Diabetes Prediction

Idea

Various medical data (BMI, Blood Glucose level, insulin, etc.) of a person is provided to the machine, and the machine predicts whether the person is diabetic or not. Dataset has 768 rows and 9 columns. Each row represents a person's medical data and whether that person is diabetic or not. Each column represents a certain medical result : Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome. Outcome column has 0 or 1, 0 representing NOT diabetic, and 1 representing diabetic.

Types Of Supervised Learning	Info and Example
Classification	Used to predict a class or a discrete value. eg. Male/Female, T/F, Mine/Rock, etc
Regression	Used to predict a quantity or continous values. eg. Salary, house price, age, etc.

Work Flow

(1) Collect the diabetes data

Real life data is used as sample to create the .csv file. This contains all the features mentioned above, and whether the patients were diabetic or not. This data is fed to the model to train and test it.

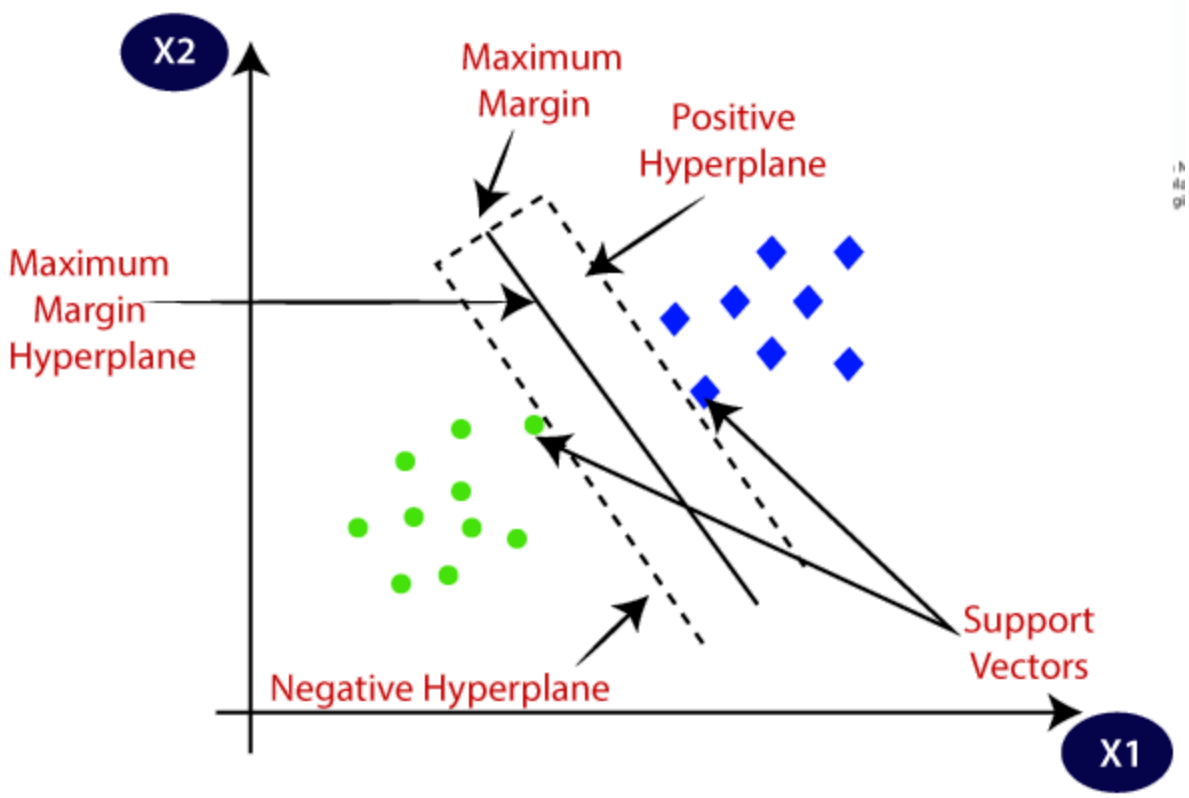
(2) Data Preprocessing

Machine cannot read the raw .csv data, and hence must be cleaned/processed. this includes finding and modifying null values, encoding the various features, and performing functions on the data which makes it easier for the model to understand. In this case, due to irregularity of ranges across columns, standardization of data must be done to fit all columns to the same range. Eg, Pregnancies (0 to 8) and Blood Pressure (60+). Not consistent throughout.

(3) Train Test Split

The data is split in 2 categories, training data, and testing data. We build the model based on the training data, and then evaluate the model accuracy on the remaining test data. eg, if there are 100 samples, and our split size is 10%, then we train our model based on 90 values and labels, and test it on the remaining 10 to check its accuracy

(4) SUPPORT VECTOR MACHINE MODEL



Support vector machines are a set of supervised learning methods used mainly for classification, regression and outliers detection. One of the biggest advantages of using SVMs is that they are still effective where number of features exceeds the number of samples, however, this also causes overfitting the model, so kernel function selection is very crucial. Here, linear function is used.

Linear kernel is the most commonly used kernel functions. This is because it is very effective when the given data is separable, and can be separated by a single line (in this case, as shown in the diagram above, the hyperplane distinguishes Diabetic and Non-diabetic.

(5) Trained Support Vector Machine model

We enter brand new data in our trained model, and this will predict whether person is diabetic(1) or not(0).

(6) Model evaluation

First, the accuracy score of the fitted model is tested on the trained data. In this case, the accuracy score for the trained data is 0.786 (about 78.6%) which is good.

Then, the accuracy score of the fiited model is tested on testing data. Here, that is 0.772 (77.2%).

The accuracy score of tested data is expected to be lower than that of trained data, as the model is fitted according to trained data first. However, since the number of features in this model were very high, and linear kernel type SVM was used, it is common to see trained accuracy score almost the same as (or even lower by a bit) as test accuracy score.

The scores are pretty similar, hence this model is not classified as overfit.