**Name-Aman Gupta**

**##cwid-20018346**

**Instructions:**

**You can use Word, Excel, Power Point, R and/or Python to answer the questions in this exam. There are a total of seven (7) multi-part questions, with point values noted for each question. You must use Excel if it is specified in the problem statement.**

**Please show your calculations, or the details of your program(s) for each problem. You must supply the R/Python programs, and the programs should be commented so that each step is clearly explained.**

**Combine all of your answers/files into a single zipped file and post the zipped file to CANVAS.**

**#1** (10 Points)

**Measure the distance between (0, 0, 0) and (1, 1, 1) using the following distance formula. Is the following function a proper distance function? Why? Explain your answer.**

**Ans :- x=(0,0,0) and y =(1,1,1)**

**d(x,y)=(0-1)^3 +(0-1)^3 +(0-1)^3=-3**

**the function is a proper distance function when**

**d(x,y)>=0 and here the value of d(x,y) is -3**

**So, d is not a distance function.**

**#2** (15 Points)

**Load the “CS513\_targeting\_num.csv” dataset, from the raw\_data module in CANVAS, into R/Python. This is a fictional customer targeting dataset. Perform the EDA analysis by:**

1. **Summarizing each numerical column (e.g., min, max, mean)**
2. **Identifying missing values**
3. **Replacing the numerical missing values with the “median” of the corresponding columns**
4. **Displaying the scatter plot of “Age”, and “Income”**
5. **Show the box plots for columns: “Age” and “Income”**

**Ans:-**

**In python file.**

**#3** (15 Points)

**Use EXCEL and the "CS513\_targeting\_num\_sml2.csv" file, which contains a different version of fictional customer targeting data for this problem. Employ the weighted k-nearest neighbors (k-NN) algorithm with k=3 in Excel to impute the missing "Income" values for customers whose monthly income ranges between 0 and 25,000 and whose age is between 20 and 70 years old.**

**Ans:-In excel sheet**

**#4** (15 Points)

**Load the CANVAS “CS513\_targeting\_cat\_full.CSV” dataset into R/Python. Remove the missing values. Construct a Naïve Bayes model to classify “Purchase” (Purchase =’yes’). Predict purchase for a random sample (30%) of the data (test dataset). Measure the following: accuracy, precision, specificity, recall and F1.**

**Ans:- In python file**

**#5** (15 Points)

**Use EXCEL and the "CS513\_targeting\_cat\_sml2.csv" file, which features another version of fictional customer targeting data for this question. Apply the Bayesian and Naïve Bayes algorithms to predict/classify whether a customer in the north who is presented with an offer for product “B” will buy the product or not.**

**Ans:- In excel sheet**

**#6 (**15 Points)

**Load the CANVAS “CS513\_targeting\_num.CSV” dataset into R/Python. Remove the missing values. Construct an unweighted knn (k=3) model to classify “Purchase” (Purchase =’Yes’). Predict the purchase for a random sample (30%) of the data (test dataset). Measure the following metrics: accuracy, precision, specificity, recall and F1.**

**Ans:- In python file**

**#7** (15 Points)

**A COVID test was administered to 1,000,000 individuals. The test correctly identified 95% of those who were sick (P[positive/sick] = 0.95) but also produced a positive result for 10% of those who were not sick (P[positive/not sick] = 0.10). If the prevalence of COVID in this population is 20%, what is the probability that an individual who tested positive is actually sick? What is the probability that an individual who tested negative is actually sick?**

**Ans:-**

**Probability of prevalence of COVID is P(S)= 0.20**

P[¬S] = 1 - P[S] = 0.80.

**Probability of true positive of covid is P(+|S)= 0.95**

P[-|S] = 1 - P[+|S] = 0.05

**Probability of False positive of covid is P(+|-S)= 0.10**

P[-|¬S] = 1 - P[+|¬S] = 0.90.

P[+] = P[+|S] \* P[S] + P[+|¬S] \* P[¬S] =0.95 \* 0.20 + 0.10 \* 0.80 = 0.23

P[-] = P[-|S] \* P[S] + P[-|¬S] \* P[¬S]= 0.05 \* 0.20 + 0.90 \* 0.80 = 0.77

**the probability that an individual who tested positive is actually sick =P(S|+)=** P[+|S] \* P[S] / P[+] (Bayes' theorem)

= 0.95 \* 0.20 / 0.23 = 0.826

**the probability that an individual who tested negative is actually sick=P(S|-)=**

P[-|S] \* P[S] / P[-]. (Bayes' theorem)

=0.05 \* 0.20 / 0.77 = 0.013

the probability that an individual who tested positive is actually sick is 0.826 or 82.6%, while the probability that an individual who tested negative is actually sick is only 0.013 or 1.3%

**(Show your calculations)**

**Datasets: CS513\_targeting\_num.csv, CS513\_targeting\_num\_sml2.csv, CS513\_targeting\_cat\_full.CSV, CS513\_targeting\_cat\_sml2**