**Experiment 12: Support Vector Machine**

**Objective** :To implement Support Vector Machine (SVM) classification

**Time Required** : 3 hrs

**Programming Language** : Python

**Software Required** : Anaconda

**Introduction**

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. It is primarily used for classification problems and is particularly effective in handling complex datasets with high-dimensional feature spaces.

The main objective of SVM is to find the optimal hyperplane that separates data points from different classes in a way that maximizes the margin between the classes. The hyperplane is determined by a subset of the training data called support vectors. These support vectors are the data points closest to the decision boundary. In SVM, a hyperplane is a decision boundary that separates the classes in the feature space. For binary classification, it is a line in 2D or a plane in 3D. In higher dimensions, it is referred to as a hyperplane.

Overall, SVM is a versatile algorithm that offers robust classification performance, especially in complex and high-dimensional datasets, making it a popular choice in various domains.

**TASK:**

Apply SVM on the Breast Cancer data set. You will use principal component analysis to transform the data to a lower dimensional space.

***Steps to follow:***

1. Download the Breast Cancer data set from the following webpage: <https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset?resource=download>
2. Load the dataset.
3. Pre-process and visualize the data.

* Replace the ‘?’ mark in the ‘bare’ column by np.nan and change the type to ‘float’.
* Fill any missing data with the median of the column.
* Drop the ID column.
* Using Pandas, Matplotlib, seaborn (you can use any or a mix) generate 3-5 plots and add them to your written response explaining what the key insights and findings from the plots are.
* Separate the features from the class.
* Split your data into train 80% train and 20% test, use the last two digits of your student number for the seed.

d) Build Classification Models -

* Support vector machine classifier with linear kernel
* Train an SVM classifier using the training data, set the kernel to linear and set the regularization parameter to C= 0.1. Name the classifier clf\_linear\_firstname.
* Print out two accuracy score one for the model on the training set i.e. X\_train, y\_train and the other on the testing set i.e. X\_test, y\_test. Record both results in your written response.