# Tools for Data Science

Task 2 Classification

Dev Moxaj Desai
*Seasons of Code*
*Web and Coding CLub, IIT B*

## Question 1

Classification is used to solve the task of predicting a label, e.g., is a given email spam or not spam, will it rain on a given day or not, et al.
For this task, use the **iris** dataset (code: `datasets.load_iris()`) and to keep the running times short use only the first **three** features.
We will be looking at primarily two algorithms to do the task.

1. Using *logistic regression* fit the model. Report the accuracy of the model.
2. Remember Bayes theorem from JEE days? Guess what, we can use that for classification. Use the *Gaussian Naive Bayes* algorithm and report the accuracy of the algorithm.
   [Note: Ignore the Gaussian part for now]
3. We are using three features to fit the model. What happens to the accuracy if we increase the number of features? In general, is it always better or worse to have more features?
   [Note: look up terms like overfitting and underfitting]

## Basic Resoruces

1. Naive Bayes
2. Logistic Regression
   (a) Algorithm: 6.1 to 6.5
   (b) Implementation: a simple implementation