University of Westminster

School of Computer Science & Engineering

6BUIS001W	Business Intelligence – Coursework 2 (2020/21)
Module leader	Dr. V.S. Kontogiannis
Unit	Coursework 2
Weighting:	50%
Qualifying mark	30%
Description	Show evidence of understanding of various Business Intelligence concepts, via the interaction of database, decision trees and visualisation principles in one case study. Performance indicators for choosing the appropriate decision strategy will be performed. Students need to perform queries in a database while they will also need to perform standard OLAP operations for a given case study. These MySQL & OLAP tasks will be performed only via R environment.
Learning Outcomes Covered in this Assignment:	This assignment contributes towards the following Learning Outcomes (LOs): • LO1 analyse data resource architectures, management process for information resource integration and the process of establishing data models in order to build data warehouse that assists management in decision making process; • LO2 reflect data discovery strategies in order to recommend an appropriate method for incorporating business intelligence in industrial environments; • LO5 implement a dynamic and interactive decision support system that applies the concept of knowledge discovery and information retrieval on a large scale business information resource;
Handed Out:	06/11/2020
Due Date	04/01/2021, Submission by 13:00
Expected deliverables	Submit on Blackboard only a pdf file containing the required details. All implemented codes should be included in your documentation together with the results/analysis/discussion.
Method of Submission:	Electronic submission on BB via a provided link close to the submission time.
Type of Feedback and Due Date:	Feedback will be provided on BB, on 25th January 2021 (15 working days)
BCS CRITERIA MEETING IN THIS ASSIGNMENT	• Problem solving strategies • Knowledge and understanding of mathematical and/or statistical principles • Knowledge and understanding of facts, concepts, principles & theories

Assessment regulations

Refer to section 4 of the "How you study" guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work which obtains a mark in the range 40 – 49%, in which case the mark will be capped at the pass mark (40%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid.

It is recognised that on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board that will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website:**http://www.westminster.ac.uk/study/current-students/resources/academic-regulations**

Instructions for this coursework

During marking period, all coursework assessments will be compared in order to detect possible cases of plagiarism/collusion. For each question, show all the steps of your work (codes/results/discussion). In addition, students need to be informed, that although clarifications for CW questions can be provided during tutorials, coursework work has to be performed outside tutorial sessions.

Coursework Description

1st Question (MySQL/R Queries/Visualization)

While SQL is one of the most widely used programming languages for querying datasets, R is quickly growing the top for analysts and data scientists looking to enrich their statistical repertoires. The rise of data-driven marketing and sales operations has swiftly forayed R into the start-up and enterprise worlds. Owing to its massive suite of libraries and visualizations, boasting over 10,000 packages on CRAN alone, R is just as powerful, if not more, as any leading BI tool when it comes to exploring correlations and causal relationships in your customer datasets. As an IDE, RStudio's GUI is also perfect for querying datasets in syntax not dissimilar to SQL.

A specific dataset (attached file: cars_info.xls) contains related information about a number of cars. The dataset contains 398 observations based on 10 variables. The specific features/characteristics of vehicles in the dataset are: *Mpg (miles per gallon), cylinders, displacement, horsepower, weight, acceleration, model, origin, car_name and price*.

The main objective in this question is to import the provided cars_info.xls dataset into a database (MySQL) and then, via R environment, to request some queries to the database and illustrate the outcome/result of those queries in RStudio environment. You need to explore two different ways of exploring the database via R. The first (classic) methodology is using the functions contained in the RMySQL/DBI packages while the second via the dplyr/dbplyr packages. Finally, you need to import back the contents of the "cars_info table" into R and produce/illustrate some simple visualization of the various parameters included in this dataset.

Specific Tasks:

1. Import the provided cars_info.xls to the database (you may need to save it first as csv file) and inspect what you have loaded via an appropriate command
2. Using the RMySQL/DBI methodology, "ask" via R the following queries and present the appropriate outcome of each query:

- Get the first 10 rows in the imported table
- Get all eight-cylinder cars with miles per gallon greater than 18
- Get the average horsepower and mpg by number of cylinder groups
- Get all cars with less than eight-cylinder and with acceleration from 11 to 13 (inclusive both limits)
- Get the car names and horsepower of the cars with 3 cylinders

3. Explore the dplyr/dbplyr alternative methodology and provide results for all above queries too. You need to read/explore the role of %>% pipe operator.

4. Extract all information from database to R and produce/illustrate the following visualization schematics:

   - Look at the distribution of values for Mpg and cylinders by creating two relevant histograms.
   - Create a boxplot to show the mean and distribution of Mpg measurements for each year in the sample
   - Draw a scatter plot showing the relationship between weight and Mpg. What information do we get from this specific schematic? Justify your answer.

In your main text, you need to include together with your results/discussion the related segments of your R code. At the end of your report, you have to include all of your codes as an appendix.

(Marks 30)

2nd Question (OLAP Operations in R)

At the core of the OLAP concept is an OLAP Cube. The OLAP cube is a data structure optimized for very quick data analysis. The OLAP Cube consists of numeric facts called measures which are categorized by dimensions. OLAP Cube is also called the hypercube. Usually, data operations and analysis are performed using the simple spreadsheet, where data values are arranged in row and column format. This is ideal for two-dimensional data. However, OLAP contains multidimensional data, with data usually obtained from a different and unrelated source. Using a spreadsheet is not an optimal option. The cube can store and analyse multidimensional data in a logical and orderly manner. There are five types of analytical operations in OLAP:

- Roll-up
- Drill-down
- Slice and dice
- Pivot

In this specific question, we have to create a sales fact table that records each sales transaction for an imaginary multi-national company that produces electrical home appliances. This company has branches in five cities: Frankfurt-Germany, Los Angeles-USA, Sydney-Australia, Seoul-S. Korea and Cape Town-South Africa. The produced products are: Washing machine, Fridge, Vacuum Cleaner, Microwave Oven, with indicative prices 200, 500, 400 and 150 USD respectively. The time framework for this sales monitoring is from 2015 to 2020 (inclusive) – i.e. six years. We need to have also information of sales for every month for each of these six years.
Hence, you need initially to create a function, in R, to generate the Sales Table. Use 500 as the indicative number of records, while the transaction data need to be generated randomly. Through R, show the first lines of transactions, in order to verify that you have managed to generate this required sales table. Obviously, the random number of units (per appliance) needs to be an integer type. Then you need to create, again via R, the so-called revenue cube for this company. Finally, you need to utilise the R environment to demonstrate these five OLAP operations for this case study. For each one of these operations, you need to provide/illustrate the related outcomes/results.

In your main text, you need to include together with your results/discussion the related segments of your R code. At the end of your report, you have to include all of your codes as an appendix.

(Marks 20)

3rd Question (Decision Support Systems in BI)

Decision tree learners are powerful classifiers that utilize a tree structure to model the relationships among the features and the potential outcomes. This structure earned its name due to the fact that it mirrors the way a literal tree begins at a wide trunk and splits into narrower and narrower branches as it is followed upward. In much the same way, a decision tree classifier uses a structure of branching decisions that channel examples into a final predicted class value. Decision trees are built using a heuristic called recursive partitioning. This approach is also commonly known as divide-and-conquer because it splits the data into subsets, which are then split repeatedly into even smaller subsets, and so on and so forth until the process stops when the algorithm determines the data within the subsets are sufficiently homogenous, or another stopping criterion has been met. There are numerous implementations of decision trees, but two of the most well-known ones are the C5.0 algorithm and the Classification and regression tree (CART). The C5.0 algorithm has become the industry standard for producing decision trees because it does well for most types of problems directly out of the box. There are various measurements of purity that can be used to identify the best decision tree splitting candidate. C5.0 and CART utilise entropy and gini index respectively as impurity measures for selecting attribute. Both measures however have advantages/disadvantages. The process of pruning a decision tree is an important component in the process, as it involves reducing its size such that it generalizes better to unseen data.

Over a period of time, the number of people affected by diseases has gradually increased. Specifically number of people suffering from liver disease has increased because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. Data was selected from the patient records in area north-east of Andhra Pradesh, India and donated to the University of California Irvine (UCI Machine Learning Repository). The specific dataset (attached file: liver.csv) contains 416 liver patient records and 167 non-liver patient records. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors. The goal is to create a classifier that predicts whether a subject is healthy (non-liver patient) or ill (liver-patient) based on some clinical and demographic features. The specific question examines data from liver patients concentrating on relationships between a key list of liver enzymes, proteins, age and gender using them to try and predict the likeliness of liver disease.

Attribute information:

1. Age: Age of the patient
2. Gender of the patient
3. Total Bilirubin
4. Direct Bilirubin
5. Total Proteins
6. Albumin
7. ag_ratio = Albumin and Globulin Ratio
8. Sgpt: Alamine Aminotransferase
9. Sgot: Aspartate Aminotransferase
10. Alkphos: Alkaline Phosphotase
11. is_patient: Disease State (our target)

Decision trees are widely used in biomedicine due to their high accuracy and ability to formulate a statistical model in plain language. R Tool is an excellent statistical and data mining tool that can handle any volume of structured as well as unstructured data and provide the results in a fast manner and presents the results in both text and graphical manners. This enables the decision maker to make better predictions and analysis of the findings. The dataset consists of 583 observations. In this question, you need to develop two classification models using C5.0 and CART decision trees respectively. The aim is to decide, at the end, which one of these two models is preferable for this specific biomedical task. Both models need also to be optimised (via tuning/pruning like methods) in order to obtain even better results. **You need to perform the following tasks**:

- Import the provided .csv file in your MySQL and check that it has been imported properly. **The following tasks however need to be performed only via R environment**.
- Import the contents of this file from your MySQL into R (using R-related commands) and designate a specific name-variable to store them.
- Explore & prepare the data:
  - The name-variable where you stored the information, contains 583 observations (rows) and 11 features (columns). There are some missing values; therefore you need to remove entirely those affected sample rows. As some of these features are non-numerical in nature, you may consider in transforming them into numerical form, if you think is more convenient to you. The target feature is located at the last column.
  - Calculate for each gender (male and female) separately, the following stats for the age column: median, mean, quantile (for probs=0.25) and quantile (for probs=0.75).
  - The data visualization process will focus on identifying patterns of specific features. Thus, via R environment, you need to:
    - ✓ Perform a histogram of the frequency of patients per age
    - ✓ Perform a histogram of the frequency of patients per sgpt
    - ✓ Perform a boxplot of Gender (Male & Female) vs age
  - You need to shuffle and re-order the provided data, so that rows are randomly sorted. Then, you need to split the dataset into training and testing sets. Use 80% of the samples for training purposes and the remaining 20% for testing. In this way, each student will have different training/testing sets. These specific training/testing sets will be used for both decision trees models.
- You need to create a decision tree model based on C5.0 algorithm to predict the type of outcome from liver. You need to use the training set for the creation of that model. The model will be then tested using the testing dataset you have already created. The evaluation of your model will be made through all of the following tools: confusion matrix (CM), Area under the Curve (AUC) and F1 Score.
- You need to create a decision tree model based on CART algorithm to predict the type of outcome from liver. You need to use exactly the same training/testing sets, you used for the C5.0 case. The evaluation of your CART model will be made again with the same, as before, tools.
- You need to provide a short discussion, based on these results and decide which model is more suitable for this specific case study.
- You need to improve your current models (C5.0 and CART) via Adaptive Boosting and Pruning schemes respectively. Develop, in R again, the necessary models and eventually perform another performance evaluation using the same testing dataset. Provide a short discussion of any improvement you may have compared to your previous models.

In your main text, you need to include together with your results/discussion the related segments of your R code. At the end of your report, you have to include all of your codes as an appendix.

(Marks 50)

Coursework Marking scheme

The Coursework will be **marked** based on the following marking criteria:

1st Question (MySQL/R Queries/Visualisation)

- **Import the dataset to MySQL** **4**
- **Perform queries via RMySQL/DBI package and display results** **8**
- **Investigate dplyr package for similar style of queries and display results** **10**
- **Provide requested visualisation schematics for this dataset** **8**

2nd Question (OLAP Operations in R)

- **Generation of sales function** **5**

- **Revenue Cube Creation** **5**
- **OLAP operations (5 operations x2)** **10**

3rd Question (Decision Support Systems in BI)

- **Import dataset to R from MySQL** **2**
- **Data exploration & Preparation**
  - **Statistics** **4**
  - **Visualisation** **6**
  - **Training/Testing dataset creation, missing information** **2**
- **C5.0 model (7 marks) + evaluation (4 marks)** **11**
- **CART model (7 marks) + evaluation (4 marks)** **11**
- **Discussion, comparison of methods based on results** **6**
- **Improvement of current two models** **8**