**Batch :** BE COMP SS P Batch

**Group no. :**  03

**Names with Sr. No :**

1. Deep Shende
2. Shweta Pal
3. Kanchan Bhil
4. Hemal Asapuri
5. Rohit Late
6. Praddyumn Wadekar
7. Pranit Kamble
8. Mahesh Kanchangire

**Project title**: Classification By Using Support Vector Machine

**Problem Statement :** Apply the Support Vector Machine(SVM) for classification on a dataset(Loan Prediction Analysis) obtained from Kaggle repository.

**Theory**

**Algorithm -** Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three. One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes.

**Working of SVM Algorithm -** Support Vector Machine algorithm is mainly used to solve classification problems. Support vectors are nothing but the coordinates of each data item. Support Vector Machine is a frontier that differentiates two classes using hyper-plane.

### **Linearly Separable Data :**

* + SVM takes all the data points in consideration and gives out a line that is called ‘**Hyperplane’** which divides both the classes. This line is termed as ‘**Decision boundary’**.
  + There can be many hyperplanes that you can see but the best hyper plane that divides the two classes would be the hyperplane having a large distance from the hyperplane from both the classes. That is the main motive of SVM to find such best hyperplanes.
  + There can be different dimensions which solely depends upon the features we have. It is tough to visualize when the features are more than 3.

### **Linearly Non-separable Data:**

* + If the data is non linearly separable as shown then SVM makes use of kernel tricks to make it linearly separable.
  + The concept of transformation of non-linearly separable data into linearly separable is called Cover’s theorem- “given a set of training data that is not linearly separable, with high probability it can be transformed into a linearly separable training set by projecting it into a higher-dimensional space via some non-linear transformation**”**. Kernel tricks help in projecting data points to the higher dimensional space by which they became relatively more easily separable in higher-dimensional space.

### **Kernel Tricks:**

* + Kernel tricks also known as Generalized dot product. Kernel tricks are the way of calculating dot product of two vectors to check how much they make an effect on each other. According to Cover’s theorem the chances of linearly non-separable data sets becoming linearly separable increase in higher dimensions. Kernel functions are used to get the dot products to solve SVM constrained optimization.
  + While using the svm classifier we can take the kernel as ‘linear’, ’poly’, ‘rbf’, ‘sigmoid’. Let us see which are the most used kernels that are polynomial and rbf (Radial Basis Function). You can refer here for documentation that is present on sklearn.

### **Degree of tolerance in SVM :**

* + The penalty term that is passed as a hyper parameter in SVM while dealing with both linearly separable and non linear solutions is denoted as ‘C’ that is called as Degree of tolerance.
  + Large value of C results in the more penalty SVM gets when it makes a misclassification. The decision boundary will be dependent on narrow margin and less support vectors.

**Dataset/Repository –** For this mini project, we are using Kaggle repository. The Kaggle repository is a collection of databases, domain theories, and data generators that are used by the many people in the world for their algorithm. It is used by students, educators, and researchers all over the world as a primary source of data sets. As an indication of the impact of the archive, it has been cited over 15000 times.

**Software/ Tools**:

* Jupyter Notebook
* Python

**Input** : Loan Prediction Analysis Dataset

**Expected output:**

Here, In the code below we saw how classification is done after applying the SVM algorithm

and how we remove the noise and unwanted data by using the preprocessing feature.

