| **LP-III Machine Learning (2024-25)** | |
|---|---|
| **Assignment 2:** Classify the email using the binary classification method. Email Spam detection has two states: a) Normal State – Not Spam, b) Abnormal State – Spam. Use K-Nearest Neighbors and Support Vector Machine for classification. Analyse their performance. | |
| **Student Name:** | **Roll No. :** |
| **Batch:** | **Division :** |

**Assignment 2:**
Classify the email using the binary classification method. Email Spam detection has two states: a) Normal State – Not Spam, b) Abnormal State – Spam. Use K-Nearest Neighbors and Support Vector Machine for classification. Analyze their performance.

**Title:** Implement KNN and SVM classification algorithm to predict Normal and Abnormal emails using dataset available at https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv.

**Aim:** Predict and Analyse Results of KNN and SVM algorithm for Classification.

**Prerequisites:** Binary Classification, KNN, SVM.

**Theory:**

**K-nearest Neighbours (KNN) Algorithm:**

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

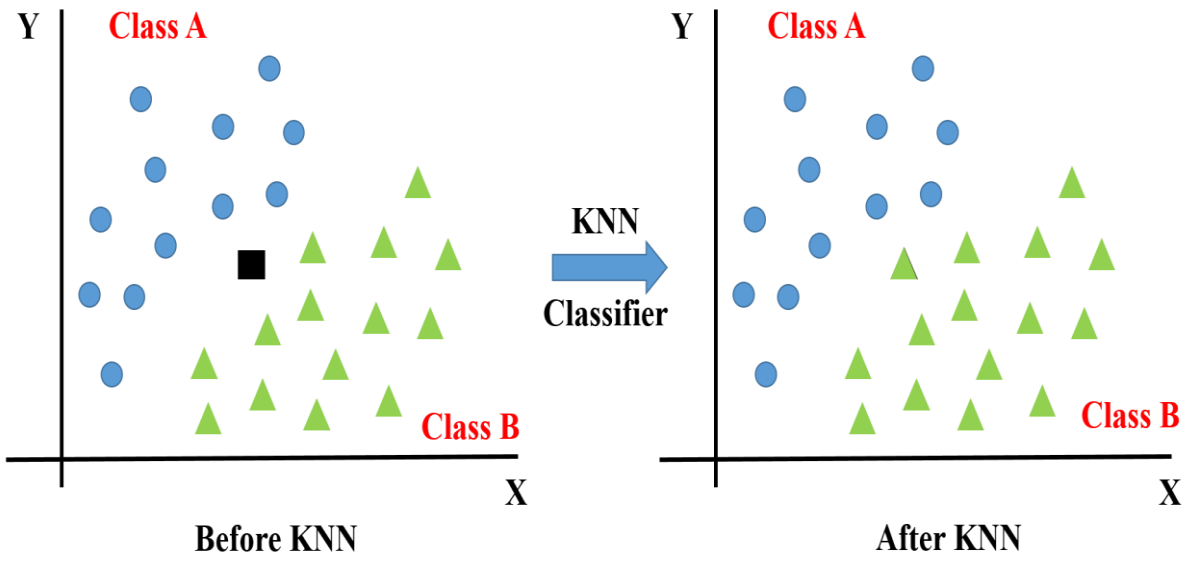


**Figure 1: Visualization of KNN.**

The impact of selecting a smaller or larger K value on the model

**Larger K value:** The case of underfitting occurs when the value of k is increased. In this case, the model would be unable to correctly learn on the training data.

**Smaller k value:** The condition of overfitting occurs when the value of k is smaller. The model will capture all of the training data, including noise. The model will perform poorly for the test data in this scenario.

When the problem statement is of 'classification' type, KNN tends to use the concept of "Majority Voting". Within the given range of K values, the class with the most votes is chosen. When the problem statement is of 'regression' type, KNN employs a mean/average method for predicting the value of new data. Based on the value of K, it would consider all of the nearest neighbours. The algorithm attempts to calculate the mean for all the nearest neighbours' values until it has identified all the nearest neighbours within a certain range of the K value.

**Algorithm:**

**The K-NN working can be explained on the basis of the below algorithm:**

**Step-1: Select the number K of the neighbors**

**Step-2: Calculate the Euclidean distance of K number of neighbors**

**Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.**

**Step-4: Among these k neighbors, count the number of the data points in each category.**

**Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.**

**Step-6: Our model is ready.**

**Support Vector Machine (SVM) Algorithm:**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the **best line** or **decision boundary** that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a **hyperplane**. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as **support vectors**, and hence algorithm is termed as **Support Vector Machine**.

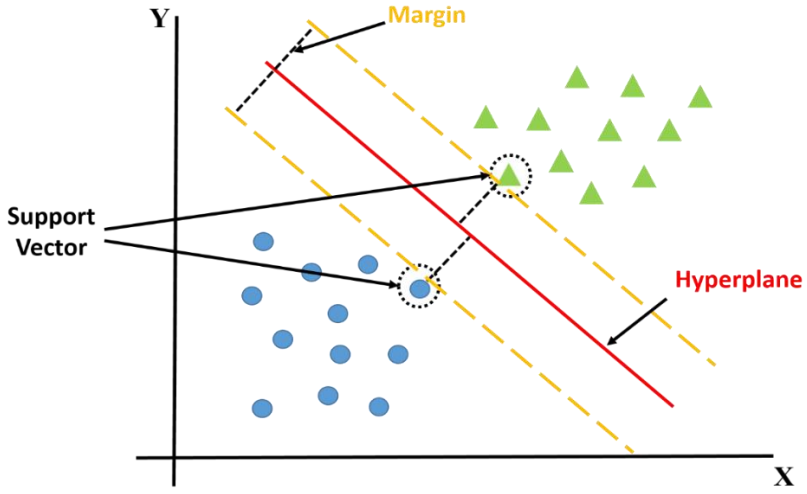


**Figure 2: Terminologies in SVM.**

There are two types of Support Vector Machines:

1. **Linear SVM or Simple SVM:** Linear SVM is used for linearly separable data. If a dataset can be classified into two classes with a single straight line, then that data is considered to be linearly separable data, and the classifier is referred to as the linear SVM classifier. It is typically used for linear regression and classification problems.

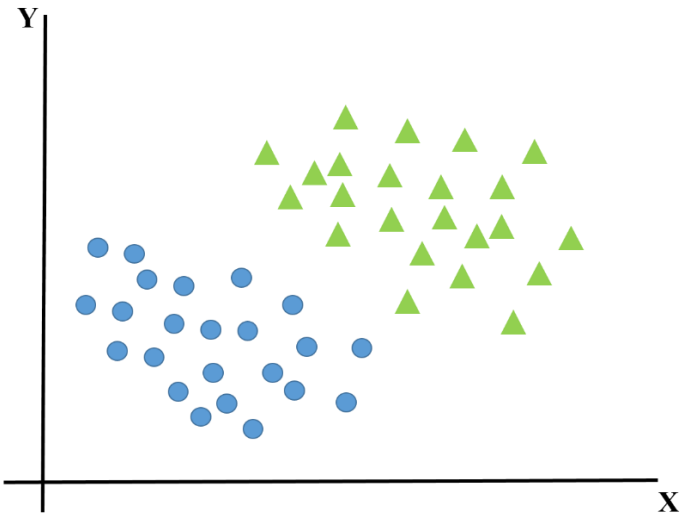


**Figure 3: Linearly Separable Dataset.**

2. **Nonlinear SVM or Kernel SVM:** Nonlinear SVM is used for nonlinearly separated data, i.e., a dataset that cannot be classified by using a straight line. The classifier used in this case is referred to as a nonlinear SVM classifier. It has more flexibility for nonlinear data because more features can be added to fit a hyperplane instead of a two-dimensional space.

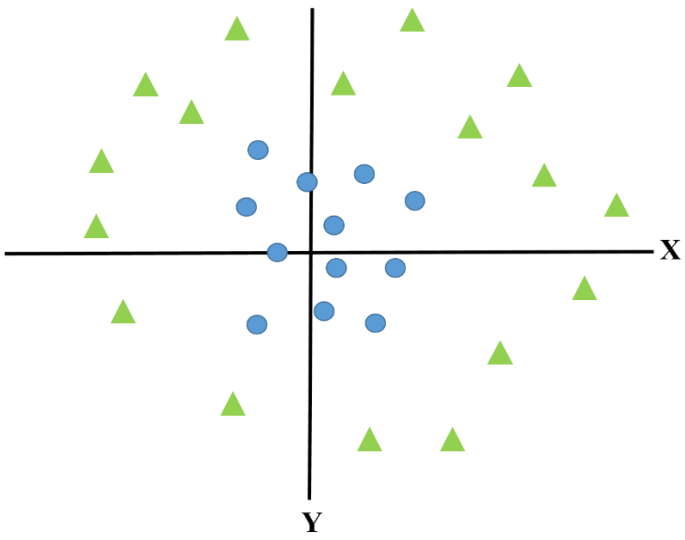


**Figure 4: Linearly Non-Separable Dataset.**

**Conclusion:** Using concept of KNN and SVM classification algorithms, we have classified emails into two class normal (non-spam) and abnormal (spam) and compared both, KNN and SVM, using evaluation metrics.

**Questions:**

1. What is decision boundary?
2. Explain train_test_split() function in detail.
3. What is the significance of MSE and MAE.
4. Explain parameters used for SVC.
5. List out Applications of KNN and SVM?