**An Analysis of the Abalone DataSet**

**Background**:- .  Abalone is a common name for any of a group of small to very large sea snails and molluscs. The age of abalone is determined by cutting the shell through the cone and counting the number of rings through a microscope.

**Dataset description**

1. We are using popular Abalone dataset and **the number of rings is the value to Predict**.

The dataset contains **4177 observations and 9 variables** which are as follows:-

* Sex - M, F, and I (infant)
* Length - longest shell measurement (mm)
* Diameter - measured perpendicular to length (mm)
* Height - with meat in shell (mm)
* Whole - whole abalone weight (g)
* Shucked - weight of meat only (g)
* Viscera - gut weight, after bleeding (g)
* Shell - weight after being dried (g)
* Rings - +1.5 gives the age in years

**Class Distribution:**

* Class Examples
* ----- --------
* 1 1
* 2 1
* 3 15
* 4 57
* 5 115
* 6 259
* 7 391
* 8 568
* 9 689
* 10 634
* 11 487
* 12 267
* 13 203
* 14 126
* 15 103
* 16 67
* 17 58
* 18 42
* 19 32
* 20 26
* 21 14
* 22 6
* 23 9
* 24 2
* 25 1
* 26 1
* 27 2
* 29 1
* ----- ----
* Total 4177

## Goal

Formulate and justify a best model for predicting Rings . You have access to all other variables as predictors.

Pre-processing

1. The variable related to “sex” is in char,so we have to change in integer by creating the **numpy as arrray** ( M=0,F=1,I=2).
2. Now,the given dataset have more classes e.g.29 and the examples of few classes is less e.g.1,2,6,9;(**Imbalance State**)

Therefore ,it can’t be trained and tested with less no. of data. So,the dataset is over-sampled(**Over-Sampling Technique**) to create the equal no of examples of each classes.

1. Further, the **feature\_importances** of each Variables(or,column) is calculated by **RandomForestClassifier** to know the importance that is required or not in the final result.
2. Again,the feature\_importances of variable named “sex” can be neglected.(e.g.0.04). so this variable is removed from dataset.
3. Now,again dataset is over-sampled.
4. Hence,the dataset is trained and tested which as follows:

The data to be tested**:-1253 (30%)**

The data to be trained**:-(4177-1253)=2924**

**a-> One hyperplane fitting the target value**

>>output=10.75%

**b->One vs All(for each classes)**

>>output=29.88%

1. Finally,**the confusion\_matrix** is created and by using **RandomForestClassifier**,the accuracy is calculated.

Final result

1. We have got an **accuracy 72%** (approx.) finally.