
Course Name: STA4102

Course Code: STA 4102

Dept: Department of Statistics and Datascience

Instructor: Alexander V. Mantzaris

Exam Date: December 9, 2022 (due 10:00am-12:50pm)

---

INSTRUCTIONS TO CANDIDATES

1. This is the 'take-home and in-class component' to the exam. Work is to be done using software such as R, SAS, Python, MATLAB, Octave, C/C++.
2. The time slot for the inclass component of the exam is: Friday, December 9, 2022 due 10:00AM - 12:50 PM
3. Work is to be done individually.
4. Partial credit will be given where for incorrect answers that include an exposition of the thought processes that went into your results and code that contains an error ('bug') which is along the correct approach. It is expected that explanations will follow the answers.
5. P-values are assumed to be significant at 0.05.
6. Present results and code together. If the output/export of the code/results is a problem screenshots are good within a word document or pdf.
7. There is an 'in class' component listed at the end.

---

# Question 1

A dataset from US store sales is taken from kaggle (https://www.kaggle.com/datasets/whenaman-codes/blood-transfusion-dataset) and is in a file called 'transfusion.csv'.

a. Load the dataset. Separate 70% of the rows to be training data and the rest to be testing data. (1 points)

b. Use the training data to fit a decision tree model where the dependent is the last column variable 'whether he/she donated blood in March 2007'. Then produce a plot of the fitted tree. Afterwards find the accuracy of the predictions on the testing data. (2 points)

c. Use the training data to fit a random forests model where the dependent is the last column variable 'whether he/she donated blood in March 2007'. Then produce a plot of the variable importance. Afterwards find the accuracy of the predictions on the testing data. (2 points)

d. Use the training data to fit a XGBoost model where the dependent is the last column variable 'whether he/she donated blood in March 2007'. Afterwards find the accuracy of the predictions on the testing data. (2 points)

e. Given the previous results, how do you gauge and compare the performances of the various models? (mention the accuracies) (1 points)

(Total: 8 points)

# Question 2

A dataset from credit card customers is taken from kaggle (https://www.kaggle.com/datasets/vrec99/life-expectancy-2000-2015) and is in a file called 'LifeExpectancy.csv'.

a. Load the dataset. Produce a training and testing dataset on a 70-30 split (randomized selections) (1 points)

b.

- Produce a decision tree model to predict the life expectancy, excluding the first 4 columns (Country/Year/Continent/Least Developed), using the training data, and find the MSE on the testing data.
- Produce a random forest model to predict the life expectancy excluding the first 4 columns (Country/Year/Continent/Least Developed), using the training data, and find the MSE on the testing data.
- Produce an XGBoost model to predict the life expectancy excluding the first 4 columns, (Country/Year/Continent/Least Developed), using the training data, and find the MSE on the testing data.
- Discuss the merits of each model based upon the MSE. Use your models to then discuss the most important variables for the predictions.

(4 points)

c. Place the data (LifeExpectancy.csv), into an SQL data base. (1 points)

d. Extract all the life expectancy values, from the DB where the year is greater than 2008, and then less than 2008 into 2 separate variables. Then produce a histogram to show the distribution of the values from the 2 variables. Afterwards conduct a t-test to assess whether they have a significant mean difference or not. (2 points)

e. Using SQL, find the average life expectancy for each country. (use Group By) (1 points)

(Total: 9 points)

# Question 3

On the designated time of the exam as stated on the official schedule you will present in person to me this presentation.

a. Produce a slide presentation of 4 slides with this content.

- Slide1 is your name and student id
- Slide2 is the main results obtained for Q1
- Slide3 is the main results obtained for Q2
- Slide4 the software issues you ran into during your assignment

(3 points)

(Total: 3 points)