**Problem statement**

The data set we’ll be using is the Iris Flower Dataset which was first introduced in 1936 by the famous statistician Ronald Fisher and consists of 50 observations from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals. The aim of this assigment is to try and distinguish between the three species of Irisusing the KNN and Naive Bayes algorithms.

The second dataset that has been provided describes the medical records for Pima Indians and whether or not each patient will have an onset of diabetes within five years. For this purpose various features like the number of prenancies, plasma glucose concentration and other symptomatic variables were taken into account. Here you are expected to fit an SVM model in order correctly identify the onset of diabetes in a patient.

The data description for both the datasets are as follows.

**Iris:**

**sepal length** - Length of the sepal (a part of the flower) in cm  
**sepal width** - Length of the sepal (a part of the flower) in cm  
**petal length** - Length of the sepal (a part of the flower) in cm  
**petal width** - width of the sepal (a part of the flower) in cm  
**class** - The species of the Flower

**Pima Indians:**

**preg =** Number of times pregnant

**plas =** Plasma glucose concentration a 2 hours in an oral glucose tolerance test

**pres =** Diastolic blood pressure (mm Hg)

**skin =** Triceps skin fold thickness (mm)

**test =** 2-Hour serum insulin (mu U/ml)

**mass =** Body mass index (weight in kg/(height in m)^2)

**pedi =** Diabetes pedigree function

**age =** Age (years)

**class =** Class variable (1:tested positive for diabetes, 0: tested negative for diabetes)

Do the following:

Iris Dataset:

1. Perform EDA and check for missing values and outliers.
2. Change all the classes to numericals.
3. Plot the scatter matrix for all the variables.
4. Split the dataset into training and test sets.
5. Build the model and train and test on training and test sets respectively using scikit-learn. Print the Accuracy of the model with different values of k=3,5,9.
6. Validate the results and plot the misclassification error.

Pima Indians:

1. Read the input file 'Diabetes.csv' using Pandas and check it's dimensions.
2. Check distribution of dependent variable 'class' and plot it.
3. Do correlation analysis and bivariate viualization with insights.
4. Do train and test split with stratify sampling on outcome variable to maintain the distribution of dependent variable.
5. Scale the data points using MinMaxScaler and fit a SVM Model on scaled data and give your observation.
6. Try improving the model accuracy using C=1000.