Assignment 2

Part A

Data Set: For Part A, we will be using the famous [Iris flower data set](http://en.wikipedia.org/wiki/Iris_flower_data_set). The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Fisher in the 1936 as an example of discriminant analysis. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor), so 150 total samples. Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

The iris dataset contains measurements for 150 iris flowers from three different species.

The three classes in the Iris dataset:

Iris-setosa (n=50)

Iris-versicolor (n=50)

Iris-virginica (n=50)

The four features of the Iris dataset:

sepal length in cm

sepal width in cm

petal length in cm

petal width in cm

Implement the following steps using SVM machine learning model:

1. Display the image of each type of flower with dimension 300 *300.
2. Get the data. Use sns.load_dataset method for this task.
3. Do some Exploratory analysis of the data set and answer the following questions:
   a. Which flower species seems to be the most separable?** (Hint: Create a Pairplot)
   b. Analyse sepal_length versus sepal width for setosa species of flower. (Hint:Use Kde plot). What is your analysis?
4. Split your data into a training set and a testing set.
5. Train the Model.

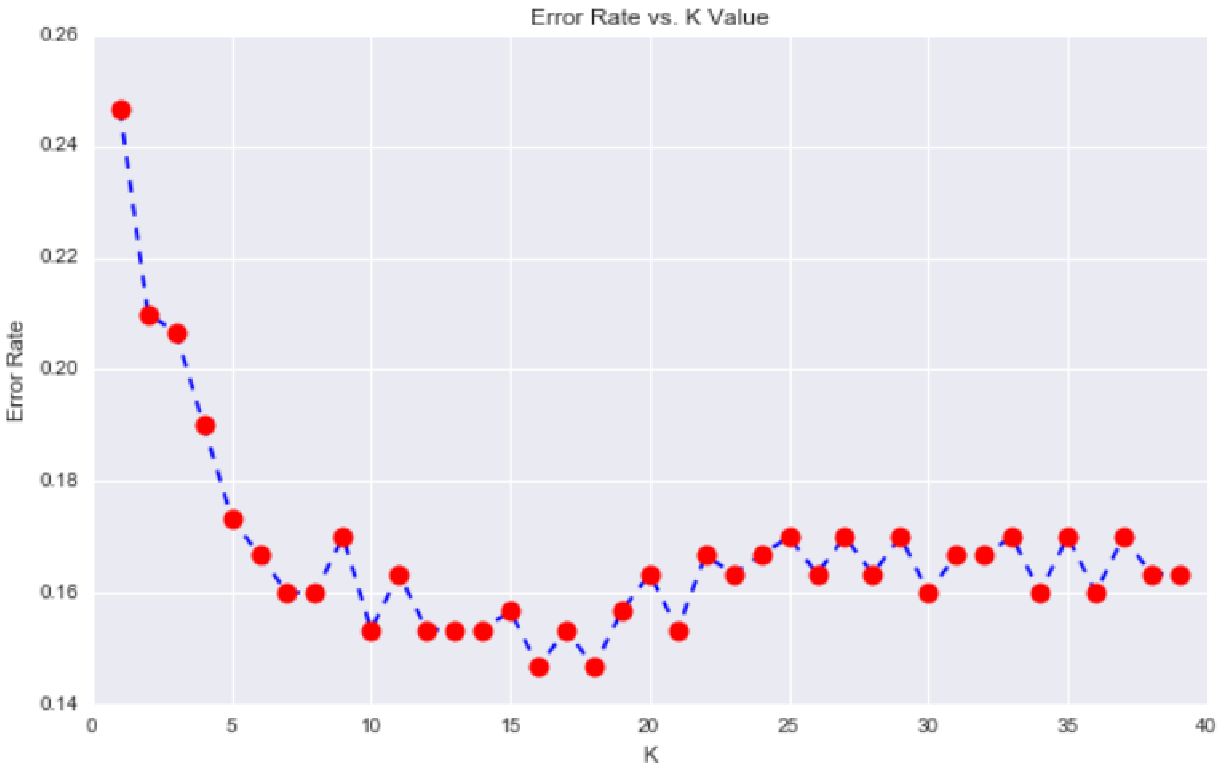
6. Model Evaluation: Now get predictions from the model and create a confusion matrix and a classification report.
7. Now it’s time to tune the parameters of the model to get better results. Use GridSearchCV for tuning.
8. Now take that grid model and create some predictions using the test set and create classification reports and confusion matrices for them. Were you able to improve?**

Part B

In this section, you will be implementing **K Nearest Neighbors for the assigned task.**

Data Set: Use KNN_Project_Data for this problem.

1. **Pre-requisite steps**
   a. Import the libraries
   b. Get the data
   c. Do some analysis like analyze what’s the dimension of the data, mean, sum , is this null etc. (You can do any number of steps for this task).
2. Do some Exploratory analysis of the data set and try to find out of the different columns are related to each other or not?
3. As the data is of different scales. So apply standardization over the data set. Crosscheck after this step of the data is actually standardized or not.
4. Use the predict method to predict values using your KNN model and X_test.Also create confusion matrix and a classification report.
5. At step 4, we may not be satisfied with the results. Let's go ahead and use the elbow method to pick a good K Value!
   a. Create a for loop that trains various KNN models with different k values, then keep track of the error_rate for each of these models with a list.
   b. Now create a plot (a sample graph is given below) using the information from your for loop.



6. Retrain your model with the best K value (up to you to decide what you want) and re-do the classification report and the confusion matrix.
7. Summarize your findings.