QUEST INTERNATIONAL UNIVERSITY
PERAK

FACULTY OF SCIENCE AND TECHNOLOGY

School of Computing

Bachelor in Computer Science (Hons)

Study Intake: 202011
FINAL ASSESSMENT
Submission Date: 19 March 2021

BCS3222 Data Mining and Data Warehousing

General instructions to candidates: **TOTAL MARKS: 100**

1. Please check that this assessment consists of **EIGHT (8)** pages of printed material including the cover page before you begin the assignment.

2. Answer **ALL** questions.

3. Use Turnitin's plagiarism check and ensure similarity detected is under 30%.

4. **NO** marks will be awarded if any part of the assignment is found to be copied directly from any materials or from another student.

5. Complete your personal and course details in a **COVER PAGE** as specified by your course instructor and attach it with your assignment report.

6. Submit your program's source codes/scripts and any related files for this assignment to the course Google Driver folder provided by your course instructor.

7. **Late submission will not be accepted.**

General Guidelines

Academic Honours Code applies to this assignment. Plagiarised work would not be marked. Refer to Academic Honours Code at http://goo.gl/k1HROA

The use of any third-party libraries and/or frameworks is strictly prohibited.

Excessive use of inappropriate comments, line breaks or unused variables in the source code will be penalised. 'Spaghetti' code is not wanted.

A detailed illustration of flowcharts and concise description of functions description are desirable.

The Design Document/Report must be concise in headings, with consistent alignment, clear description and illustration used. No extra mark will be given for superb graphic design

Marks will be deducted for wrong spelling and grammar used in the User Manual. Use UK English, i.e. 'colour' instead of 'color'. If acronyms are used, all acronyms must be defined in Glossary of Terms.

To use Turnitin, go to http://www.turnitin.com/en_us/home and log in using your own credentials. Please ensure that similarity detected is fewer than 30%.

Turnitin Class Name: BCS3222/BAS3163 Data Mining and Data Warehousing (Nov2020)

Turnitin Class ID: 27356444

Turnitin Enrollment Key: DMDW202011

Question (100 marks)

Classification, which is also known as supervised machine learning, is one of the data mining techniques that extracts the patterns from a collection of data, and build models describing important data classes. It is a two-stage process, consisting of a learning/training stage where a classification model is constructed, and a classification stage where the model is used to predict class labels for given data. There exist various types of classification techniques, such as decision tree, naïve Bayesian, decision rules, artificial neural networks, Bayesian belief networks, support vector machines and k-nearest neighbours. Among these techniques, decision tree and naïve Bayesian have been reported as the popular techniques used in many classification applications.

Decision tree has been well-known with its capability to force the consideration of all possible outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis. Furthermore, its high interpretability has contributed to its popularity in many applications of classification.

Naive Bayesian is a statistical classifier that predicts class membership probabilities. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. Furthermore, its high accuracy and speed when applied to large databases has contributed to its popularity in many applications of classification.

The objective of this assignment is to develop a classification program using console application programming language, such as C, C++, C# and Java. The program should consist of TWO modules, which use the C4.5 algorithm (one of the most popular algorithms of decision tree) and naive Bayesian technique respectively to build a classification model based on a training set (defined in text file). The trained trained models are then evaluated by using a testing set (defined in text file). The contents and the classification results of each of the testing data by the trained models should be displayed to the terminal screen. The metrics for evaluating the performance of each of the trained classifiers based on the testing set should then be displayed to the terminal screen as well.

The deliverables of this assignment are as follows:

1. **The executable file (.exe) of the program.**
2. **The source codes of the program.**
3. **The Deployment & Installation Manual, and the User Manual of the program.**
4. **A text file (.txt) which consists of the training set*.**
5. **A text file (.txt) which consists of the testing set*.**
6. **The technical report (.pdf) of the assignment.**
7. **The presentation video of the demonstration and operation of the program.**

***** Create the training set file to construct the classification model, and the testing set file to evaluate the performance of the trained model. Both files are suggested to be in the format of .txt or .csv. The program must be flexible to accept any number of data tuples, attributes and classes/labels in the selected dataset.

The format of training set and testing set is defined as follows:

The first row defines of the list of attributes and class label of the data tuples. The following rows define the data tuples at a finite count. The attributes and class label of the data tuple are separated by semicolon (;) and the data tuples are separated by new line. The template is as shown below.

No ; Attribute_1 ; Attribute_2 ; Attribute_3 ; ..... ; Attribute_N ; Class_Label

1 ; Data1_Value1 ; Data1_Value2 ; Data1_Value3 ; ..... ; Data1_ValueN ; Data1_Label
2 ; Data2_Value1 ; Data2_Value2 ; Data2_Value3 ; ..... ; Data2_ValueN ; Data2_Label
3 ; Data3_Value1 ; Data3_Value2 ; Data3_Value3 ; ..... ; Data3_ValueN ; Data3_Label
4 ; Data4_Value1 ; Data4_Value2 ; Data4_Value3 ; ..... ; Data4_ValueN ; Data4_Label
5 ; Data5_Value1 ; Data5_Value2 ; Data5_Value3 ; ..... ; Data5_ValueN ; Data5_Label
6 ; Data6_Value1 ; Data6_Value2 ; Data6_Value3 ; ..... ; Data6_ValueN ; Data6_Label
.
.
.
.
.
.
M; DataM_Value1 ; DataM_Value2 ; DataM_Value3 ; ..... ; DataM_ValueN ; DataM_Label

THE ASSIGNMENT COMPRISES TWO(2) PARTS, AS FOLLOWS:

Part A: Classification Program [80 marks]

Module 1: C4.5 Classification Module [40 marks]

The program should allow the user to enter a file name which consists of the training set, to load the training set to the program. The program will then develop a decision tree model using C4.5 algorithm and the training set. The program should illustrate each of the stages of developing the decision tree model, by displaying all the measures for every iteration in building the decision tree model.

The measures for C4.5 are defined as follows:

1. $\boldsymbol{Info(D)}$ of the training set.
2. $\boldsymbol{Info_A(D)}$ of each of the attributes in the training set.
3. Information gain $\boldsymbol{Gain(A)}$ of each of the attributes in the training set.
4. Split information $\boldsymbol{SplitInfo_A(D)}$ of each of the attributes in the training set.
5. Gain ratio $\boldsymbol{GainRatio(A)}$ of each of the attributes in the training set.

The program should then illustrate the resulting decision tree model in an appropriate tree structure to the terminal screen.

After the decision tree model has been built, the program should enable the classifying module which allows the user to enter a file name which consists of the testing set, and load the testing set to the program, to execute the classification operation. The classification results should then be illustrated to the terminal screen, by displaying every testing data with its resulting label.

The confusion matrix and the metrics for evaluating the performance of the classifier should then be displayed to the terminal screen as well. The common evaluation measures of classifier performance are defined as follows:

1. Accuracy or Recognition Rate
2. Error Rate or Misclassification Rate
3. Sensitivity (Recall) or True Positive Rate
4. Specificity or True Negative Rate
5. Precision
6. F, F_1, *F-score* or Harmonic Mean of Precision and Recall

Module 2: Naive Bayesian Classification Module [40 marks]

The program should allow the user to enter a file name which consists of the training set, and load the training set to the program. The program will then construct the naïve Bayesian model based on the training set. The program should illustrate computational process in developing the naïve Bayesian model, by displaying the value of all the probability notations for the trained model to the terminal screen.

The probability notations are defined as follows:

Let C_i be the category, and x_i be the attribute value of the attribute vectors which form the data tuples,

1. $P(C_i)$, the prior probability of each of the categories.
2. $P(x_i \mid C_i)$, the posterior probability of each of the attribute values conditioned on each of the categories.

After the classification model has been built, the program should enable the classifying module which allows the user to enter a file name which consists of the testing set, and load the testing set to the program, to execute the classification operation. The program should illustrate computational process in classifying the testing data tuples, by displaying the value of all the probability notations for each of the data tuples in the testing set and their resulting label to the terminal screen. The program should include the functionality for Laplacian correction to avoid computing probability values of zero.

The probability notations are defined as follows:

Let C_i be the category, X_i be the data tuple, and x_i be the attribute of the data.

1. $P(X_i \mid C_i)$, the posterior probability of each of the data tuples conditioned on each of the categories, by taking all the relevant $P(x_i \mid C_i)$ into calculation.
2. $P(X_i \mid C_i).P(C_i)$ of each of the data tuples for each category, which is used to determine the class label of a particular data tuple, based on Bayes Decision Rule.

The confusion matrix and the metrics for evaluating the performance of the trained classifier based on the testing set should then be displayed to the terminal screen as well. The common evaluation measures of classifier performance are defined as follows:

1. Accuracy or Recognition Rate
2. Error Rate or Misclassification Rate
3. Sensitivity (Recall) or True Positive Rate
4. Specificity or True Negative Rate
5. Precision
6. F, F_1, *F-score* or Harmonic Mean of Precision and Recall.

Note: The source codes, the training set file, the testing set file, and other relevant files of the program must be submitted to the course folder in Google Drive.

Part B: Technical Report [20 marks]

Write a report for the assignment, which comprises the sections as below:

Chapter 1: Introduction

- Introduction
- Problem Statement
- Objectives
- Scope

Chapter 2: Literature Review

- Overview of Machine Learning concept and techniques.
- Literature on Decision Tree and its variants (eg. ID3, C4.5 and CART).
- Literature on Bayes Theorem and naïve Bayesian classification.

Chapter 3: Methodology

- Detailed discussion on the selected methodology in relation to the design and development processes of the program.

Chapter 4: Implementation, Results Analysis and Discussion

- Detailed description and explanation for each of the stages in the classification operations.
- Detailed presentation, analysis and discussion of the evaluation of the trained classification models.

Chapter 5: Conclusion

- Conclusion
- Strengths and weaknesses of the developed classification program.

Note: The assignment report and the plagiarism report must be submitted to the course folder in Google Drive.

Submission of Deliverables:

All the deliverable materials must be submitted to the Google Drive folder of the course "2020 November – BCS3222_BAS3163 DMDW". The Google Drive folder's link for Final Assessment materials submission is provided as below:

https://drive.google.com/drive/folders/15cnRbxkRq-N-7d8xcs0Ki8ycTtaiX97q?usp=sharing

- The Complete Package of Source Codes and Relevant Files (eg. Training Set and Testing Set) of the classification program, together with the Detailed Deployment and Installation Manual, and the Detailed User Manual must be uploaded to the designated folder of *"2020 November – BCS3222_BAS3163 DMDW" -> "Final Assessment" -> "00 - System Source Codes and Files" -> "MatricNo_Name"*. Folder "*MatricNo_Name"* is to be created by students respectively, and named after their matric no. and name.

- The Technical Report must be uploaded to the designated folder of *"2020 November – BCS3222_BAS3163 DMDW" -> "Final Assessment" -> "01 – Documentation" -> "MatricNo_Name"*. Folder "*MatricNo_Name"* is to be created by students respectively, and named after their matric no. and name.

- The Presentation Video must be uploaded to the designated folder of *"2020 November – BCS3222_BAS3163 DMDW" -> "Final Assessment" -> "02 – Presentation Video" -> "MatricNo_Name"*. Folder "*MatricNo_Name"* is to be created by students respectively, and named after their matric no. and name.

- END OF QUESTION PAPER -