6339- Advanced topics in database systems

"Methods of Knowledge Discovery in Healthcare" Project 1

In order to complete this project you are strongly advised to have the following two applications installed on your computer:

- MySQL server
- MS Excel (or any other equivalent Spreadsheet program which can handle .csv files)
- Weka Environment for Data Analysis (version 3.7.1 or newer). You can download Weka 3.7 from here http://sourceforge.net/projects/weka/files/

These instructions have been drafted with the use of MySQL and Weka into consideration, so you are strongly encouraged to use Weka for your data analysis. You therefore need to install MySQL server and Weka in your computer. You do not need to deliver any Weka file. You are going to transfer your output (i.e. from Weka) on your paper using the print screen windows option. For the Weka output, you need to present the whole Weka explorer table. The datasets are samples from hospital claims data. Before you start, go through the list below, which briefly describes the data included in the datasets. After you understand the nature of the data, you are asked to prepare answers for three tasks. The marks are indicated in detail into the deliverable tables.

Feature ID	Feature Name	Description
1	AGE	Age group of the patient. The higher the number, the older the age
2	SEX	1-male, 2-female
3	RACE	different codes represent different ethnicity
4	DAY_OF_ADMISSION	each code represents a different day of the week
5	DISCHARGE_STATUS	A-patient discharged alive B-patient discharged dead
6	STAY_INDICATOR	L-long stay, S-short stay
7	DRG_CODE	a code which is used to charge the patient
8	LENGTH_OF_STAY	number of days that the patient stayed in the hospital
9	DRG_PRICE	the $ amount of the DRG code
10	TOTAL_CHARGES	total charges ($)
11	COVERED_CHARGES	covered charges by the insurance company ($)
12	POA_DIAGNOSIS_INDICATOR_1	defines whether DIAGNOSIS_CODE_1 was present on admission
13	POA_DIAGNOSIS_INDICATOR_2	defines whether DIAGNOSIS_CODE_2 was present on admission
14	DIAGNOSIS_CODE_1	patient diagnosis (using ICD-9)
15	DIAGNOSIS_CODE_2	second patient diagnosis (using ICD-9)
16	PROCEDURE_CODE_1	code describing a procedure that the patient underwent
17	PROCEDURE_CODE_2	code describing a second procedure during hospital stay
18	DISCHARGE_DESTINATION	a code that shows where the patient was sent upon discharge
19	SOURCE_OF_ADMISSION	a code that shows where the patient came from when he was admitted
20	TYPE_OF_ADMISSION	a code defining an emergency or programmed admission
21	ADMITTING_DIAGNOSIS_CODE	initial diagnosis given to a patient upon admission

Before you start

First of all, you will need to download the file **6339_Dataset_1.csv** from Blackboard. This is the dataset you will be analyzing during the project. We assume that you already have installed MySQL Server and Weka 3.7

Task 1

Use your methodology of choice to answer the following research questions:

1. What is the **most common** disease for each age group? What is the prevalence of the **top three** diseases for each age group? Consider that each patient has up to two diagnosis codes.

2. Compare (i) the in-hospital mortality of men and women and (ii) the in-hospital mortality for each of top three diseases between men and women.

3. (i) Are there any demographic factors that are found to be different between long and short hospital stays? (ii) What is the most common long stay primary diagnosis (DIAGNOSIS_CODE_1) and which is the most common short stay primary diagnosis (DIAGNOSIS_CODE_1)

4. What is the effect of the length of stay to (i) the total cost (ii) the in-hospital mortality?

5. Investigate the relationship between the Discharge Destination and the Age Group.

Deliverables

The method (code/queries etc.) you used to find the answers and your output, followed by a short narrative explaining the result **(20)**

Task 2

1. Observe the data and design an appropriate relational schema. Then create the schema on the DBMS system of your choice (preferably MySQL server). The schema should have the following merits:
    a. Normalization principles should be applied to avoid duplicates
    b. Appropriate Data types should be defined

2. Import the data into your newly developed schema

Deliverables

The code (SQL) you used to develop the database schema **(8)**

Description of the method you followed to import the data into the database **(3)**

Your relational diagram with the correct notations **(4)**

3. Create the following queries
    a. An appropriate query which returns a result which is **identical** to the given csv file. This way you are demonstrating how one can extract data from an Electronic Medical Record database, to use for data analysis.
    b. Queries which return the Coverage Ratio

(Coverage Ratio = COVERED_CHARGES/TOTAL_CHARGES) of patients who stayed in the hospital for a period of time longer than 5 days. How does this compare to the Coverage Ratio of patients with a Long Stay?

c. Is there any variation in the average Length of Stay of patients admitted to the hospital in different days of the week (DAY_OF_ADMISSION)? Showcase this with an appropriate SQL query and design an appropriate graph comparing the average Length of Stay between Friday admissions (DAY_OF_ADMISSION=6) and Monday admissions (DAY_OF_ADMISSION=2). Discuss possible reasons that may contribute to what you have found.

d. Using an appropriate method, explore a possible relationship between the DRG_PRICE and the TOTAL_CHARGES. Is there a linear relationship between these two properties?

Deliverables

The code for the queries you created to respond to the above problems **(12)**

Short paragraph answering the theoretical part of question c and d **(4)**

Remember what we discussed in class regarding preprocessing. For Tasks 3 and 4, make sure you have your data in an appropriate data type.

Task 3

Using the Weka implementation of k-means, please find out:

(a) The appropriate number of clusters which are required to adequately 'describe' the discharge characteristics of the patients (discharge destination, discharge status, stay indicator). Use the elbow method to define the number, by evaluating the 'within cluster sum of squared errors' you get as a result in your Weka output. Draw an appropriate graph to explain your answer.

(b) Based on the number of clusters you specified in the above step, please calculate those clusters.

(c) Briefly discuss two interesting (in your opinion) profile groups you have just found.

(d) Is the method we have followed a supervised or an unsupervised data mining technique? Please explain your answer.

Deliverables

Your methodology to define the appropriate number of clusters and the sum of squared errors-cluster number graph, indicating the elbow **(5)**

A screenshot of the Weka output showing the cluster results **(5)**

Short paragraph answering (c) **(5)**

Short paragraph answering (d) **(4)**

Task 4

We need to use data mining to predict whether the Diagnosis Related Code (DRG) price of an admission is going to be higher or lower than $80,000. Make sure your features **are not** in string data type.

(a) Use any appropriate method to modify the class attribute values to be only of two values, either zero (DRG price less than $80,000) or one (DRG price more than $80,000) so that the problem will be **binary classification**. Integrate the new attribute (DRG_PRICE_BINARY) into your dataset.

(b) Observe the available features and explain how the data are acquired in a temporal manner during the healthcare procedure in the real hospital. Specifically, define what do clinicians/administrators already know:

1. At the time when the patient enters the hospital
2. At the time when the patient is discharged from the hospital

Scenario 1: we only know the admission information about the patient. In other words, we only know the Features with ID 1,2,3,4,19,20,21. Only keep those features in Weka before you proceed.

Scenario 2: we know what was known in scenario 1 plus all the information the clinicians and administrators acquired during the hospital care. In other words, we know the Features with ID 1,2,3,4,5,6,8,12,13,14,15,16,17,18,19,20,21. Only keep those features in Weka before you proceed.

(c) Answer, for the two above scenarios, the following questions:

1. Why did we exclude the attributes 7, 10 and 11?
2. Undergo the feature selection process **CfsSubsetEval** to select the appropriate features **for each of the two scenarios**. This way, you will be able to know which features will be included in your classification later on.
3. Use the classifiers (i) Naïve Bayes and (ii) Logistic Regression to classify the DRG_PRICE_BINARY, **for each scenario**, by using the features you found to be useful during feature selection.
4. Discuss the accuracy of the classification for the two classifiers in each scenario, in terms of:
   4a. the overall accuracy
   4b. the accuracy prediction of expensive (>$80,000) DRG costs
   4c. the accuracy prediction of not so expensive (less than $80,000) DRG costs.

Deliverables

The updated dataset with the DRG_PRICE_BINARY attribute with explanation **(4)**

Short paragraph answering (b) **(4)**

Short paragraph answering (c1) **(4)**

Screenshots from Weka showing the results of the feature selection process for question (c2) with brief explanation. One screenshot for each scenario should be included. **(4)**

Screenshots from Weka showing the classification results for question (c3) with brief explanation. Two screenshots (one for each classifier) for each scenario should be included (total 4 screenshots) **(8)**

Detailed explanation of the classification results (questions 4a, 4b, 4c) **(6)**