STAT 5428 Midterm - Skills Part **March 24- March 28, 2022**

Name: _____

Student ID #: _____

Please read the following instructions carefully.

1. There are two questions in this exam and each carries 50 marks.
2. Please create a single pdf file with your answers. Please note that software code is never an answer to the question. You must write your answer explicitly and document your interpretations. The necessary code will be moved to appendix.
3. You must show all the work and provide sufficient explanation to justify all answers. Correct answers with inconsistent work will not be given any credit.

Number	Max Possible	Points
1	50	
2	50	
Total	100	

1. Consider the *winequality-red.csv* data. Please see the following link for a brief description about the data: https://archive.ics.uci.edu/ml/datasets/Wine+Quality. The data includes the following attributes:

   Input variables (based on physicochemical tests):

   (a) fixed acidity
   (b) volatile acidity
   (c) citric acid
   (d) residual sugar
   (e) chlorides
   (f) free sulfur dioxide
   (g) total sulfur dioxide
   (h) density
   (i) pH
   (j) sulphates
   (k) alcohol

   Output variable (based on sensory data): quality (score between 0 and 10)

   please do the following and provide your comments in each case:

   (a) Using graphical methods, can you say something about the distribution of the wine quality? Do many of them have low/high quality?
   (b) Which are the independent variables have either positive or negative association with the output variable "quality"? Do you observe any nonlinear association/ no association at all?
   (c) Perform a simple linear regression of *sulphates* on the outcome variable *quality*? Is the effect significant? Is the association linear at all?
   (d) Perform a multiple linear regression with all the Input variables as predictors and the output variable as the response variable. Check the regression assumptions. Are all the assumptions are satisfied? If not how do you amend them? How much variation in *quality* is explained by the model?
   (e) For the model in (e), rerun the regression model using the Box-Cox transformed *quality* as the response variable. Please check the assumptions. Do you see any improvements? What is the optimal transformation? What is the R^2 and is it better than the original model?
   (f) Please fit the ACE model and check the R^2 value. Do you see any improvements? Please show what are the transformations considered for each variable. You may use scatter plots.

2. Autistic Spectrum Disorder (ASD) is a neurodevelopment condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Please find the attached dataset related to autism screening of adults that contained 20 features to be utilised for further analysis especially in determining influential autistic traits and improving the classification of ASD cases. This dataset includes ten behavioural features (AQ-10-Adult) plus ten individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science. For more details, please see the attached description file:

   (a) What is the response variable of interest in this study? What are the predictor variables that we can consider? Are there any missing values? what do you want to do with them.

   (b) Please explore the association graphically between the response variable and the predictor variables. Which behavioral variables are positively associated with ASD? Which individual characteristics variables are positively associated with ASD?

   (c) Fit a logistic regression model with autism as the dependent variable and all the other useful variables as predictors? Which predictor variables are significant? which of them positively associated with the outcome variable? Interpret your findings in the context of the given problem.

   (d) Can you drop the insignificant variables from the model? Does it improve?

   (e) What is the accuracy of your model at threshold 0.5? Please construct an ROC curve and identify an appropriate threshold. Justify your answer.

   (f) Construct the confusion matrix at the identified threshold from the previous step and compute: accuracy, sensitivity, specificity, ppv and npv. Interpret each of the statistic in the context of the problem.