ML project 1 - predicting diabetes

Problem type - Classification problem

Softwares used - python 3.10, numpy, pandas, sklearn

Editors used - jupyter notebook

Ralle® Technologies

Project steps

Step 1 : loading patients data using pandas

Spte 2 : cleaning data

Step 3 : Training the algorithm

Step 4 : Testing the algorithm

Step 5 : Improving the algorithm

Step 1 - loading the data

- We are using pandas software to load patients data (xls) into jupyter notebook.
- pd.read_xls() function is used to load data

Step 2 - cleaning data

- We will remove duplicate column or correlated columns by using df.corr() function
- We will remove rows with NAN values using dropna() or else we can fill NAN values with fillna() mean strategy.
- We will convert text data into number data using dictionary mapping technique

Step 3 - training the algorithm

- We will split the data into training data and testing data in 70% and 30% ratio.
- We will give 70% data to algorithm for training
- Since it is classification problem, I have tried with naive bayes algorithm, random forest algorithm, and logistic regression algorithm.
- For training algorithm we will use fit() function

Step 4 - testing the algorithm

- We will test the algorithm by using predict() function.
- We can also measure the accuracy of the algorithm using confusion matrix and also by using accuracy.
- I have picked logistic regression algorithm for my project because it was giving better accuracy and also better recall % (in confusion matrix), compared to other algorithms.

Step 5 - improvising algorithm

- We can improve the accuracy by using tuning parameter which is supported for logistic regression algorithm.
- We can also improve the accuracy by getting more data or else by cleaning the data more carefully.