

**IMU UNIVERSITY MALAYSIA**

**Bachelor's in Digital Health**

**Health Data Engineering**

**(BDH 2253)**

**Assignment**

**Student Name:**

**Yap Jun Song (00000042067)**

**Describe the source and format of the dataset used**

The dataset used for this assignment is the diabetic\_data.csv file. This dataset mentions the medical records of diabetic patients from different age groups and the medications they used. It also includes attributes such as admission details, diagnoses, hospital readmission status, etc.

|  |  |
| --- | --- |
| Description | Amount |
| Attributes | 50 columns |
| Instances | 101,767 rows |
| File size | 18,711 KB |

**Demonstrate the steps to extract/import the data**

**Step 1:**

Firstly, import the pandas library, then put the correct file path to let the computer locate the file. Then using the ‘pd.read\_csv()’ to access the file and use the df.head() to print out the first 10 instances inside the dataset.

A computer screen shot of a computer code

AI-generated content may be incorrect.

Output:

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer

AI-generated content may be incorrect.

**Mention any challenges encountered during extraction**

During the data extraction and import process, there are several challenges that we will face, which are large file size, missing and inconsistent values, and file path issues. Due to the dataset preserving over 100,000 patient records, which will cause the computer to become slightly delayed when loading the file, and it requires enough system memory to execute it. Then, some columns have the missing value, like the weight and medical specialty columns need to do data imputation if needed, and there are some columns that can do one-hot encoding during data preprocessing. Lastly, because the file will need to work on different machines, it is crucial to set the correct file path for each machine to ensure the smooth data import and extraction.

**Objective**

From the dataset, we can uncover the hidden association between medication strategies (including medication types, medication combinations, and medication change status) and hospital readmission among diabetic patients in different age groups. There are 4 sub-objectives that we can analyse:

1. Determine which medication is the most commonly used in different age groups.

2. Discover the age groups with the highest readmission rates among diabetic patients

3. Find that the medication change rate happens in different age groups

4. Diabetes medication usage rate in different age groups.

**Data Preprocessing**

**Step 2:**

To meet the objectives that we set, the first step we need to take is to drop those columns that are not related to our objectives. After removing the columns that are not related to our objectives, like patient ID, race, gender… This can reduce the column and simplify the dataset, just keep the variables that are related to medication factors, age, and readmission. Therefore, this can help us to directly focus and analyze the relationship between medication types, medication combinations, medication change, age, and readmission.

A screen shot of a computer program

AI-generated content may be incorrect.

Output:

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Step 3:**

After dropping the unnecessary column, we need to check whether the remaining columns have missing values or not. This step is important because it can ensure the results of our analysis, like the distribution of different drugs in different age groups, are reliable and not affected by missing values. Besides that, this can avoid errors or biases caused by missing data during analysis.

**A screen shot of a computer code

AI-generated content may be incorrect.**

Output:

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

**Step 4:**

This step is to avoid the logical conflicts. For example, the row of the dataset shows the ‘diabetesMed=No’, but the patient is having some medication. This shows that the row of data records is wrong or the system problems with the data record. If not checked properly, the analytics of the frequency of medication use and the combination strategy will be contaminated. These logical errors may become the source of the error and lead to the result of getting misled when analysing the relationship between drug factors and readmissions.

**A screenshot of a computer program

AI-generated content may be incorrect.**

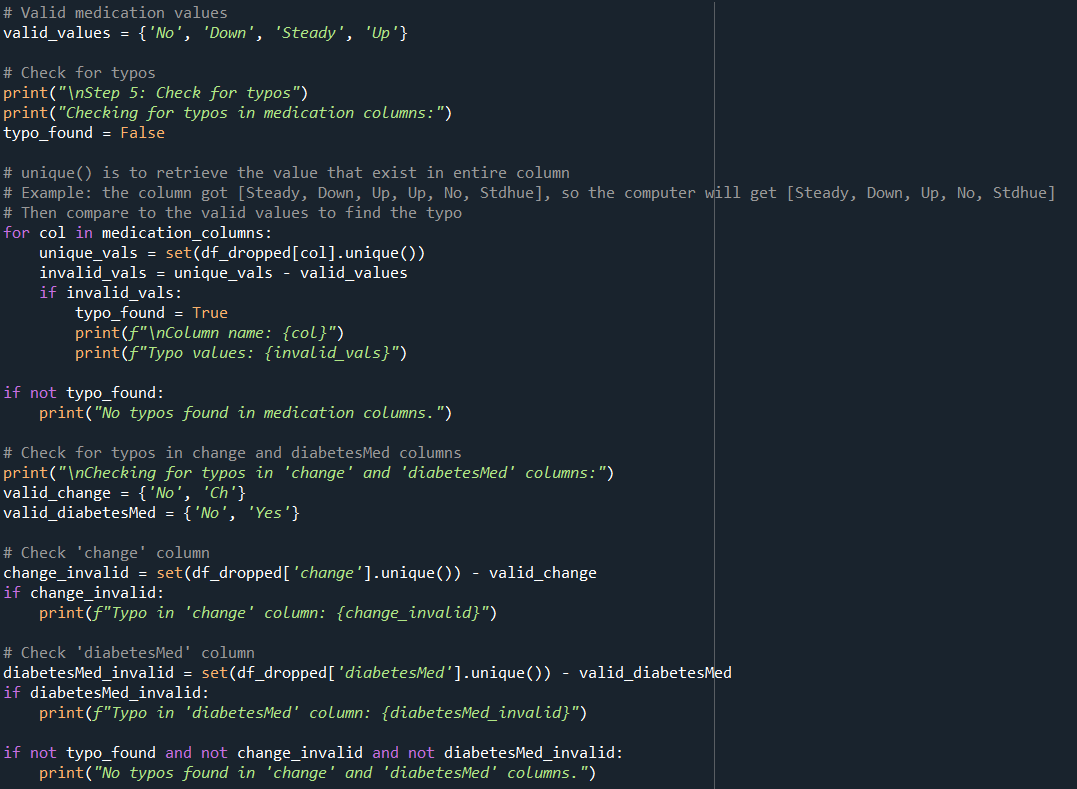
Output:

A screenshot of a computer code

AI-generated content may be incorrect.

**Step 5:**

After that, move to the crucial part of data preprocessing, which is checking the spelling errors or inconsistencies, like the data ‘No’ is written as ‘Nu’. The data with spelling errors will cause coding failures because the computer cannot detect and it will split the categories into several incorrect groups when counting. This can ensure the consistency in drug and readmission statistics to avoid category confusion for statistical analysis. Apart from that, as the remaining column is categorical data so there is no need to find the outlier. If the dataset is numerical data, we will need to plot a box plot graph to find the outliers and handle them, whether to remove the outlier or change the outlier to the nearest data point.

****

**A screen shot of a computer code

AI-generated content may be incorrect.**

Output:

**A screen shot of a computer

AI-generated content may be incorrect.**

**Step 6:**

Move to the most important step in the data preprocessing, which is one-hot encoding. Due to my objective, I just need to know how many medications are used, the medication factors situation, and the hospital readmission. Therefore, I just do the one-hot encoding for these columns. This can ensure that medication usage patterns can be accurately counted, compared across different age groups, and linked to hospital readmission outcomes in subsequent analyses. For the readmission column, there are three types of data, which are ‘NO’, ‘<30’, and ‘>30’. To meet my third objective, I need to find the association between medication factors and hospital readmission, so I will set the patient with readmission no matter ‘<30’ or ‘>30’, to become 1 because they can be seen as readmitted, meanwhile, no readmission record will set as 0. This binary encoding can transform the qualitative data into a structured format can enhancing the clarity and statistical power of subsequent uncover hidden patterns and associations with allowing the straightforward comparison of readmission rates between different medication strategies.

**A screen shot of a computer program

AI-generated content may be incorrect.**

Output (Before doing one-hot encoding):

A screenshot of a computer

AI-generated content may be incorrect.

Output (After doing one-hot encoding):

A screen shot of a computer

AI-generated content may be incorrect.

A blurry image of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

**First objective: Determine which medication is the most commonly used in different age groups**

From the dataset that has already been encoded and cleaned, we need to specifically extract and aggregate the significant columns to meet Objective 1. The objective is to determine which medication is mostly used in different age groups. To achieve this, we need to group the data by age and calculate the occurrences of each medication’s usage within each age group. The output can empower the researchers to plot a graph with the age groups on the X-axis and the amount of selected medication on the Y-axis, which can provide a clear perspective into medication distribution trends across age ranges.

A screen shot of a computer program

AI-generated content may be incorrect.

Output (Objective 1):

A screenshot of a computer

AI-generated content may be incorrect.

**Second objective: Discover the age groups with the highest readmission rates**

To meet this objective, we will need to calculate the total number of patients with readmission, then calculate the patients with readmission in different age groups. By computing the readmission rate for each age group, we can determine the percentage of patients readmitted among the total number of patients in that particular age group. So that we can compare and identify which age group has the highest probability of hospital readmission. The output can provide a visualized graph when putting the age group on the X-axis and the readmission rate on the Y-axis, then can provide an obvious highlight of the age groups with the highest rate of readmission.

A screenshot of a computer program

AI-generated content may be incorrect.

Output (Objective 2):

A screenshot of a computer screen

AI-generated content may be incorrect.

**Third objective: Find the medication change rate in different age groups**

Next, we will need to count the total patients within each age group, then compute the amount of medication change, which is ‘change\_encoded = 1’ in different age groups. After that, we can know the medication change rate among different age groups. By knowing the medication change rate, we can discover and compare the frequency of medication adjustments across age groups to get clear insights into how medication change practices vary with age and lead to provide a precise treatment to patients.

A screen shot of a computer screen

AI-generated content may be incorrect.

Output (Objective 3):

A screenshot of a computer

AI-generated content may be incorrect.

**Fourth objective: Diabetes medication usage rate in different age groups**

The original dataset includes all diabetic patients, so we will need to find the patients who are taking diabetic medication in different age groups to meet this objective. By calculating the percentage of patients with diabetes medication in each group, it can provide a clear insight for the pharmaceutical company to let them understand the diabetes medication usage changes by age and potentially guide the age-specific treatment or product planning.

A screen shot of a computer program

AI-generated content may be incorrect.

Output (Objective 4):

A screenshot of a computer

AI-generated content may be incorrect.

**Preprocessing pipeline**

A diagram of a diagram

AI-generated content may be incorrect.

This preprocessing pipeline is specifically designed for the diabetic dataset. The purpose of it is to ensure the dataset is clean, consistent, and directly relevant to the preset objectives. First, we need to import the dataset and define the objectives clearly. Next, we need to observe whether each attribute contributes to the objectives or not; otherwise, we need to remove the irrelevant columns to simplify the dataset and help us focus on the objectives. After that, we need to check the existence of a missing value. If it has a missing value, we need to use different imputation techniques like mean, median, or mode based on the attribute’s distribution. This step can ensure that our result will not be affected by the missing value and lead to bias. Then, logical issues are a part of something not easy to find out. We need to verify the consistency between columns, like no patients are recorded as not taking medication, but the medication columns show the usage. We will remove the rows with the logic errors because they cannot be solved and will lose the reliability of the results. Additionally, need to identify and correct the spelling errors to ensure the data is consistent and allow the subsequent step to work. Lastly, we will change some columns (medication types, change status, diabetesMed, and readmission) into categorical data to achieve quantitative analysis. After doing all the steps of preprocessing, we can get the clean data to meet our preset objectives, and the analysis can provide a reliable and clear comparison and visualization based on high-quality and structured data.

**Reflection**

Data preprocessing is essential in health data analysis. As datasets are growing extremely and becoming more complex, they have a large volume and are heterogeneous, which need to be processed and transformed the raw data into a cleaner, more structured format meanwhile removing the noise to make the dataset easier to process (Khan, 2024). Besides that, it is also a procedure of the data cycle that can ensure the collected information is reliable and the data can be standardized, normalized, or processed using other methods to present a visualization in an appropriate way to stakeholders (Shania, 2023).

First, for Objective 1, if the spelling errors are not handled properly, such as not removing the typo in preprocessing, like the drug name being recorded with the wrong spelling. The total amount of the same drug used will be divided into various categories and leading to inaccurate statistics of the most commonly used drugs among each age group. This will let the researcher underestimate the actual usage rate of the drugs and misjudge the usage trend of drugs.

Secondly, the readmission rate of different age groups cannot be fairly compared if we do not set the correctly coded for readmission column, like <30 and >30, to ‘1’ to meet our Objective 2. This will let the doctor misjudge the readmission risk of some age groups and affect the identification of high-risk groups because the readmission rate is not accurate.

Lastly, if we do not handle the missing values or logical errors correctly, it will severely affect our objectives 3 and 4, then lead to bias and unreliable results. The medication change rate and the use rate of diabetes drugs will be distorted. The consequences of it will lead the doctors to a misunderstanding of medication adjustments for different age groups and affect the precision medicine programs, thereby influencing drug production and supply chains.

In summary, preprocessing can help to streamline the data, reduce the dimensional complexity and processing time, thus letting the clean, well-structured data tell a clearer story, which allows the organization's analysis to be faster and more efficient (Ash Lei, 2025). On the other hand, if the preprocessing is not done in a good way, the reliable foundation of each objective will be destroyed, and the truth will be misled. So that the data-driven decisions will lose their meaning and hurt patient health or medical resource assessment. Therefore, data preprocessing not only plays a technical role, but it is an important guarantee to ensure the quality and value of health data analysis, to allow the doctor to provide precise treatment and empower the patient to receive better treatment.

**References**

Ash Lei. (2025). *Benefits of Data Preprocessing*. Byteplus.com. https://www.byteplus.com/en/topic/402858?title=benefits-of-data-preprocessing-transforming-raw-data-into-actionable-insights

Khan, F. (2024, April 30). *What Is Data Preprocessing: Definition, Importance, and Steps*. Astera. https://www.astera.com/type/blog/data-preprocessing/

Shania, K. (2023, September 27). *The Healthcare Data Cycle: Generation, Collection, and Processing | TechTarget*. Healthtech Analytics. https://www.techtarget.com/healthtechanalytics/feature/The-Healthcare-Data-Cycle-Generation-Collection-and-Processing