3. Problem solving

Guidance note:

Firstly importing the libraries. Pandas library is for dataframe, numpy for working with arrays, seaborn for making statistical graphics, sklearn for machine learning models, matplotlib for visualizations.

Use shape command for number of rows and columns in the data. describe() for getting statistical measures of the data. There are 8500 diabetic and 91500 nondiabetic persons. Then check there is any missing value in the data. we found that there is not any missing value in the data. Check there is any duplicate value in the data. we found that there are 3854 duplicate values in the data. Using drop command we eliminate duplicate values. Then separating data and labels as X and Y. Y is column 'Diabetes' and X is except 'diabetes' all other columns. Then X3 is data frame of numerical variables. Scaling numerical variables using StandardScalar() , X5 is data frame of categorical variables. Then encoding categorical variables. Merge the dataset using concatenate. Find correlation using corr() and create heatmap. In heatmap, dark colours represent high positive correlation and light colour represents low positive correlation. Intensity of correlation decreases from dark colour to light colour. In our data all variables are positively correlated. Splitting data into train and test set.

Then use machine learning algorithms to fit the model.

Firstly use logistic regression. Fit and predict the model. Accuracy is 0.959.Then plot Roc curve.We found that area under curve is 0.8028. So, logistic regression is good fitted model for the data. Use Random forest. Fit and predict the model. Accuracy is 0.965. So, Random forest is good fitted model for the data. Use SVM. Fit and predict the model. Accuracy is 0.958. So, SVM is good fitted model for the data. Use Naïve bayes. Fit and predict the model. Accuracy is 0.609. So, Naïve bayes is not good fitted model for the data.

Among all the algorithms accuracy of Random forest is high so it is best classifier.