**Instructions:**

**You can use Word, Excel, Power Point, R and/or Python to answer the questions in this exam. There are a total of Five (5) 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 your answers/files into a single zipped file and post the zipped file to CANVAS.**

**#1** (10 Points)

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

**#2** (15 Points)

**Load the “Admission\_v2\_missing” dataset, from the raw\_data module in CANVAS, into R/Python. This is a fictional undergraduate admission data set (see the description below). Perform the EDA analysis by:**

1. **Summarizing each numeric column (e.g., min, max, mean).**
2. **Identifying missing values.**
3. **Replacing the numerical missing values with the “mean” of the corresponding column.**
4. **Displaying the scatter plot of “GRE”, “GPA” and “Rank”, one pair at a time.**
5. **Showing box plots for columns: “GRE”, and “GPA” one graph at a time.**

**#3 (25 Points)**

**Use Excel and the “****Admission\_v2\_missing\_small.csv” (Excel file containing another variation of the fictional undergraduate admission dataset) to replace the missing “GRE” score using weighted knn (k=3).**

**#4** (25 Points)

**Load the CANVAS “Admission\_v2.CSV” dataset into R/Python. Remove the missing values if necessary. Discretize the “GRE” scores into “up to 500 inclusive”, “above 500 and up to 600”, “above 600 and up to 700” and “above 700”. Also discretize the GPAs into “up to 2.5 inclusive”, “above 2.5 and up to 3”, “above 3 and up to 3.5”, and “over 3.5”. Construct a Naïve Bayes model to classify admission (admission=’yes’) based on the other variables. Predict admission for a random sample (30%) of the data (test dataset). Measure the accuracy of the model. use (30%) of the data as test dataset**

**#5 (**25 Points)

**Load the CANVAS “Admission\_v2.CSV” dataset into R/Python. Remove the missing values. Construct an unweighted knn = (3,5,7) model to classify admission (Admit=’YES’) based on the other variables. Predict admission for a random sample (30%) of the data (test dataset). Measure the accuracy of the model for each K.**

**Admission Data Set Description**

**Applicant: Applicant ID.**

**Admit: if applicant was admitted (YES, NO).**

**GRE: GRE score, ranging from 200 to 800.**

**GPA: GPA score, ranging from 0 to 4.**

**Rank: Rank of applicant’s undergrad school, ranging from 0 to 4.**

**Datasets:** **Admission\_v2\_missing ,Admission\_v2\_missing\_small.csv, Admission\_v2.CSV**