Assessment Item 1: Problem Solving Task

Exploration of a Large Medical Dataset

Due date: 4 September 2022
Weighting: 20%

Introduction

The tasks set in Assessment item 1 allow you to display your knowledge and understanding of data exploration and preparation for data mining. In this assessment, you will use **Python libraries and code** to understand and analyse a dataset to identify common trends, highlight the data quality issues and correct those problems to enhance the data quality for the future data mining task. **You should mostly use the codes and libraries introduced in the subject**. You will be able to display your technical competence gained from the practicals and the knowledge gained from the lectures and your readings. You will be using the prepared dataset to build data mining models in Assessment Item 2.

Instructions

1. The work your group hands in must be your own; **no collaboration or borrowing from other groups or outside help is permitted**. Read the Assessment Policies on Blackboard or the QUT Website.
2. The assessment report is due on 4th September. You should submit the assessment via Blackboard Assessment. Additionally, in Week 7 practical labs you will have a demo of your answers to the questions defined in the tasks. All members of the group should attend the lab and be available to answer the tutor's questions. Failing to attend the Week 7 lab to showcase the demo (5-10 min per group) will incur an individual penalty of 2 marks.
3. The dataset required for this assessment can be found on Blackboard with the file named **D1.csv**.
4. This is a group assessment. The team size is three. Select the "Groups" tool from the left side of the panel on Blackboard and choose one of the IFN509 groups to register. This should be done by week 2.
5. The group is to be MANAGED by you. As in real life, the performance of the individuals in the team shall be judged by the performance of the team together, so choose your partners carefully.
6. To ensure that everyone agrees to their responsibilities in the team, we have asked the team to complete a Team Agreement form. You can find the team agreement template under the Assessment Item 1 link. Once the team is formed, complete the team agreement form and register the team on Blackboard. It should clearly state if there is an unequal distribution of marks between the team members. This should be agreed upon by all members as shown by the signature of all members.
7. Only a single submission per group is required. The group project report should be submitted via online submission answering each question of the tasks. The report should be correctly formatted and should be easy to navigate through answers provided for all questions defined for all tasks.

There is no need for including an introduction, summary, conclusion, or references in the report. The report should just include responses to the questions set in the case study. Some answers may require screenshots. Use them as needed, but you may include a table detailing those results. While you may like to go into extreme detail, you will not have the space to do so. Rather, write down the important points and attach the important screen dumps, to show that you have thought the matter through. The report is expected to be about 15-20 pages long.

Name the report as **assessment1-report.doc**. The word file should include a cover page with the Student ID number and full name (as in QUT-Virtual) for all students, along with the group number. **Combine the report with your team agreement, Jupyter notebook, and the final dataset** after pre-processing (named as D1-processed.csv) and name the compressed file as **assessment1.zip.** Submit this file on **Blackboard (under the assessment item 1 link)**.

Marks Distribution (Total 20 marks)

The assessment will be marked as two components. The first component, (18% worth), the group work, will be assessed via the report that you will submit on Blackboard. Only a single report should be submitted per group. The second component, (2% worth), the individual work, will be assessed via the questions in the lab demo during Week 7 practical. You should be able to answer the questions to selected tasks raised by the tutor. This will test your understanding of the tasks.

The marks as per the group report are distributed as follows.

Group Assessment (Report) Components	**Marks (18)**
Task 1: Variable Data Types (Initial and Final)	1
Task 2: Exploration with Statistical Measure	4
Task 3: Exploration with Visualisation Plots	4
Task 4: Data Preparation	5
Task 5: Feature and Task Selection	3
Team Agreement	0.5
Report Presentation	0.5
Individual Assessment (Lab Demo) Components	**Marks (2)**

Data Analysis Scenario

eHealth is having a tremendous impact on the way healthcare systems are operated. Practitioners are investing significant resources into how best to utilize the large volumes of data that eHealth utilities are gathering. Electronic Health Record (EHR) data analysis offers great hope in the discovery of useful knowledge for better health service management and improving patient outcomes. Data mining techniques have been commonly applied to EHR data for knowledge discovery in the recent decade. The most popular application is risk prediction such as predicting readmission, length of stay, heart failure, etc.

For this assignment, you are given an EHR dataset in CSV format and are expected to explore the data to get a good understanding of the data before conducting data mining. This dataset consists of 50,031 instances of diabetic patient encounters represented by 39 variables about patient information and hospital outcomes. Each encounter is an inpatient encounter (i.e., a hospital admission). The laboratory tests were performed during the encounter and the medications were administered during the encounter. The data was collected from more than 100 American hospitals from 1999 to 2008. The variables are described below:

No.	Variable Name	Description
1	encounter_id	Unique identifier of an encounter
2	patient_nbr	Unique identifier of a patient
3	race	Race of the patient
4	gender	Gender
5	age	Age quantile
6	weight	Weight in pounds
7	admission_type_id	Identifier corresponding to 9 distinct admission types (see IDs_mapping in Appendix 1)
8	discharge_disposition_id	Identifier corresponding to 29 distinct values (see IDs_mapping in Appendix 1)
9	admission_source_id	Identifier corresponding to 26 distinct values (see IDs_mapping in Appendix 1)
10	Length_of_stay	Number of days between admission and discharge
11	payer_code	Unique identifier assigned to each insurance company
12	medical_specialty	Indicates a specialty of the admitting physician, for example, cardiology, surgeon, etc.
13	num_lab_procedures	Number of lab tests performed during the encounter
14	num_procedures	Number of procedures (other than lab tests) performed during the encounter
15	num_medications	Number of distinct generic medication names administered during the encounter
16	number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter
17	number_emergency	Number of emergency visits of the patient in the year preceding the encounter
18	number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter
19	diag_1	The primary diagnosis (coded as the first three digits of ICD9)
20	diag_2	Secondary diagnosis (coded as the first three digits of ICD9)
21	diag_3	Additional secondary diagnosis (coded as the first three digits of ICD9)
22	number_diagnoses	Number of diagnoses entered into the system

23	diabetes	Indicates if the patient's primary diagnosis is diabetes or not. Values include: "Yes" and "No"
24	max_glu_serum	Glucose serum test result. Indicates the range of the result or if the test was not taken ("none")
25	A1Cresult	A1c test result. Values include: '>8' if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.
26	metformin	These are 10 variables for diabetes medications. The values of these variables indicate whether the drug was prescribed or there was a change in the dosage. Values include: "Up" if the dosage was increased during the encounter, "Down" if the dosage was decreased, "Steady" if the dosage did not change, and "No" if the drug was not prescribed.
27	repaglinide	
28	nateglinide	
29	chlorpropamide	
30	glimepiride	
31	acetohexamide	
32	glipizide	
33	glyburide	
34	tolbutamide	
35	insulin	
36	change	Change of medications. Indicates if there was a change in diabetic medications (either dosage or generic name). Values include: "Ch" and "No"
37	diabetesMed	Diabetes medications. Indicates if there was any diabetic medication prescribed. Values include: "Yes" and "No"
38	Readmitted	Days to inpatient readmission. Values include: "<30" if the patient was readmitted in less than 30 days, ">30" if the patient was readmitted in more than 30 days, and "No" for no record of readmission.
39	single_day_admission	Indicates if this encounter is a single-day admission or not Values include: "Yes" and "No"

Table 1 Variable Description

Tasks

Suppose you have been hired as a data analyst consultant by the company. Before you conduct data mining, your task for this assignment is to understand the data, explore it for any quality issues using summary statistics measures as well as visualisation plots, find common patterns, and correct any data problems you have identified.

Answer the following (add screenshots as appropriate).

1. Examine the data types assigned by the `Pandas' library for each variable. Check them with the dataset description provided in Table 1. If there is a mismatch in the data type assigned by the library and the data type as per the description for a variable, correct the data type. Report them. Attach a screenshot showing the correct data types of variables.

2. Using suitable statistical measures and functions:
    1) Identify and report the skewness present in the variables.
    2) There may be inconsistencies, errors, or missing values in the data. List the errors identified and detail how you have identified them.
    3) Answer the following questions: call method, calculate values to show the result
        a. What is the average length of stay in the hospital of a male patient who was readmitted in less than 30 days?
        b. Which age group has the highest number of encounters whose primary diagnosis is diabetes? What is the number?
        c. How many encounters whose admission type is Emergency? How many of these emergency encounters are readmitted within 30 days?
        d. What are the top-three race categories according to the number of readmission cases (including both less than or larger than 30 days)?

    Do grouping and calculate value

    Note: Do not just answer. Elaborate on how you have reached the answer.

3. Using suitable visualisation plots: (August 8th - 14th) mathlabplot
    1) Understand the distribution of variables and identify data quality problems.
    2) Determine if there is any relationship between the variables **`dibetes`** and **`diabetesMed`**? How would you handle these two variables in the data modelling if a relationship exists?
    3) Identify the highly correlated variable pairs and elaborate on how to treat these variables in the mining process in such a case.

do conversion => convert category to numerical value

4. Data preparation: Week Aug 15th
    1) Summarize your findings based on data exploration.
    2) Elaborate on the data preparation steps required (e.g., data cleaning and transformation) to address the data quality problems that you encountered during data exploration.
    3) Demonstrate the data preparation by including a screenshot(s) of the Python code and its outputs that show the steps on how you corrected all the identified data quality problems in this dataset.

5. Selection of data mining task and feature selection: Week Aug 22nd
    1) Identify the most suitable data mining task (i.e. classification, clustering or association mining) that can be performed on this dataset. Justify your choice.
    2) What variables will you include in this data mining task and why? Describe here if you will create any derived variables. Identify the roles (i.e., input or target) of each variable.

Assessment 1 Marking Sheet

	Tasks	Requirements	Marks
Report Marking (group)	1. Variable data types	• Identify mismatched data types • Correct mismatch data types • Screenshot or table	1
	2. Exploration with statistical measures	• Identify skewness in variables • Identify inconsistencies and errors in variables • Briefly describe how you have identified them • Answer the questions	4
	3. Exploration with visualization plots	• Use appropriate visualization methods • Appropriate plots, screenshots • Observations regarding variable distribution and data quality • Identify relationships and how to handle them in the mining process	4
	4. Data preparation	• Summarize exploration findings • State data preparation steps • Correct data quality problems • Outputs, screenshots	5
	5. Feature and task selection	• State suitable data mining tasks with justifications • Select variables with explanations	3
	6. General	• Team agreement • Report presentation	1
Lab demo Marking (individual)	Random Q&A during the lab demo		2

Appendix 1 IDs Mapping

admission_type_id	description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped
discharge_disposition_id	description
1	Discharged to home
2	Discharged/transferred to another short-term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA
8	Discharged/transferred to home under the care of Home IV provider
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonatal aftercare
11	Expired
12	Still patient or expected to return for outpatient services
13	Hospice / home
14	Hospice / medical facility
15	Discharged/transferred within this institution to Medicare-approved swing bed
16	Discharged/transferred/referred to another institution for outpatient services
17	Discharged/transferred/referred to this institution for outpatient services
18	NULL
19	Expired at home. Medicaid only, hospice.
20	Expired in a medical facility. Medicaid only, hospice.
21	Expired, place unknown. Medicaid only, hospice.
22	Discharged/transferred to another rehab fac including rehab units of a hospital.
23	Discharged/transferred to a long-term care hospital.
24	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
25	Not Mapped
26	Unknown/Invalid
30	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere
27	Discharged/transferred to a federal health care facility.

28	Discharged/transferred/referred to a psychiatric hospital or psychiatric distinct part unit of a hospital
29	Discharged/transferred to a Critical Access Hospital (CAH).
admission_source_id	description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critical access hospital
11	Normal Delivery
12	Premature Delivery
13	Sick Baby
14	Extramural Birth
15	Not Available
17	NULL
18	Transfer From Another Home Health Agency
19	Readmission to Same Home Health Agency
20	Not Mapped
21	Unknown/Invalid
22	Transfer from hospital input/same fac result in a sep claim
23	Born inside this hospital
24	Born outside this hospital
25	Transfer from Ambulatory Surgery Center
26	Transfer from Hospice