# Linear Regression for Classification

### Created by:

Arijit Ganguly

1001871460

# Problem Statement

The use of iris data set for the prediction of species is a classic example for classification problem. This classification problem needs to be solved by the Linear Regression which is a supervised learning problem. A linear regression algorithm needs to be developed that can predict the species of input provided to the algorithm with almost certainty (close to 100% accuracy).

# Data

The Iris Data set contains 5 columns – Sepal Length, Sepal Width, Petal Length, Petal Width and Species.

We use the first 4 columns as our features – Sepal length, Sepal width, Petal length and Petal width. All the columns for features are float values. The values in these columns make up the A matrix.

Last column is going to be used as the labels. It states the species of the plant which have the corresponding features. The data set contains 3 species – Iris setosa, Iris virginica and Iris versicolor.

Data Distribution based on species:

|  |  |
| --- | --- |
| **Species** | **Number of Records** |
| Iris Setosa | 50 |
| Iris Virginica | 50 |
| Iris Versicolor | 50 |

Information on the data received from function pandas.info()

|  |  |  |  |
| --- | --- | --- | --- |
| **Index number** | **Column** | **Non-Null Count** | **Data Type** |
| 0 | Sepal Length | 150 | Float64 |
| 1 | Sepal Width | 150 | Float64 |
| 2 | Petal Length | 150 | Float64 |
| 3 | Petal Width | 150 | Float64 |
| 4 | Species | 150 | category |
| 5 | Species\_cat | 150 | Int8 |

Data visualization:

# Strategy

The strategy for implementing a solution is to use k-folds cross validation to create k number of bins in the data and then train and test data on each of these bins which will give **Beta** for each the bins. To reduce the overfitting, we will find the mean of each Beta value generates for each feature as **Beta Mean**.

## Method Used:

### Fit Method

The fit method is the method use for obtaining the **Beta** values from the matrix **A**, which is the matrix created from the records in the features for each bin and **Y**, which is the label encoded values for the species.

Using the formula below, we get **Beta** values:

**B = (ATA)-1ATY**

### Predict Method

The predict method uses the B values and the input matrix **A** values to be used to get predicted **Y** values.

Using the formula below, we get **Y** values:

**Y = AB**

## Process