Practical work 3 - ML : SVM

Training a binary classification model using support vector machines

The objective:

The aim of this work is to learn how to :

– Train a support vector machines model on a dataset,
– Find the hyperplane equation and the support vectors,
– Drawing the hyperplane and the margins
– Use the kernel trick,
– Defining the best hyperparameters using GridSearch method,
– Evaluate the SVM model before and after using GridSearch.

The required work :

The construction of a binary classification model using support vector machines to diagnostically predict whether a person is diabetic or not, based on 5 features: (*Glucose*, *BloodPressure*, *Insulin*, *BMI* and *Age*) among the 8 diagnostic measures in the dataset (*diabetes.csv*).

1. Create the X matrix containing the five columns (features) mentioned above and the Y vector.
2. Use the **Scikit-learn** machine learning tool to create and train a **support vector machines** model in a new notebook in Jupyter.
3. Implement a function that mimics the function **train_test_split()** of **Scikit-learn**, This allows the dataset to be divided into two random subsets: training and testing.
4. Find the **hyperplane** equation and the **support vectors**,
5. Write a function to compute the **confusion matrix** from scratch without using **Scikit-learn**.
6. Calculate the rates: **True Positive (TP), False Negative (FN), False Positive (FP), True Negative (TN)** from the **confusion matrix** and then calculate the **Accuracy** in a function.
7. Use **GridSearch** method in order to find the best hyperparameters of the **SVM** model.
8. Based on the **GridSearch** method, what is the appropriate hyperparameters that gives the best **accuracy** ?
9. Repeat question 6 after having applied the **GridSearch** method.
10. Since our model is trained on 5 features of the dataset, create 2D scatter plots for every possible pair of features (e.g., (*Glucose*, *Insulin*), (*BMI*, *age*), etc.). On each plot, draw the SVM hyperplane (*decision boundary*) and the margin lines that separate the classes.
11. Using the **SVM** model, what is the prediction of a person who has the following measurements: (*Glucose*=100, *BloodPressure*=90, *Insulin*=0.5, *BMI*=55 et *Age*=63).