**UNIVERSITY OF TECHNOLOGY AND EDUCATION**

**FACULTY FOR HIGH QUALITY TRAINING**

**Instructor: Dr Lê Văn Vinh**

**Project 03**

**PROJECT REPORT**

**Support Vector Machine**

**Nguyễn Hoàng Danh – 17110009**

**Đỗ Nguyễn Hoàng Hải – 17110024**

**SCHOOL YEAR: 2020 – 2021**

**HO CHI MINH CITY – December 2020**

Instructor's evaluation

*Ho Chi Minh City, January, 2021*

**Instructor**

**Contents:**

[Instructor's evaluation 2](#_Toc61680106)

[Chapter 1: Support Vector Machine 4](#_Toc61680107)

[**1.1.** **Introduction:** 4](#_Toc61680108)

[**1.2.** **Classification:** 5](#_Toc61680109)

[**1.3.** **How does SVM work?** 5](#_Toc61680113)

[1.4. Types of SVM: 6](#_Toc61680115)

[**1.5.** **Advantages:** 6](#_Toc61680116)

[**1.6.** **Disadvantages:** 6](#_Toc61680117)

[Chapter 2: Demo 7](#_Toc61680118)

[**2.1.** **Dataset** 7](#_Toc61680121)

[**2.2.** **Implementation and result** 7](#_Toc61680122)

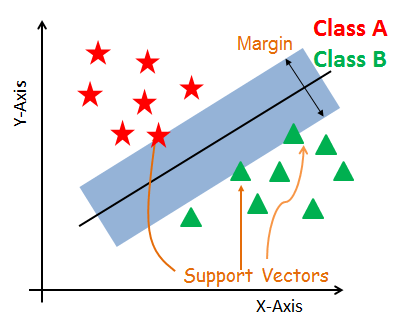
# **Chapter 1: Support Vector Machine**

## **Introduction:**

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.



***Support Vectors***

Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.

***Hyperplane***

A hyperplane is a decision plane which separates between a set of objects having different class memberships.

***Margin***

A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.

## **Classification:**

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y).

Classification belongs to the category of supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketing etc.



## **How does SVM work?**

The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:

* Generate hyperplanes which segregates the classes in the best way.
* Select the right hyperplane with the maximum segregation from the either nearest data points.

# **Types of SVM:**

* Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

## **Advantages:**

* Effective in high dimensional spaces.
* Still effective in cases where number of dimensions is greater than the number of samples.
* Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
* Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

## **Disadvantages:**

* If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
* SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

# **Chapter 2: Demo**



## **Dataset**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

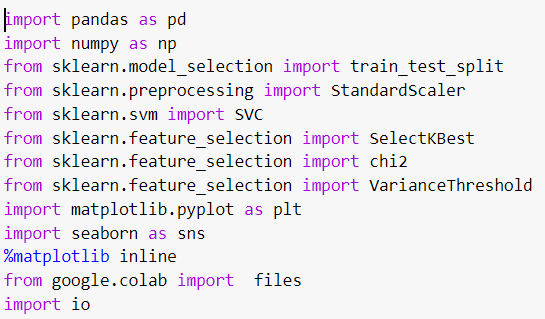
Table

Description automatically generated

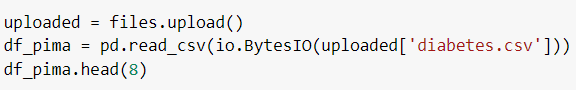
## **Implementation and result**

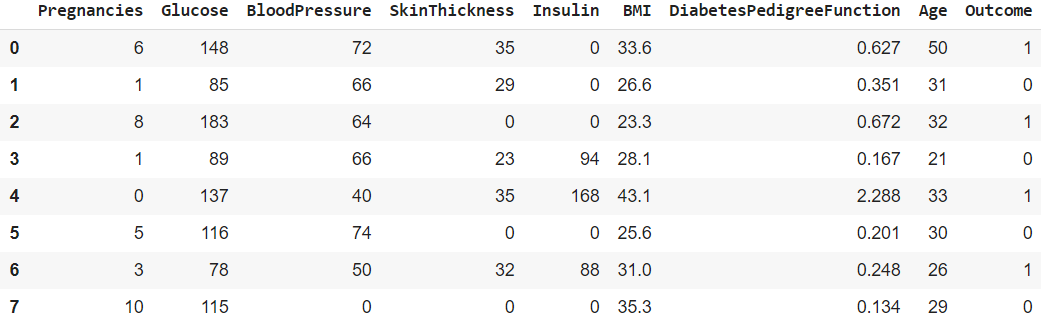
We use Google Colab to implementation

First of all, we need to import a lot of libraries

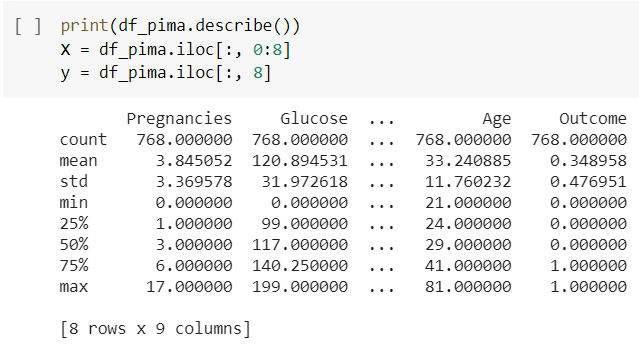


Next, we will load file from computer and show data

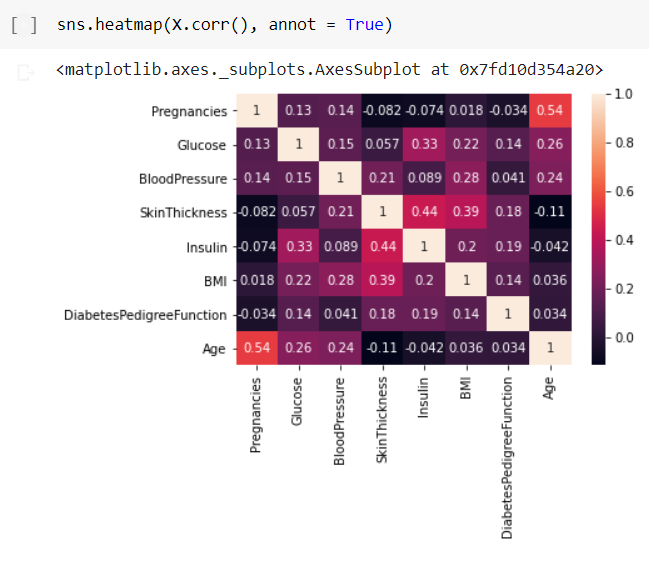




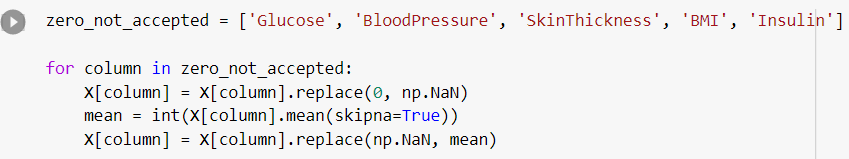
Describe the data with python



View the correlation



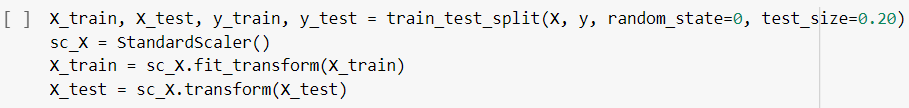
Replace Zeroes with the median value of the column



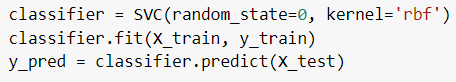
Feature extraction



Split dataset in 80(train data)/20(test data)



Implement SVM classification with Radial Basic Function Kernel



Evaluates the algorithm with Accuracy, F1-score, Precision, Recall and shows classification report

