Question 1

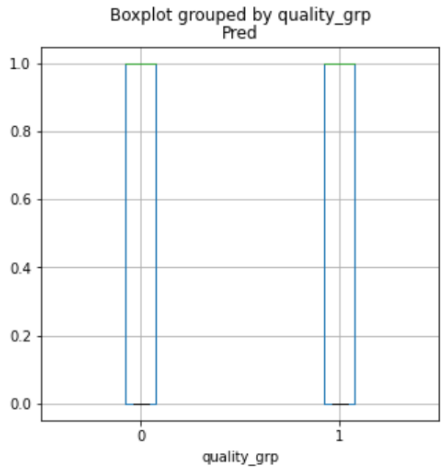
a) What is the Misclassification Rate of the classification tree on the Training data at Iteration 0 (i.e., when all the weights are one)?
Answer: 0.1673

b) What is the Misclassification Rate of the classification tree on the Training data at Iteration 1?
Answer: 0.1537

c) What is the Misclassification Rate of the classification tree on the Training data when the iteration converges?
Answer: 0

d) What is the Area Under Curve metric on the Testing data using the final converged classification tree?
Answer: AUC= 0.32229

e) Generate a grouped box-plot for the predicted probability for *quality_grp* = 1 on the Testing data. The groups are the observed *quality_grp* categories.



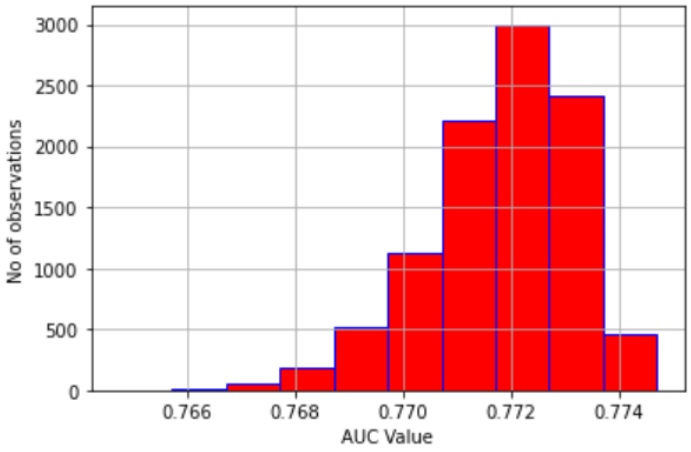
Question 2

a) Use the Forward Selection method to select input features into the model. The final model must include the Intercept term. Use $\alpha = 0.05$. Which input features did you enter into the model?

```
Model = Intercept + alcohol + citric_acid + free_sulfur_dioxide + residual_sugar + sulphates
```

b) What is the Area Under Curve metric on the Testing data?
Answer: AUC: 0.7723613128085971

c) Generate 10,000 Bootstrap samples from the Training data. Your random seed is 20210415. Then train a logistic regression model on each Bootstrap sample. The model will contain the input features that you have selected in (a). After each logistic regression model converges, calculate the predicted probabilities and the Area Under Curve metric on the Testing data. Generate a histogram of the 10,000 AUC metrics. The histogram width is 0.001.



d) Using the numpy.percentile function, calculate the 2.5th percentile and the 97.5th percentile of the 10,000 AUC metrics. What are the two percentile values?

95% Confidence Interval: 0.7687049, 0.7738949

e) The two percentiles in d) will be the lower and the upper limits of the 95% confidence limits for the AUC on the Testing data. If the value 0.5 falls within the confidence limits, then statisticians will conclude that the AUC on the Testing data is not significantly different from 0.5. Based on your 95% confidence limits, what is your conclusion?

Answer: The AUC on testing data is significantly different from 0.5. This is because 0.5 doesn't fall in intervals of confidence levels.