Experiment. No. 08

Title : 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 neighbours and support vector machine for classification. Analyze their performance

objective : To classify email using the binary classification and implemented email spam detection technique by using k-Nearest Neighbours and support vector machine algorithm :

Dataset Description :
The csv file contains 5172 rows. each row for each email. These are 3002 columns. The first column indicates Email name. the name has been set with numbers and not recipient's name to protect privacy. The last column has the labels for prediction : 1 for spam, 0 for not spam.

Theory :

Data Preprocessing :
A real-world data generally contains noises, missing values & maybe a in an unusable format which cannot be directly used for machine learning models.
Data preprocessing is required tasks for cleaning the data & making it suitable for a ML model which also increases the accuracy & efficiency of a ML model.

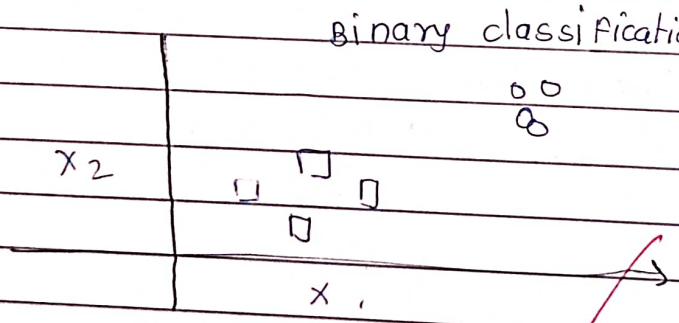
steps

1. Get dataset
2. Import libraries
3. Import dataset
4. Finding missing data
5. Encoding categorical data
6. splitting dataset into training & test set
7. Feature scaling

Binary classification

In binary classification, the goal is to classify the input into one of two classes or categories.

Example - on the basis of the given health conditions of a person, we have to determine whether the person has a certain diseases or not



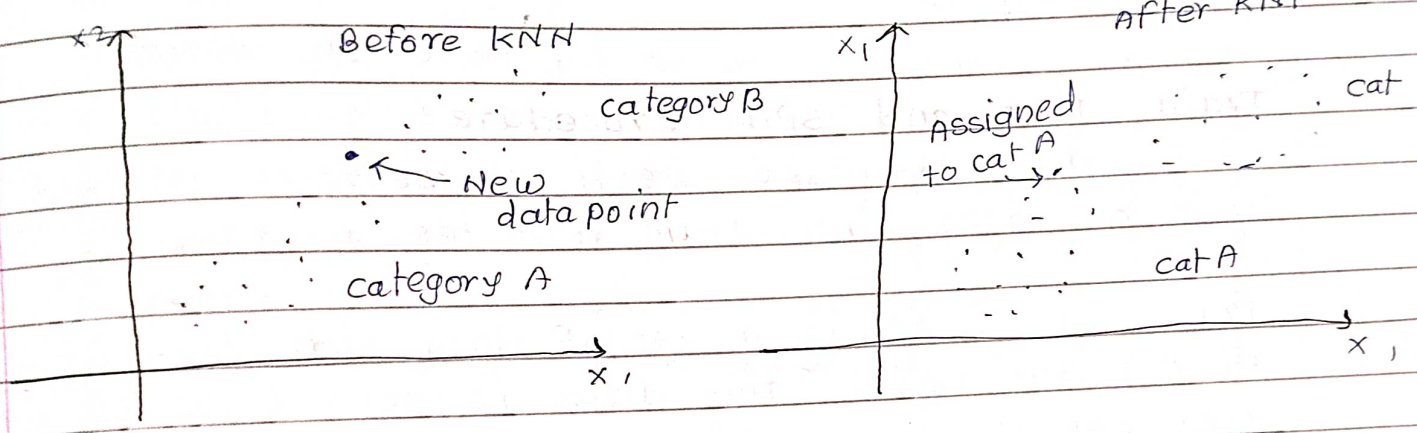
K-nearest neighbours -

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that much similar to the new data.

Example - suppose, we have an image of creature that looks similar to cat & dog. bat

want to know either it is a cat or dog. our KNN model will find similar features of the new data set to the cats & dog images & based on the most similar features.

It is a non-parametric as well as lazy algorithm.



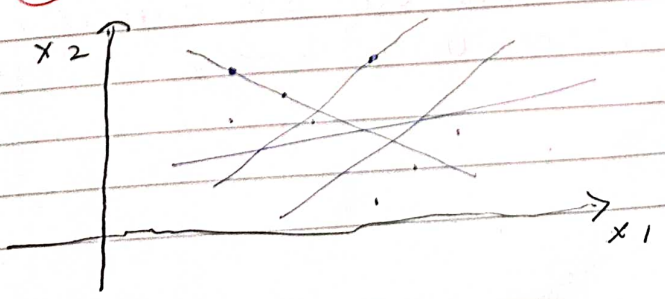
Support Vector Machine:

It is a powerful machine learning algorithm used for linear or non linear classification, regression and even outlier detection tasks

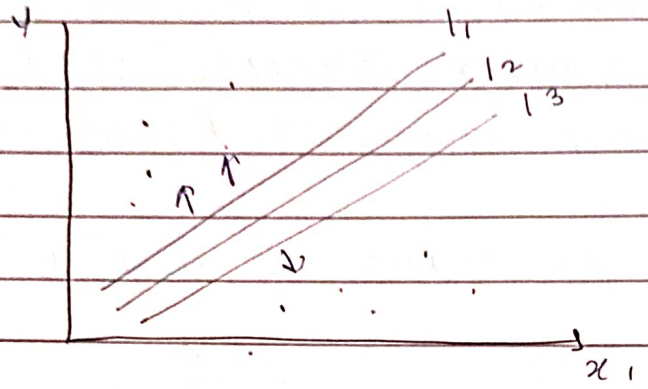
SVM can be used for a variety of tasks such as text classification, image classification, spam detection, handwriting identification, etc.

It is adoptable and efficient in a variety of a applications because they can manage high-dimensional data and non linear relationships.

main objective of SVM algorithm is to find the optimal hyperplane in an N-dimensional space.



It then separate data into two different datas



Train, Test and split Procedure:

The train_test_split() method is used to split our data into train and test sets.

Train set : It is a set of data that was utilized to fit the model. This data is seen & learned by mo

Test set : It is a subset of the training datas that is utilized to give an accurate evaluation of Final model fit

First, we need to divide our data into features (x) and labels (y)

Dataframe gets divided into x_train, x_test, y_train & y_test

x_train & y_train sets are used for testing the model if it's predicting the right outputs.

Conclusion:

In this way we classify email either it is spam or not using binary classification.