**Problem Set 2**

**Due February 10, 2025**

**• For the entire Problem Set 2, you need to save and submit your work in Canvas as a single .R file or a python script.**

**• Important: Use comments throughout your code to:**

**- Indicate which question you are answering**

**Example: #Question 1A, #Question 1B, etc**

**- Annotate your work**

**Examples: #Calculate the mean for ‘height’**

**#P(being on time) = 1/6**

**#Apply binomial distribution**

**\*\*\*If you do not provide annotations, we will not be able to give you full credit for solving the problems.**

**• If you have any questions about what is being asked or what you need to do in order to solve the problems, please reach out to a TA or instructor as soon as possible. If it is necessary to provide additional information or corrections, updated information will be posted on the PS2 Assignment page in Canvas.**

**1.** For Problem 1 questions, use the dataset for “Risk Factors Associated with Low Infant Birth Weight.” You should use R or Python to examine the dataset and answer the following questions. (**30pts)**

Additional information about the dataset:

• Source: Hosmer, D.W. and Lemeshow, S. (1989) \_Applied Logistic Regression.\_ New York: Wiley

• This data frame contains the following columns:

‘low’ indicator of birth weight less than 2.5 kg.

‘age’ mother's age in years.

‘lwt’ mother's weight in pounds at last menstrual period.

‘race’ mother's race (‘1’ = white, ‘2’ = black, ‘3’ = other).

‘smoke’ smoking status during pregnancy. 1 = yes, 0 = no

‘ptl’ number of previous premature labours.

‘ht’ history of hypertension.

‘ui’ presence of uterine irritability.

‘ftv’ number of physician visits during the first trimester.

‘bwt’ birth weight in grams.

5 pts

A. How many observations are in the dataset?

5 pts

B. Examine each column(variable), determine what type of variable each represents, and indicate whether each one is discrete or continuous. Then, go on to determine the following distribution or descriptive statistics as appropriate:

For Discrete Variables:

• Indicate whether the feature is nominal, ordinal, or binary

• How many levels each variable has

For Continuous variables:

• Determine the mean, standard deviation, and median

5 pts

C. How many individuals older than 30 smoke?

5 pts

D. Plot a histogram for birth weight.

5 pts

G. Calculate the probability of randomly selecting an individual that has either a low birth weight or a mother who was a smoker.

5 pts

H. Calculate the probability of randomly selecting an individual that is white and has more than 3 physician visits during the first trimester.

**2.** Answer the following probability questions using R or Python. (**30pts)**

A. What is the probability that given a positive mammogram exam, a woman has a positive cancer diagnosis? Assume that the breast cancer incidence rate is 1%, the positivity rate for the exam if a patient has cancer is 90%, and there is a false positive rate of 8% for the exam. (10 pts)

B. For every attempt to call your friend, there is a 70% probability of actually speaking with them. Calculate the probability of having exactly 12 successes in 20 attempts. (10 pts)

C. The cholesterol levels of a group of patients are normally distributed with a mean of 200 mg/dL and a standard deviation of 25 mg/dL.

1. What is the probability that a randomly selected patient will have a cholesterol level between 180 mg/dL and 220 mg/dL?
2. Additionally, calculate the interquartile range (IQR) of the cholesterol levels.
3. Discuss how these statistics can be used to identify patients at risk of cardiovascular diseases
4. How the distribution might change if the standard deviation were reduced to 15 mg/dL? (10 pts)

**3.** Naïve Bayes classifier

The dataset consists of measurements taken from breast masses using characteristics of cell nuclei extracted from digitized images. Your goal is to build a **Naïve Bayes classifier** to predict whether a tumor is **benign or malignant**. The dataset should be available on canvas (**20pts)**

**Part 1: Data Preparation and Exploration (5 pts)**

a) Load the dataset in your preferred language and perform an initial exploration. What are the dimensions of the dataset? Are there any missing values? (2 pts)  
b) Summarize the key characteristics of the dataset using appropriate descriptive statistics and visualizations. What are the distributions of numerical features? (2 pts)

**Part 2: Naïve Bayes Classifier Implementation (15 pts)**

1. Preprocess the data for Naïve Bayes classification. Make sure to encode categorical variables (if any) and split the data into training and testing sets (70-30 split). Show your code and explain your preprocessing steps. (2 pts)
2. Train a **Naïve Bayes classifier.** Which are the features with the highest contribution to the classifier? Report the accuracy, sensitivity, and specificity of your model. (3 pts)
3. Create and interpret the **confusion matrix**. What does it tell you about the model's performance? Are there any potential concerns regarding false positives or false negatives? (5 pts)
4. Plot the **ROC curve** for your Naïve Bayes model and compute the **AUC (Area Under the Curve)**. What does the AUC tell you about the classifier's ability to distinguish between benign and malignant cases? (5 pts)