Nima Kelidari
SN: 98108124

Second homework of analysis of regression

I decided to use the second dataset which data is 50-50 for diabetic and non-diabetic persons. this balance will help us to decrease bias in our models. And it is cleaner and has less same row and NA rows so much. (But we try to detect and delete them anyway)

1. Yes, by using the proper features and model, we surely can get a nice prediction and then in the next part, use it for inference. The main reason for this is the data given to us, has some features which probably have some relationship with the probability of whether one has diabetes or not. We can show it by plotting those features by the diabetes factor. For example, we plotted BMI, high blood pressure, high cholesterol, general health, age, education, and income by diabetes factor in the type of histogram density, histogram, and box plot. Here we can see the different distribution of these factors for two groups of diabetic and non-diabetic persons. Moreover, we fit a logistic regression to show t-value is so big for this problem and that these factors and responses are correlated strongly. At last, we fit a model by these factors and see almost all of them, have strongly related to the diabetes factor. So as least some of them can be used for the prediction of whether one has diabetes or not. In the second part, we introduce some of these features as good ones for prediction.
2. We can answer this question by feat a model like logistic regression on whole data and all of the features, then detect those features which have big enough t-values. Here we can see BMI, high blood pressure, high cholesterol, general health, age, cholesterol check, history of heart attack, sex, age, heavy alcohol consumption, and income have an absolute value of t-value more than 10; We can detect these factors as important factors compared to others and check them more carefully. we will see how we can select features more reliably by feature selection in part 3.
3. Yes, absolutely we can do it by feature selection methods. As best subset selection can be very time-consuming and computationally heavy in practice (in real cases), we just use two methods forward and backward selection by AIC. In each one, we want to use 5 features maximum utmost. so we make a full model and null model first and try to extract the five most important features. By each of these two methods, we can see five features below BMI, high blood pressure, high cholesterol, general health, and age, have the most importance in prediction as we guss before in part 2; in other words, these have most dependency to pred factor. On another hand, these do have not a significant relation to each other, because, in a model that we created before, they didn't decrease each other t-value strongly. so we use these five features for use in the next parts.
4. We will use the five features that are mentioned in the last part. If we use all of the features or more than five, our efficiency for the trained model will drop. in this part, we use 70 percent of the data for the training and 30 of the data for the test. here we will not use a validation set and we prefer to use cross-validation, more specifically 5fold-cv for it,s time-consuming. then we will try to train the model, tune its hyperparameters, and then test it and get accuracy and precision, and recall from them, then make the confusion matrix for each model based on the test set. We here will use QDA, LDA, Randomforrest, Neural Network, Tree, Linear SVM, Logistic regression, and KNN

models. after the test, we will save the accuracy of each model and show and compare them at last and choose one of them. (Here we can see the Tree model had excellent performance besides its simplicity)

5. Yes, we can do this thing. For this target, we need to use a simple and appropriate inferentially. As we want to ask some questions as less as possible from a person, I prefer to use the Tree model. In the last part, we understood Tree can do nice predictions too. so use a big tree and use cross-validation to get how many nodes and terminals we need for a good model. So we choose 5 and get accuracy, a plot of the Tree, and a confusion matrix for this model. By this tree, we can by asking just some questions, we can predict whether one has diabetes or not, with acceptable accuracy.