Introduction:

In this Mini project for the given dataset (Abalone dataset) we have to predict the age from physical measurements, building feature set, creating train test split, evaluate the accuracy of decision tree with max_depth=3, and have to find the best value of maximum depth for better performance.

Abalone is a shellfish considered a delicacy in many parts of the world. An excellent source of iron and pantothenic acid, and a nutritious food resource and farming in Australia, America and East Asia. 100 grams of abalone yields more than 20% recommended daily intake of these nutrients. The economic value of abalone is positively correlated with its age. Therefore, to detect the age of abalone accurately is important for both farmers and customers to determine its price. However, the current technology to decide the age is quite costly and inefficient. Farmers usually cut the shells and count the rings through microscopes to estimate the abalones age. Telling the age of abalone is therefore difficult mainly because their size depends not only on their age, but on the availability of food as well. Moreover, abalone sometimes form the so-called 'stunted' populations which have their growth characteristics very different from other abalone populations This complex method increases the cost and limits its popularity. Our goal in this report is to find out the best indicators to forecast the rings, then the age of abalones.

Dataset description:

Dataset characteristics:	Multivariable	Number of Instances:	4177
Attribute Characteristics:	Categorical, Integer, Real	Number of attributes:	8
Associated Tasks:	Classification	Area:	Life

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). 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 dataset: **4176**
Total number of variables in dataset: **8**
Metadata and 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 as a continuous value.

Name	Data Type	Meas.	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

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.

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

Steps taken in Projects:

Importing the dataset using pandas library

```
df = pd.read_csv ('tarun.csv', names= column_names)
```

1. Building feature set:

   In this part we have to build the feature sets for our projects such as what are the attributes we have to describe them and also describe the target.

   ```
   X = np.array(df.loc[:,['Sex', 'len', 'dia', 'height', 'whole wt', 'shucked wt',
   'Viscera wt','shell wt']])
   y = np.array (df.loc [:,['Rings']])
   ```

2. Creating train test split:

   Simply using the sklearn library we import train_test_split function and use it to train and test our data.
   With test size of 80 for train and 20 for test.

   ```
   X_train, X_test, y_train, y_test= train_test_split (df, y, test_size=0.2)
   ```

3. Evaluate the accuracy of decision tree with max_depth 3:

   Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

   Using sklearn library we import :

   ```
   from sklearn.tree import DecisionTreeClassifier
   ```

Than we use the criterion gini and given max depth 3 for the decision tree to calculate the accuracy at depth 3.

We use sklearn library for calculating the accuracy :

```
from sklearn.metrics import accuracy_score
```

```
print('Accuracy =%2f'%accuracy_score(y_test,y_pred))
```

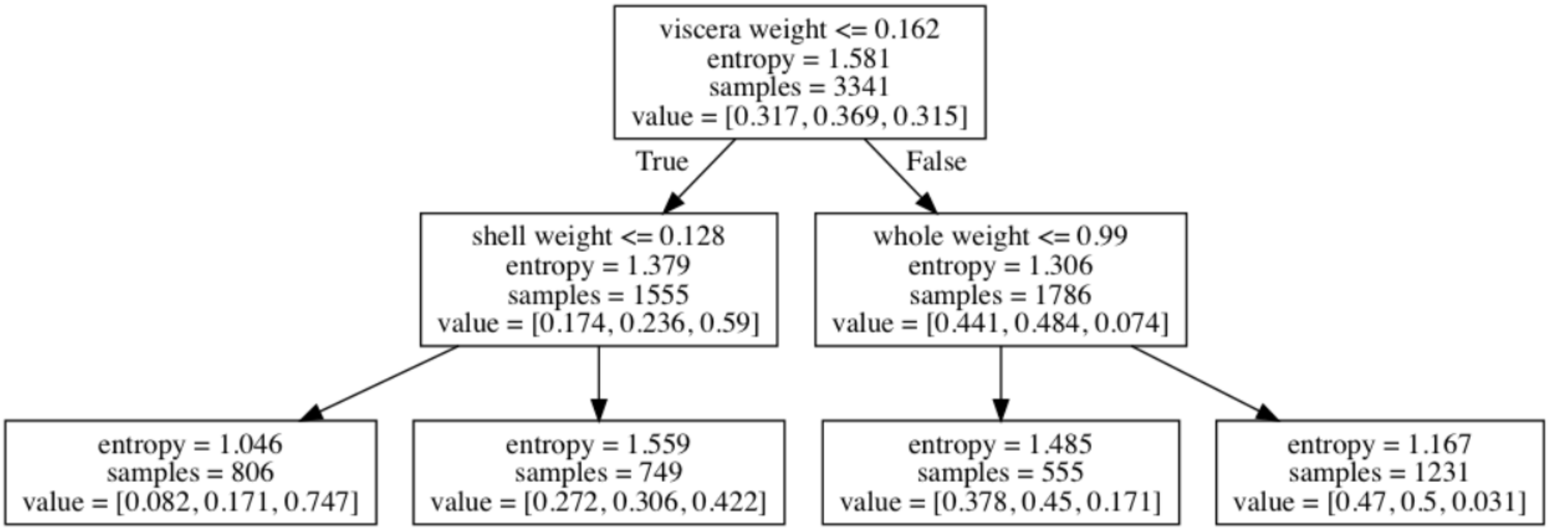
Here we calculate the accuracy by matching our prediction with the testing class and we use gini criterion for more accuracy.

4. Find best value of maximum depth for better performance:

   Here we calculate the maximum depth value for which our accuracy is maximum.
   In this case the accuracy is maximum at depth 25.

Decision tree:

Let's use the abalone data set as an example. We will try to predict the number of rings based on variables such as shell weight, length, diameter, etc. We fit a shallow decision tree for illustrative purposes. We achieve this by limiting the maximum depth of the tree to 3 levels.



To predict the number of rings for an abalone, a decision tree will traverse down the tree until it reaches a leaf. Each step splits the current subset into two. For a specific split, the contribution of the variable that determined the split is defined as the change in mean number of rings.

We can view the graph by using sklearn library known as graphviz:

```
from sklearn.tree import export_graphviz
```

```
dot_data= export_graphviz(tree,out_file='tree.dot',feature_names=['sex', 'length',
'diameter','height', 'whole weight', 'shucked weight', 'viscera weight’, ‘shell
weight', 'rings'])
```