Mini Project 3

MATH 4530/5530

Spring , 2024

Problem 1
In this problem you will work with the US health insurance data set that is uploaded in Canvas, and develop a predictive model to estimate individual medical charges based on personal attributes using linear regression.Follow the following steps.

1. Do some descriptive statistics and data visualization to explore the variables in the data.
2. Split the data into train-test data sets.
3. Fit at least 3 different regression models to the train set using different features or different combinations of the features.
4. Find MSE of each model by applying these models to test data set, and choose a model that has the smallest MSE.

Problem 2
In this problem you will work with the heart disease data set uploaded in Canvas, and a method that works best in this data set to classify whether a person has heart disease or not depending upon the features. Follow the following steps.

1. Do some descriptive statistics and data visualization to explore the variables in the data.
2. Split the data into train-test data sets.
3. Fit the 3 methods for classification we have studied: Logistic Regression Model,K-nearest neighbor classifier and Naive Bayes Classifier to the train dataset.
4. Apply the fitted methods to the test data set.
5. Find the following for each: TPR,TNR,FPR,FNR and Accuracy for each. For better performance, TPR, TNR,ACC should be high and FNR, FPR should be low.
   TPR = $\frac{TP}{TP+FN}$
   TNR=$\frac{TN}{TN+FP}$
   ACC=$\frac{TP+TN}{TP+TN+FP+FN}$
   FNR=$\frac{FN}{FN+TP}$
   FPR=$\frac{FP}{TN+FP}$
6. Draw a ROC Curves on the same plot.
7. Find the method that works best on this data set.