**STUDENT’S DECLARATION OF ORIGINALITY**

By submitting this online assessment, I declare that this submitted work is free from all forms of plagiarism and for all intents and purposes is my own properly derived work. I understand that I have to bear the consequences if I fail to do so.

Final Online Assessment Submission

| Course Code: | AACS2383 |
| --- | --- |
| Course Title: | Introduction to Data Mining |
| Signature |  |
| Name of Student | My name is PYQ 20202105 |
| Student ID |  |
| Programme of Study |  |
| Date | 22 September 2021 |

# 

# **Mark Summary**

# 

| **Question 1** |  |
| --- | --- |
| **Question 2** |  |
| **TOTAL** |  |

# 

# 

# 

# 

# 

**AACS2383 - Final Online Assessment**

# **QUESTION 1 a) (10 marks)**

| i)  Accuracy refers to the degree to which information accurates reflects an event or object. For example, if a customer’s age is 30, but the system says she is 24, that information is inaccurate.  Completeness refers to whether all required information is present in the dataset. For example, if the customer information in a database required both first and last names, any record without one of the attributes is marked as incomplete.  Consistency refers to all representations of that item across data stores. For example, a birthdate is entered in a system using U.S format which is mm/dd//yyyy. Hence, any date entered using European standaard (dd/mm/yyyy) causes inconsistency.  ii)  Timeliness. The data should be updated timely so that the data is correct when it is needed.  Validity. The data should follow a specific format of business rules. |  |
| --- | --- |

# **QUESTION 1 b) (8 marks)**

| 3 methods to handle missing attributes to proceed with the data for further analysis are data preprocessing, data transformation and data reduction. Data preprocessing included data cleaning which replaced the missing value with mode value, ignored the tuple, binning method, regression and clustering to handle the noisy data. Next, the data transformation was done to normalize the data and become measurable and able to proceed data analysis. By transforming the data, concept hierarchy can be applied. After data transformation, the next step to handle missing attributes is data reduction by using dimensionality reduction. As a result, the outcome of an attribute will be easier to categorise and match with other attributes.  always data preprocessing only  replaced the missing value with mode value  ignored the tuple  bining convert discrete data to dediscretized data  how to replace missing value?  -tell how this method work  -example  mean  median  mode  for example, age we take mean value of age and replace all missing value with mean  average income of customer, we take missing value of income to replace with all this missing value  ignored tuple  we can manually filled in the missing value, although it is not recommended  fill in the missing value by global constant  we can find the missing value and replace it by Unknown  we can use certain machine learning algorithm such as linear regression to predict a value on the regression line. Naive bayes decision tree to predict what is going / the most likely value to replace with missing value |  |
| --- | --- |

# **QUESTION 1 c) (3 marks)**

| 22,12,61,57,30,1,32,37,37,68,42,11,25,7,8,16  Rearrange:1,7,8,11,12,16,22,25,30,32,37,37,42,57,61,68  16 numbers  Equal depth:16/4=4  Bin 1:1,7,8,11  Bin 2:12,16,22,25  Bin 3:30,32,37,37  Bin 4:42,57,61,68  Equal width:(68-1)/4=16.75:approximately 17  Bin 1(1-18):1,7,8,11,12,16  Bin 2(18-35):22,25,30,32  Bin 3(35-52):37,37,42  Bin 4(52-69):57,61,68 |  |
| --- | --- |

# 

# **QUESTION 1 d) (10 marks)**

| - Key write at first row  - Relationship arrow point to the key |  |
| --- | --- |

# **QUESTION 1 e) (10 marks)**

| | Schema | Star Schema | Snowflake Schema | | --- | --- | --- | | Similarity | A central fact table surrounded by dimension tables | | | Differences | The dimension tables are denormalized. | The dimension tables are normalized. | | Benefit | Queries are simpler. All the data will connect through the fact table, then the multiple dimension tables are viewed as a large table of information. | Less disk space used. Because data is normalized and there is a minimal data redundancy. | | Drawback | Decreased data integrity. Due to the denormalized data structure, star schemas do not enforce data integrity very well. | Maintenance can be complex. It is due to there being a large number of different tables in the data warehouse. |   I think snowflake schema might be more empirically beneficial if we compare which schema is better for maintainability. It is because maintainability for a data warehouse is heavily dependent on the amount of redundant data. The more redundancies of data, the more maintenance needs to take place. Comparing these two schemas, snowflake schema has the least data redundancies. |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

# 

# 

# **QUESTION 2 a) (32 marks)**

# 

| Naive Bayes is a linear classifier while K-NN is not; It tends to be faster when applied to big data.In comparison, k-nn is usually slower for large amounts of data, because of the calculations required for each new step in the process. If speed is important, choose Naive Bayes over K-NN.  In general, Naive Bayes is highly accurate when applied to big data. Don't discount K-NN when it comes to accuracy though; as the value of k in K-NN increases, the error rate decreases until it reaches that of the ideal Bayes. |  |
| --- | --- |

# **QUESTION 2 b) (4 marks)**

| (i)Root = class  -It is because class attributes symbolize the decision of two categories which are Oak tree and Pine tree.  -class attribute is the data label of this data set analysis  -Other attributes cannot become root node is because they symbolize intervention and outcome of class attributes.  (ii) decision tree      (iii) 1. Pine  2. Oak |  |
| --- | --- |

# **QUESTION 2 c) (14 marks)**

| 1. There will be more outliers 2. Some clusters that exist when using the first parameter setting will be deleted or split into several smaller sub-clusters. Therefore the number of clusters could increase or decrease |  |
| --- | --- |

# **QUESTION 2 d) (14 marks)**

| | Extrinsic | Intrinsic | | --- | --- | | The ground truth is available. | The ground truth is unavailable. | | Compare clustering against the ground truth using certain clustering quality measures. | Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are | | Example: BCubed precision and recall metrics. | Example: Silhouette coefficient. | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

# **QUESTION 2 e) (14 marks)**

| i)50%  ii)0%  iii)20% because for each fold, only 1 of the 5 test examples is classified correctly, so the average accuracy on the two folds is 20%. |  |
| --- | --- |