# IS603 Final EXAM

Name: Rohan Salvi

ID number: VH10935

**MANDATORY ACADEMIC INTEGRITY AGREEMENT:**
**I will not discuss the exam with anyone except Dr. Quader until after it finishes at 7:00 pm ET on 4/5/2022. All of my work is my own and is not copied from other sources (even if paraphrased). I have acknowledged and cited all sources. I have not committed any act of cheating, fabrication, or plagiarism, and I have not helped others to do so. I have not used software to attempt to hide plagiarism. I will delete my copy of the exam afterward.**

**I understand that there is a zero-tolerance policy on plagiarism, and any violations will result in a 0 grade on the exam. I understand that academic integrity violations may incur additional consequences such as a failure grade in the entire course and further disciplinary action.**

Signature, certifying that I agree to the above statement: Rohan Salvi

Today’s date: 20/5/2022

**Your exam submission will not be accepted without a signature for the above.** You may attach a signed copy of this sheet or copy the complete statement to another document and sign it there. This exam is open book. You may consult other sources, including the textbook, the course slides, homework assignments and the internet, but you must follow the policies in the academic integrity agreement. For definitions of academic integrity terms including cheating, fabrication, plagiarism, and dishonesty, see `https://www.umbc.edu/policies/pdfs/UMBC%20III%201.10.03%20Undergraduate%20Student%20Academic%20Conduct%20Policy.pdf` .

Again, **you may not discuss the exam with anyone.**

1)

a. Given that A= [2,0,0] B =[0,1,3] C= [0,1,2], D= [3,0,1], E= [1,0,1]

   i. The Euclidean distance between d(A,E) = sqrt of ((2 - 1) * 2 + (0 - 0) * 2 + (0 − 1) * 2 = 1.414
   ii. The Euclidean distance between d(B,E) = sqrt of ((0 − 1) * 2 + (1 − 0) * 2 + (3 - 1) * 2) = 2.449
   iii. The Euclidean distance between d(C,E) = sqrt of ((0 - 1) * 2 + (1 - 0) * 2 + (2 - 1) * 2) = 1.732
   iv. The Euclidean distance between d(D,E) = sqrt of ((3 - 1) * 2 + (0 - 0) * 2 + (1 − 1) * 2) = 2

b. Customer A is Customer E's nearest neighbor. Customer E's predicted class label is Plan A.

c. A, C, and D are Customer E's three closest neighbors. Given K=3,. Plan B is the A class label, whereas Plan A is the C and D class labels. Plan B is E's class label.

d. When K or the number of neighbors is too low, the model selects only the values that are closest to the data sample, resulting in the very complicated decision boundary illustrated above. This model does not generalize well on the test data set, resulting in unsatisfactory results. The model begins to generalize better as the number of neighbors is larger, however increasing the value too much reduces performance.

e. We may choose the K with the lowest mistake rate. Begin calculating with a random K value as the beginning point. The decision limits become unstable when K is set to a low value. The smoother the decision boundaries get as the K value increases, the better for categorization. Plot the graph with K indicating values within a specific range and the error rate. After that, choose the K value that has the lowest mistake rate.

2)

a. Managers are close to the company's stake holders/ or in other words we can say that they know what they have to do with the data and how to use it for the company's usage. Managers should know what expertise his team of data scientists are in, so that he can assign them proper data tasks. On other side data scientists (manager's team) should know manager will show/tell them where and how the data should be dealt with. In order to generate and process appropriate models, they will need a detailed grasp of the object on which they are going to work.

b. 1st question:
We need to know the profit rate and how much profit would be returned if this idea is implemented. We should know if the data mining solution will bring benefits before, we proceed with the task.

2nd Question:
Is it the most efficient way? or are there any other ways to deal with this problem. i.e., why only this method why not any other method of data mining did we choose to move forward with.

3rd Question:
The second question concerns user data privacy information. Have the end users been notified about the data collection?

c. I think normal K- means clustering would work. Since we can attribute the plan, age, location, or occupation as a factor. This will enable them to provide tailored suggestions to a certain consumer segment.

d. It is unsupervised learning, since the data is not labeled, we will categorize it as an unsupervised learning technique.

e. The reason for choosing this method because it is easy to implement k-means and identify unknown groups of data from complex data sets. The results are presented in an easy and simple manner, it is flexible ie K-means algorithm can easily adjust to the changes. If there are any problems, adjusting the cluster segment will allow changes to easily occur on the algorithm, also suitable for a large dataset, segmentation is linear in the number of data objects thus increasing execution time. Also compared to using other clustering methods, a k-means clustering technique is fast and efficient in terms of its computational cost

O(K*n*d).

f. Considering the situation that we have now performed feature extraction from the data collected from users. We now filter out the people with what age Sim plan, occupation and location or any other factor have opted out of Ez Mobile. This makes us clear that what kind of people are leaving EZ mobile and going for Horizon. Now we can also predict depending on these factors what segment of people are changing their minds to switch to Horizon. Then we offer a new plan, or some new discounts offer or Loyalty points to retain this customer.

3.

a. Business Understanding

Let's understand this business, this VR travelling won't be available at every location, since this company recently started, the locations that they provide will be limited. The startup has to offer a trial period of experience, but before the trial period starts, the company takes bunch of information from user and his interests regarding which location to he would visit. The data scientist's business aim in this project is to detect the consumer's attitude and offer areas for them to visit and investigate. Now after the trial period ends, based on the trial experience feedback and information taken in the start we can recommend and lure customers to travel their favorite locations on VR.

b. Data Understanding

The data collected will be very vast and very exciting, the VR will collect visual VR video data, sound data, also location of the user in VR, most visited locations. Then we also have feedback from the trial experience. To make better selections, we must choose characteristics that are appropriate for the analysis while keeping the model simple and general. We can also identify the most common tourist search pattern, allowing us to design a destination based on visitor behavior.

c. Data preparations

The data that we will have that needs to be thoroughly cleaned. There were many people, who just filled the trial pack with fake details, that must be removed. Even though some information was fake but the human behavior in the VR would be very helpful, so that such kind of people can be avoided, or the company can bring in something that interests them and, in the end, also lure these people into buying the travel pack. Signal processing also needs to be done. Video data collected will be huge , along with location and other details . Keeping this file in folders or zip is not manageable. I would suggest using HDF5 in this case.

d. Modeling

We need some strong models to handle both the image and other forms of data because we are using both video(image) and numeric data. The training dataset, which is a statistical representation of the data collected across the time for which the result is decided, will be used to train the model. We may calculate a score based on how likely people are to spend time at a particular location in VR and compare it to the real-world data.

f. Evaluation

The training dataset will be split into 3 parts: Training, test and validation. The training set is a subset of the dataset that is used to create prediction models. The validation set is a subset of the dataset used to evaluate the model's performance throughout the training phase. It serves as a testing environment for fine-tuning model parameters and choosing the most effective model. A validation set isn't required for all modeling techniques. To evaluate the project, we will use the anticipated value framework. We'll achieve this by first building a confusion matrix based on the model's prediction vs. actual data, and then turning the results into probabilities.

g. Deployment

This step involves integrating a machine learning model into an existing production environment to make data-driven business choices. The model will be incorporated in a virtual reality computer design, with each correct prediction helping to improve usability and provide a more realistic travel experience for the user. The primary goal of this data generated will be to better understand user cognition and behavior in a virtual reality environment. This result should also be used to give user recommendation of next trip or destination he would like to visit.

4)

a)

The main issue that might arrive is the live location of the bot in real world. This might be security concern as the travel-bot might be stolen or damaged by people. There should be some limitation set to the bot that it cannot go beyond a point or location that it feels safe. Also, Customers must be informed that their data will be used for our business and must consent to our request to gather it.

b)  The location of bot should be restricted, but this limits the user experience. Location of the travel bot will be still a security issue; the bot is visually visible and can be tampered by anyone. The best solution to this would making a complete 3D environment based on visuals from VR data or camera. Or making contracts with a 3rd party company to maintain the travel-bot. Customers should be made sign data usage policy company policies ,letting them know their data will be used and with whom will it be shared and how will It be used .