Dataset Introduction :- Pima Indians Diabetes Database, This dataset is a publicly available database from the National Institute of Diabetes and Digestive and Kidney Diseases. It contains 768 instances and 9 features, including the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, insulin, body mass index, diabetes pedigree function, age, and the target variable indicating whether the individual has diabetes or not.

Algorithm Explanation :- I used the Support Vector Machine (SVM) technique with a linear kernel in this project. SVM is a popular machine learning technique used for regression and classification. Identifying the ideal border or hyperplane that divides the data points into multiple classes is the aim of SVM.

Finding the best hyperplane to maximize the margin between the two classes of data is important to the linear SVM's operation. The margin is the separation between each class's nearest data points and the hyperplane.

I just choose SVM because i wanted to learn the working of it and it is also easy to use, It also works best when doing binary classification.

Implementation Details :-

1) Load the data from OpenML.
2) Convert the OpenML data into pandas dataframe.
3) Description and information of the data.
4) Visualize using Box plots
5) RFE was used to get the best features.
6) An SVM model was built for different number of features.
7) Visualized the features from the most accurate model using bar plot’s.
8) I took all the medians of the features because the data is not normalized.

Results and Analysis :- The accuracy and precision of the model were calculated on the testing set. The accuracy of the model with 5 features was found to be 0.779, and the precision was 0.714.

Further analysis of the model can be done by seeing the Confusion matrix below(fig a).

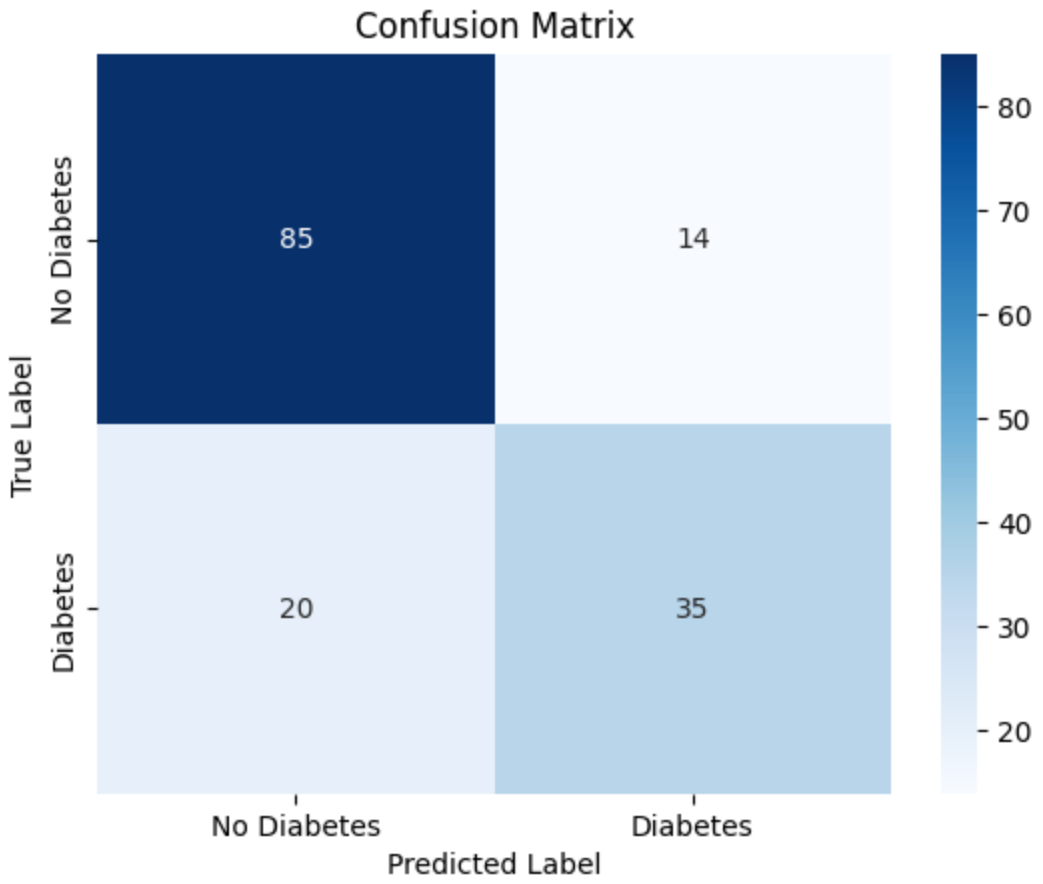

Fig : a

For future improvements we can use different feature selection methods and also the data is not completely normalized. We can try different model's on this dataset and see which one produces the better output.