/Abalone

2. UCI Machine Learning Repository: Abalone Data Set. http://archive.ics.uci.edu/ml/datasets/Abalone

3. Abalone Data Set README file. http://archive.ics.uci.edu/ml/machine-learningdatabases/abalone/abalone.names

4. Extending and benchmarking Cascade-Correlation. Sam Waugh (1995), PhD thesis, Computer Science Department, University of Tasmania.

5. A Quantitative Comparison of Dystal and Backpropagation. David Clark, Zoltan Schreter, Anthony Adams. Australian Conference on Neural Networks (ACNN'96)

6. IBM Watson Documentation

https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/spss-viz-linear.html