Project

There are two projects contributing 50% to the total assessments of this unit. The projects are to be submitted to cssubmit during the semester.

- **Project 1 submission** is an individual effort on data preprocessing, warehouse design and implementation. The deadline of submission is on Friday 23:59 pm 2nd April (cssubmit). It is worth 25% of the total assessments.
- **Project 2 submission** can be an individual or a paired effort (i.e. to complete the final submission in a group of 1 or 2 people, ideally both from CITS5504), due on Friday 11:59 pm 21st May (cssubmit). It is worth 25% of the total assessments.

Marking scheme and more details of the projects have been available in the following. The overall objectives of the projects are to build a data warehouse from real-world datasets, and to carry out basic data mining activities including association rule mining, classification and clustering.

Project 2 Pattern Discovery and Building Predictive Models

Project 2 aims to produce clean, reduced or transformed data for pattern discovery and predictive analysis. In this project, we will assess the data cleaning and predictive model building skills. You can use either **Weka** or other data analytic toolsets (e.g., R or Python) familiar to you.

Datasets and Problem Domain

For this project, we would like to use the mobile price classification dataset as the source of data. The target of this project is to predict whether the price of a mobile phone is high or not. A copy of the necessary files is here

Your tasks

1. Data cleaning and analysis
    - a. Read through the table and the table column descriptions. Understand the meaning of each column in the table.
    - b. Distinguish the type of each attribute (e.g., nominal/categorical, numerical). You may need to discretise some attributes, when completing Task 2, 3 or 4.
    - c. Determine whether an attribute is relevant to your target variable. You may remove some attributes if they are not helpful for Task 2, 3, or 4. You might create separate data files for Task 2, 3 and 4.
    - d. Identify inconsistent data and take actions using the knowledge you have learnt in this unit.

- **Note**: You may use different data processing procedures, when working on different tasks to get better results.

2. Association rule mining
   - Select a subset of the attributes (or all the attributes) to mine interesting patterns. To rank the degree of interesting of the rules extracted, use support, confidence and lift.
   - Explain the top k rules (according to lift or confidence) that have the "price_category" on the right-hand-side, where k >= 1.
   - Explain the meaning of the k rules in plain English.
   - Given the rules, what recommendation will you give to a company willing to design a high price mobile phone (e.g., should the mobile phone equipped with bluetooth)?
3. Classification
   - Use the "price_category" as the target variable and train two classifiers based on different machine learning algorithms (e.g. classifier 1 based on a decision tree; classifier 2 based on SVMs).
   - Evaluate the classifiers based on some evaluation metrics (e.g., accuracy). You may use 10-fold cross-validation for the evaluation.
4. Clustering
   - Run a clustering algorithm of your choice and explain how the results can be interpreted with respect to the target variable.
5. Data reduction
   - Perform numerosity reduction and perform attribute reduction.
   - Train the two classifiers in Task 3 on the reduced data.
   - Answer the question: "Does data reduction improve the quality of the classifiers"?
6. Comparison of association rule mining, classification and clustering
   - Put all three types of learning together (association rule mining, classification and clustering), and interpret their relations in the context of this problem.

File to submit

A zip file that contains:

1. A report in PDF containing the six tasks listed above. If you work in team, only one submission is needed and the contribution of each team member should be clearly mentioned. Clearly indicate the name and student number of yourself and the team member.
2. All the codes (e.g., python, SQL script) and/or screenshots (for Excel, or other data processing software) of data cleaning and process procedures.
3. Intermediate and final result files for all data processing procedures.

The file needs to be submitted to cssubmit.
Plagiarism is strictly prohibited. Don't submit codes downloaded from the Internet.

Marking scheme (Pattern Discovery and Predictive Analytics)

[30 marks]
[5 marks] Explain the data processing operations (e.g., remove some attributes and action on inconsistent data) that you have done.
[5 marks] Explain and interpret the top k association rules mined; based on the association rules, provide a recommendation for a company willing to design a high price mobile phone.
[5 marks] Explain how you train the classifiers and your evaluation results.
[5 marks] Clustering and interpretation of the clustering result (with respect to the target variable)
[5 marks] Explain the data reduction you have performed; compare the classifiers trained on reduced data with the classifiers trained on the original data.
[5 marks] Your answer to Task 6.

Project 1 Building a Data Warehouse

Project 1 contributes 25% to the final grade of this unit, and requires submission to cssubmit-cits5504. Project 1 is an individual effort on data warehouse design and implementation, due on Friday 23:59 pm 2nd April (cssubmit-cits5504). In this project, we will use the World Bank COVID-19 dataset as the source of data for the data warehouse. A copy of the World Bank COVID-19 dataset is here.

The overall objectives of this project are to build a data warehouse from the given data, and to **answer the following 4 business queries**.

1. What is the total number of confirmed cases in Australia in 2020? What is the number of confirmed cases in each quarter of 2020 in Australia? What is the number of confirmed cases in each month of 2020 in Australia?
2. In Sept 2020, how many recovered cases are there in the region of the Americas? How many recovered cases in the United States, Canada and Mexico, respectively, in Sep 2020?
3. What is the total number of covid deaths worldwide in 2020? What is the total number of covid deaths in large countries, medium countries and small countries, respectively, in 2020?
    - **Note**: In this project, country size is measured by population. Large countries: population >= 40 million; small countries: population <= 2 million; medium countries: 2 million < population < 40 million.
4. Do countries with a life expectancy greater than 75 have a higher recovery rate?

You may follow Kimball's four steps to designing a data warehouse. To realise the four steps, you can start by drawing and refining a StarNet.

Your tasks

1. Observe the data carefully. You can find all the information needed to answer the 4 business queries from the data provided. Draw a StarNet with the aim to identify the dimensions and concept hierarchies for each dimension. This should be based on the lowest level of information you have access to.
2. Use the StarNet footprints to illustrate how the 4 business queries can be answered with your design. Refine the StarNet if the business queries cannot be answered, for example, by adding more dimensions or concept hierarchies.
3. Once the StarNet diagram is completed, draw it using software such as Microsoft Visio (free to download under the Azure Education) or an online drawing service (i.e. draw.io) or a drawing program of your own choice. Paste it to your report.
4. Implement a suitable schema (star/snowflake) using SQL Server Management Studio (SSMS). Paste the database diagram generated by SSMS onto your report.
5. Implement Data Cleaning, Integration and ETL processes (you can use any program languages or/and any software you prefer) and load the World Bank COVID-19 dataset to populate the tables in SQL Server. You may need to create separate data files for your dimension tables.
6. Use SQL Server Data Tools to build a multi-dimensional analysis service solution, with a cube designed to answer your business queries. Make sure the concept hierarchies match your StarNet design.
7. Use Power BI to visualise the data returned from your business queries, and paste the visualisation results to your report.
8. Implement a galaxy schema using SSMS, which contains two fact tables. One fact table is for the number of confirmed, recovered, and death cases. The other fact table is for government measures. Paste the database diagram generated by SSMS onto your report. Give an example of a business query that this data warehouse can answer (**You DO NOT need to answer this query**. Giving an example query would be enough).

Try to complete as many of the above tasks as possible. You should submit a report in .pdf format. Your report should include the StarNet and query foot-prints, the star/snowflake/galaxy schema and the visualisation results to answer the business queries, as well as the explanation on how you perform data cleaning/pre-processing.

Files to submit for Project 1

The followings are the **files needed for Project 1 submission**.

- A PDF report consists of the design, implementation and usage of the data warehouse to answer the queries, the StarNet and query foot-prints, the Star/Snowflake/galaxy Schema, the description of the data cleaning/preprocessing/ETL process for data transformation, and the visualisation results to answer the business queries.
- The SQL Script file and the CSV files for building and populating the tables of your data warehouse.
- The Visual Studio solution project file of the analysis service multi-dimensional project.
- The Power BI file (.pbix).

Marking scheme [45 marks in total]

- [5 marks x 2] Data cleaning/pre-processing/ETL process for data transformation with code or screenshots or explanation
- [5 marks x 2] 4 compulsory business queries that the StarNet can answer and Power BI visualisation corresponding to the 4 business queries
- [5 marks] Concept hierarchies and corresponding StarNet
- [5 marks] Star/SnowFlake schema for date warehouse design
- [5 marks] SQL Script file for building and loading the database
- [5 marks] Coherence between the design and implementation, quality and complexity of the solution, reproducibility of the solution
- [5 marks] The correctness and the quality (for the schema and query) of task 8

Data warehousing exercises are often open-ended. In other words, there is almost always a better solution. **You can interpret the scale of marks as**:

- 5 - Exemplary (comprehensive solution demonstrating professional application of the knowledge taught in the class with initiative beyond just meeting the project requirement. I.e. a highly automated and highly fault tolerant solution for data cleaning/preprocessing and/or other data operations, a deep understanding in dataset with excellent schema design, a very clear, pretty, and convincing powerBI visualisation design, a formal style and well-written report.)
- 4 - Proficient (correct application of the taught concepts)
- 3 - Satisfactory (managed to meet most of the project requirement)
- 2 - Developing (some skills are demonstrated but need revision)
- 1 - Not yet Satisfactory (minimal effort)
- 0 - Not attempted.