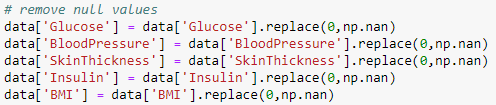
**Homework 1: Data Analysis**

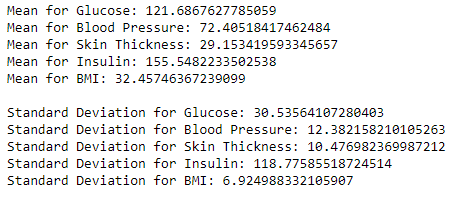
1. Download the Pima Indians Diabetes dataset. It has 8 attributes and a binary class variable, indicating the following information

Attribute 1: Number of times pregnant   
Attribute 2: Plasma glucose concentration 2 hours in an oral glucose tolerance test   
Attribute 3: Diastolic blood pressure (mm Hg)   
Attribute 4: Triceps skin fold thickness (mm)   
Attribute 5: 2-Hour serum insulin (mu U/ml)   
Attribute 6: Body mass index (weight in kg/(height in m)^2)   
Attribute 7: Diabetes pedigree function   
Attribute 8: Age (years)   
Attribute 9: Class variable (0 or 1)

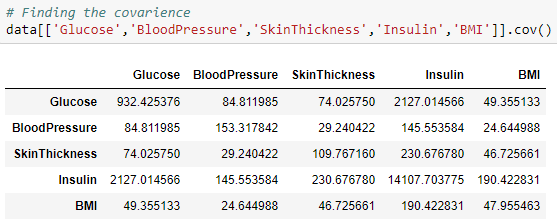
1. Remove all null values



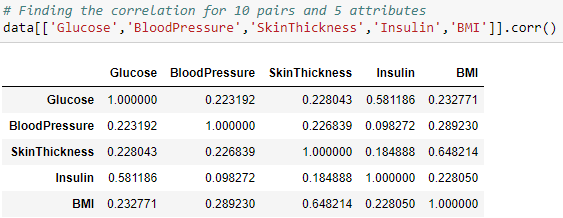
1. Compute the mean value and standard deviation for attributes 2 – 6.

****

1. Compute the covariance matrix for attributes 2 – 6.

****

1. Compute the correlations for each of the 10 pairs of the 5 attributes

****

1. Explain the results for computing the correlations and covariance matrix from questions 4 and 5.

Covariance is a way of telling if two variables are related positively or negatively.

1. Glucose and insulin are positively related
2. Blood pressure and skin thickness have the highest covariance for that column
3. Skin thickness and insulin has the highest covariance for that column.
4. Again insulin and glucose have the highest covariance
5. BMI and insulin have the highest covariance together for that column.

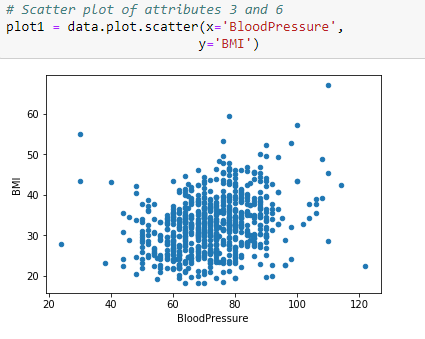
All of these covariance are positive and move together in the same direction when one increases.

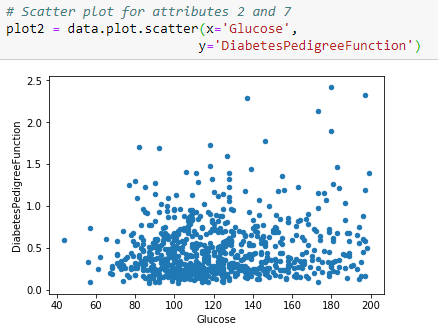
Correlation is another way of telling how variables are related. One(1) being the highest positive and -1 being highest negative correlation and 0 being the lowest.

1. Glucose and insulin are moderately correlated
2. Blood pressure and BMI are weakly correlated
3. Skin thickness and BMI are moderately correlated
4. Insulin and Glucose are moderately correlated
5. BMI and skin thickness are moderately correlated

The relation are all positive this mean when one variable moves up the other will also follow the same positive trend.

1. Create a scatter plot for attributes 3 and 6 of your dataset and another scatter plot for attributes 2 and 7.

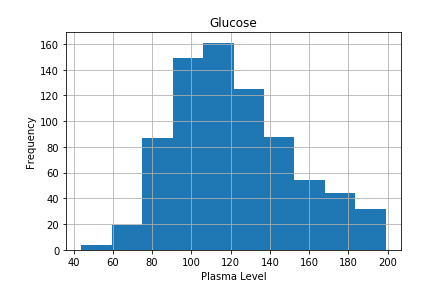
****

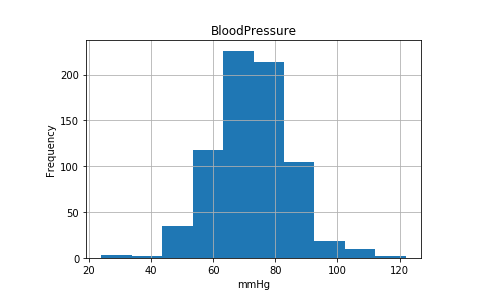
****

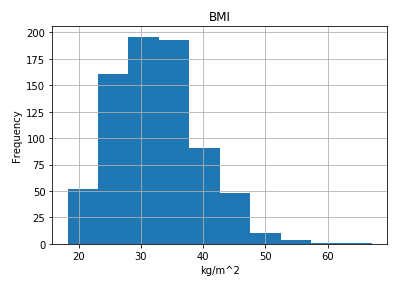
1. Interpret the results obtained in question 7.

The scatter plot for attribute 3 and 7 show no correlation and doesn’t look like there is a trend. For attributes 2 and 7 again the graph show no trends.

1. Create histograms for attributes 2, 3 and 6.

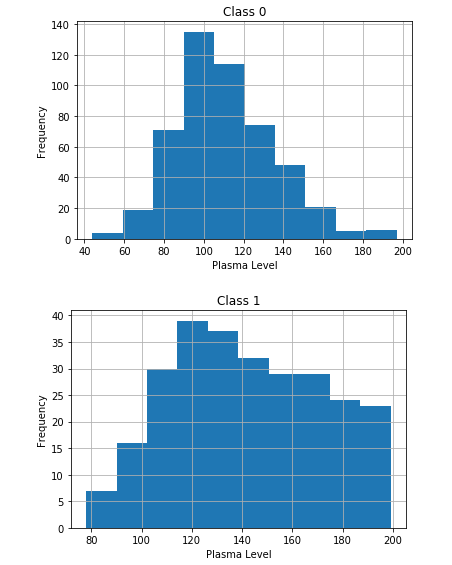
****

****

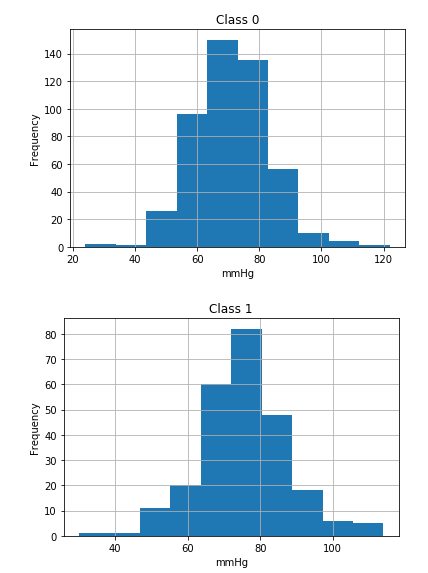
****

1. Create the same histograms for the 3 attributes for the instances of class 1 and for the instances of class 0.

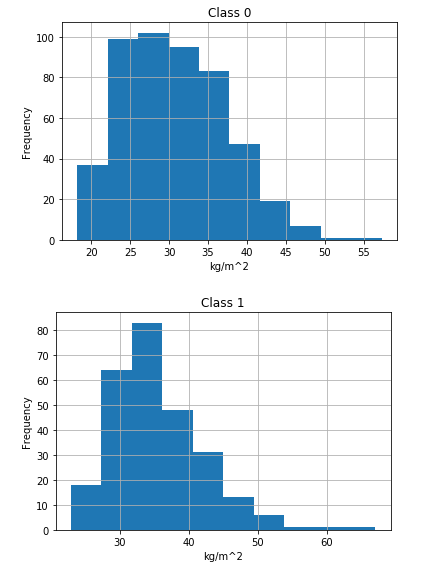
Glucose



Blood Pressure



BMI



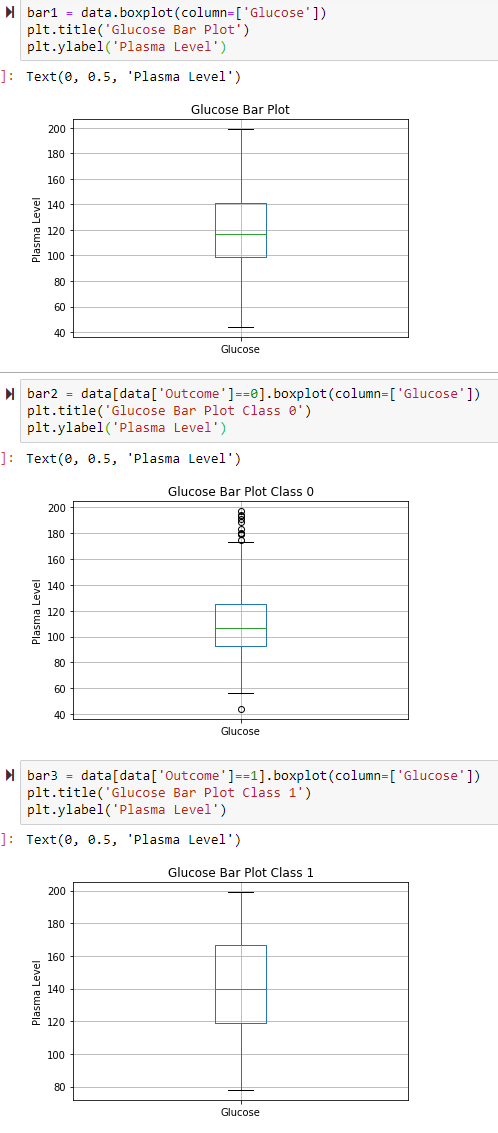
1. Interpret the 9 total histograms obtained in questions 9 and 10.

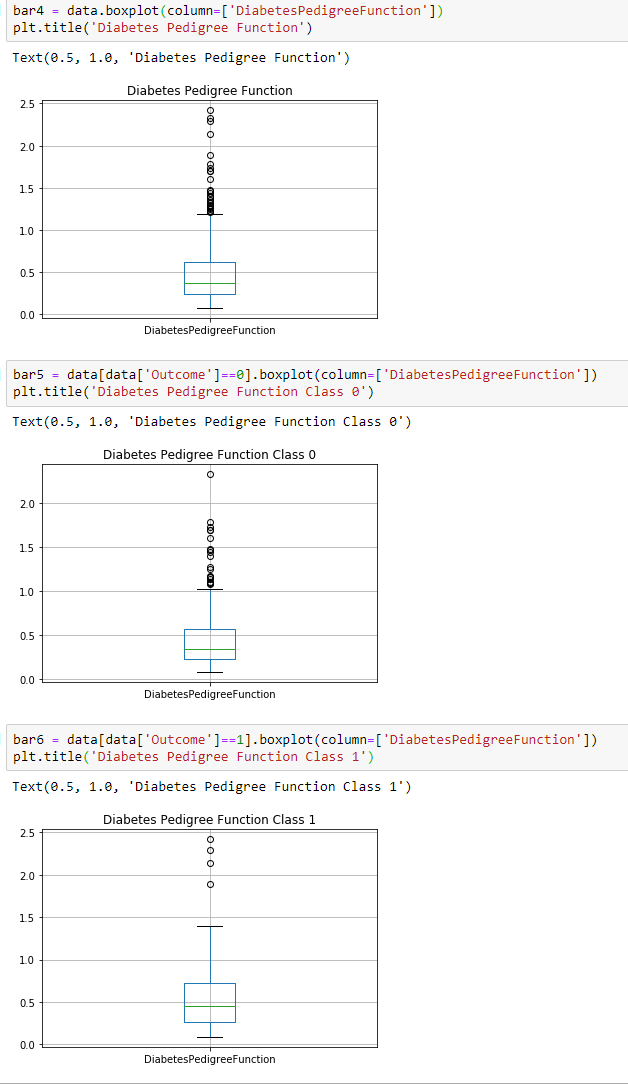
The the first bar plot indicated that the diabetes patient have higher glucose in their plasma compared to non diabetic patients.

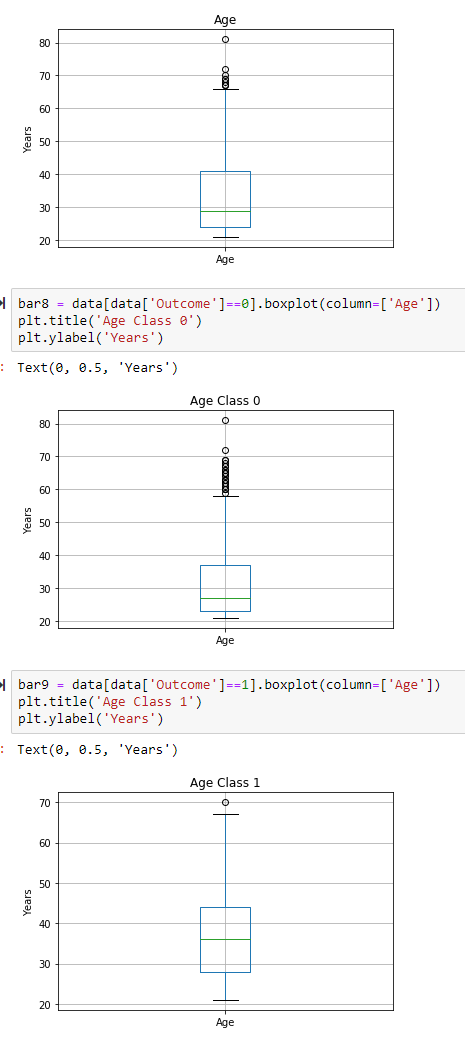
For the second bar plot both groups groups exhibit the same trends in blood pressure.

The last graph shows patients BMI and there does not look like there is significant difference between patients’ BMI.

1. Create box plots for attributes 2, 7 and 8. One for the whole dataset and one each for the instances of the two classes.

****

****

****

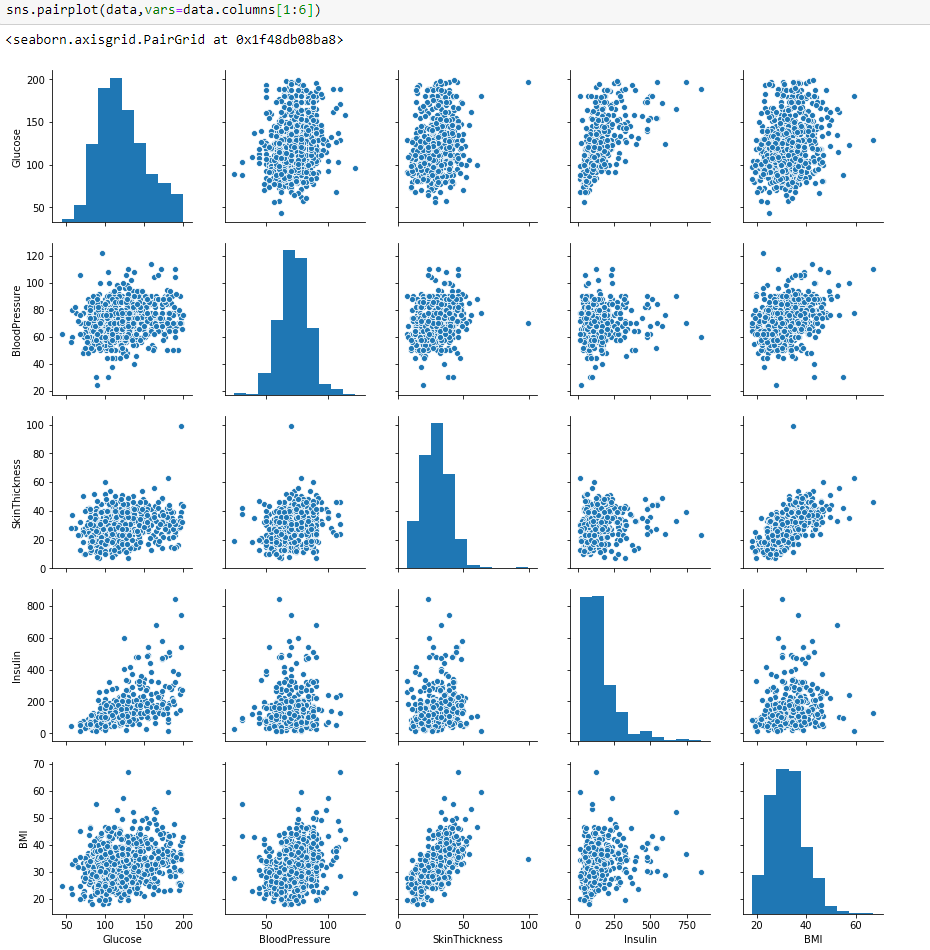
1. Interpret and compare the 9 boxplots obtained in question 12.

The glucose levels for the diabetic patients have many outliers indicating higher levels of glucose in blood.

The second set of graphs have a higher number of outliers indicating there will be high chance of developing diabetes in the next 5 years.

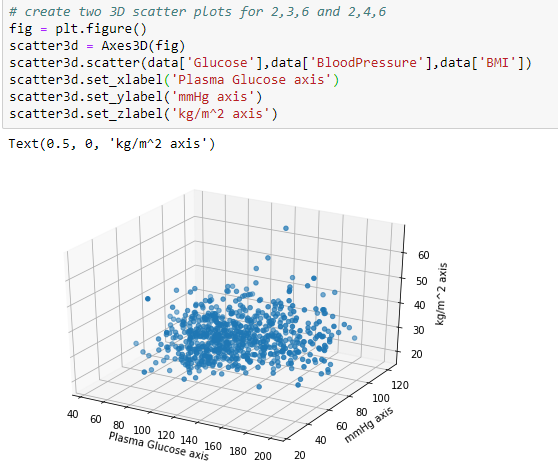
The last set of boxplots there is a higher number of outlier in the class 0 group indicating that older population will have a higher chance of developing diabetes.

1. Create supervised scatter plots or all pairs of attributes 2 – 6.

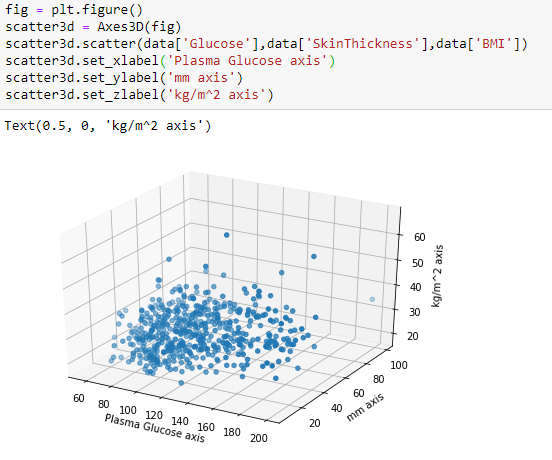
****

1. Create two 3D scatterplots: one for attributes 2, 3 and 6 and one for attributes 2, 4, 6.

**Plot 2,3,6**

****

**Plot 2,4,6**

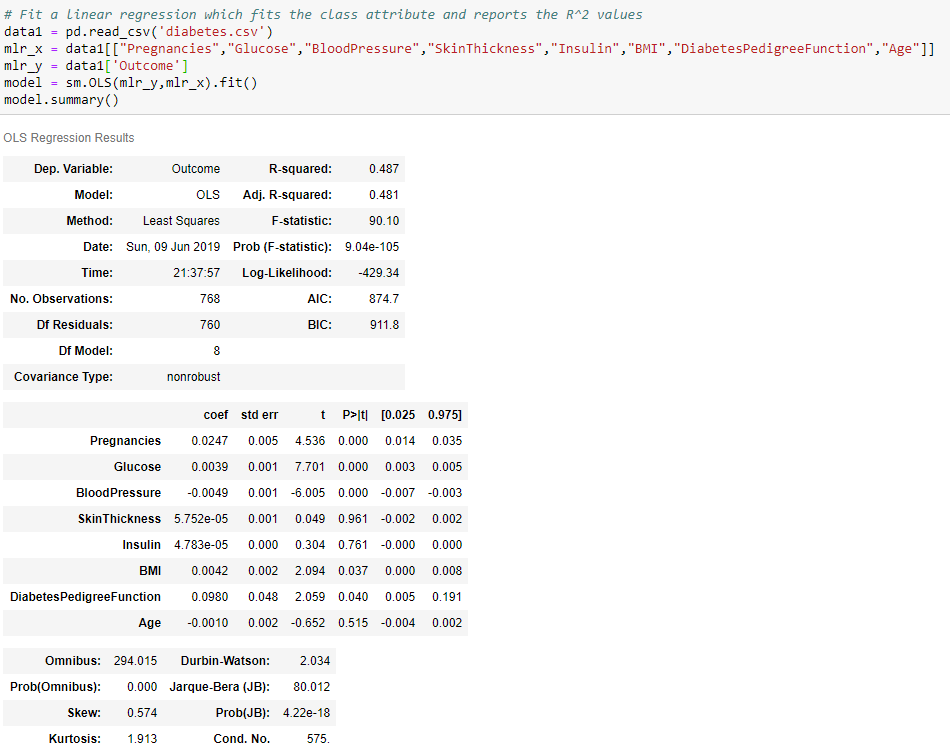
****

1. Interpret the scatter plots obtained in question 15. Address what can be said about the difficulty in predicting diabetes. Also, address the usefulness of the 3D scatterplot compared to the 2D scatterplot.

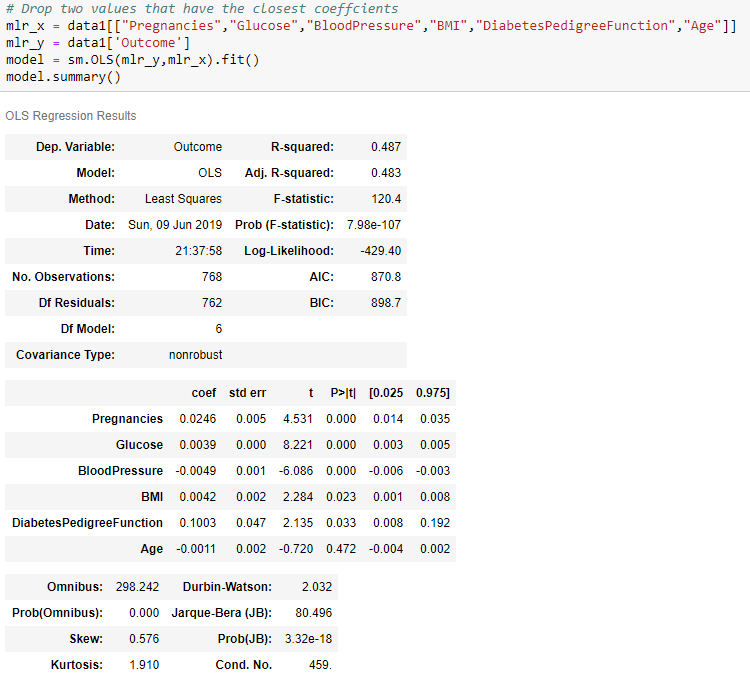
Both scatter plots do no depict any trends since none of the data points increase together or decrease together. There are some outliers that do no clump in the middle. 2D plots tend to share more about the relationship between two variables more clearly. While 3d plots the x and y axis will be the predictor values and z values are the response. As explained about the z axis show now trends and it is hard to tell if two values are correlated.

1. Fit a linear model that predicts the class attribute using the 8 attributes of the dataset as independent variables. Report the R2. Next, drop the two attributes, whose coefficients are closest to and obtain another linear model using the 6 remaining attributes as an independent variable.

8 Attributes



Drop two values



1. Do the coefficients from question 19 tell you anything about the importance of the attribute in predicting diabetes? What about negative and positive coefficients?

Both values skin thickness and insulin does not affect the R^2 since the values for the coefficients are close to zero, they have very little effect on the response variables. Negative and positive coefficients means the predictor variable degrees or increases in the case of positive to the response variables. Since these values are small, the relationship are very weak.

1. Compare the two regression functions from question 17.

The R squared values are identical for both functions so removing the coef closed to 0 didn’t change the function (0.481 vs 0.487). So the equation generate by statsmodel only explains 48% of the variance in the dependent variable. The coefficient values are all close to 0 and standard errors are also close to 0.

1. Create 3 decisions tree model with 20 or less nodes using the cleaned dataset.

1. Using the 3 decision trees from question 20. Explain how these decision trees were obtained Report the training accuracy and the testing accuracy for each decision tree. Interpret the learnt decision tree. What do they tell you about the importance of the 8 continuous attributes for the classification problem?
2. Write a conclusion (at most 20 sentences) summarizing the most important findings of the assignment. What did we learn about the dataset? Address the findings obtained related to predicting the class attribute.