MAJOR PROJECT SUMMARY

In this project we have taken a diabetes dataset and discussed about the basic machine learning workflow steps such as data exploration, data cleaning steps, model selection using Scikit Learn library. First of all we have imported the required libraries such as pandas ,matplotlib, seaborn after that performed data wrangling on this dataset and got the answers of 2 questions asked on this dataset shown below .In the machine learning part we have used 3 classification algorithms like logistic regression ,K nearest neighbour (KNN) and random forest. we have imported the train test split from sklearn.model_selection and accuracy score from sklearn.metrics for training testing and checking the accuracy of the diabetes dataset. After the application of the 3 algorithms that are logistic regression, KNN and random forest we have got accuracy score of 76%,72% and 77% respectively. From this we can conclude that random forest algorithm has the best accuracy score (77%) as compared to logistic regression and KNN algorithm on the diabetes data set.

Questions and Answers

Q.1 Calculate the Number of diabetic patient and non diabetic ?

Ans- diabetic patient =475

non-diabetic patient=249

Q.2 Which Age has the highest diabetes patients ?

Ans- At the age 25 it has highest diabetes patients .The total count is 13.