Credit Member Statement – CP3403/CP5634

This form is to recognise individual member contributions, reducing member disputes and facilitating collaboration within group. Please insert this credit member statement to the front page of your data mining report.

Member Name	Contributions (Percentage)
Nayana Ann Moni	- Preprocessing data o Transformation (40%) o Conversion (30%) o Data cleaning (35%) - Mining data o Logistic Regression (100%) - Analysing data and interpreting patterns o Visualisation (100%) - Writing the DM report o Overall framework (100%) o Preprocessing and data mining approaches (100%)
Rosmi	- Preprocessing data o Transformation (30%) o Conversion (40%) o Data cleaning (35%) - Mining data o (100%) - Analysing data and interpreting patterns o Generating tables and graphs (100%) - Writing the DM report o Analysis and discussion (100%) o Introduction & Conclusion (100%)
Adewale Arogbonlo	- Preprocessing data o Transformation (30%) o Conversion (30%) o Data cleaning (30%) - Mining data o Decision tree (100%) -Analysing data and interpreting patterns o Result interpretation and analysis (100%) - Writing the DM report o Business Scenario (100%) o Data Description (100%)

A Data Mining Approach to Predicting Hospital Readmission Rate of Diabetic Patients

Nayna Ann Moni · Rosmi · Adewale Arogbonlo

May 29, 2022

Abstract

Data mining is used to find better patterns to make better decisions. The objective of this study is to assist hospitals to improve diabetic patient care and to reduce the number of times the patients are readmitted. A patient that is readmitted within a short time frame is considered a readmission. Data mining techniques are used in this study to predict readmission rates by analysing a data set of patient admission history. Decision tree, Logistic regression and random forest were used to predict readmission. Random forest gave the best accuracy of 94% amongst the three models.

Key words: *Data mining, Hospital readmission, Diabetes readmission, Prediction*

1. Introduction

There are different types of diabetic patients regularly admitted into a hospital. Even though technology has advanced greatly over the years, diabetic patients still face the challenge of readmission. Diabetes is one of the most costly conditions to treat hence readmission is a huge drawback to both the individual and the hospital. The process of recognising patients who are likely to be readmitted within a short period of time is unpredictable with the naked eye. Without further analysis it may seem there is no relation between a readmission case of a young diabetic patient with type 1 diabetes and an elderly patient with type 2 diabetes. Data mining techniques can be used to extract largely significant information from large data sets that otherwise may seem trivial.

Readmission of patients within 30 days could be due to inadequate treatment, poor patient compliance, inadequate follow-up care, insufficient information given to patients, poor after care by caregivers or even medical errors. This could largely affect a hospital's reputation and also increase their costs. According to the HCUP Nationwide Inpatient Sample 2016, total hospital costs for patients with diabetes is up to $ 123 billion [1].

From a business viewpoint, the aim of this study is to analyse past patient readmission data using data mining models to predict the odds of readmission with high accuracy. As it is solely a medical professional's decision to discharge a patient, the decision is often biased and unreliable due to limited knowledge. Data mining approaches have been successfully applied in several fields to find hidden patterns from data sets. This could greatly assist with better decision making as it can produce evidence of treatment trends and their results.

2. Data Description

To discover the relationship between patient history and readmission rates, We obtained a data set of diabetic patients from a public repository. The data set has 50 attributes and contains 101766 rows of patient encounters. The data set contains medical record of drugs prescribed, past diagnosis, tests performed on a patient and the dates of discharge and readmission. The data set is a record of patients in 130 United States hospitals from the year of 1999 to 2008. Out of the 50 features, 27 are of the different types of drugs administered to the diabetic patients. The other attributes are as follows: 7 historical data, 5 personal details of the patient and 11 patient encounter variables for hospital usage. For a better understanding of the data set, consider the following factors:

- Each row in the data set represents a unique encounter with a patient. The same patient might have had several encounters therefore, many of the rows may have the same patient's details.

- The patient's identifying details, such as gender and race, as well as age which is maintained as categorical variables that appear as labels representing intervals of 10 years.

- ICD-9 standard codes are used to represent the diagnosis made of the patients during the consultation in columns diag_1, diag_2 and diag_3.

- A '?' symbol is used to indicate missing values in any attribute.

- Each row also has 24 attributes related with various diabetes drugs, each of which indicates if the drug was prescribed, or a change in its dosage was administered. The possible values in these colums are; 'NO' which means the drug was not prescribed, 'Steady' meaning no change was made to the dosage, 'Up' meaning the dosage increased, and 'Down' meaning the dosage was decreased during the visit.

- The max_glu_serum value represents the result range of the glucose serum It is represented by the values '>200', '>300', 'normal' and none if it was not taken.

- A1Cresult has the variables '>8', '>7', 'normal' and 'none' as the possible values.

- Change_of_medication shows whether a patient was prescribed any new medication or if they were asked to get off any medication, represented by 'change'. If there were no changes, this is represented by 'no change'.

- 'readmitted' is the most import feature, which is used to determine if the patient was readmitted within 30 days, represented by '<30, readmitted after 30 days, represented by '>30' and no record of readmission was represented with 'No'.

3. Business scenario

Hospital readmission is seen to be a good indicator of how well diabetes patients are being cared for. As a healthcare professional, you'd like to find associative and/or correlative patterns in diabetes patients' medical histories that will help you figure out when they need to be admitted to the hospital and how to limit the number of readmissions. The capacity to accurately detect patients who are at high risk of being readmitted to the hospital in the next 30 days should allow for more enquiry and possibly avert the readmission. This is a serious issue in the healthcare system since a large number of avoidable hospital readmissions are caused by poor care during patients' hospital stays and a poorly organised release procedure. Several prior research looked into the elements that influence diabetes patients' readmission rates. The majority of research, however, focuses on subgroups of diabetes populations, and solutions are generated from a lower sample size than this study. The findings were based on demographic and socioeconomic characteristics that determine readmission rates integrated in some cases [2]. In certain circumstances, the models have no specified aim and instead focus on all-cause readmission.

4. preprocessing and data mining approaches used

Data preprocessing is the first step in data mining. It is used to prepare raw data for further analysis. Unprocessed data contains a lot of missing values and unnecessary information, it may be noisy and inconsistent.This could lead to inaccurate results and errors. There are several forms of data preprocessing; data cleaning, data integration, data transformation, data discretization and data reduction. The steps taken to clean the selected dataset are described in the following subsections.

4.1. Data Cleaning

Missing values have a significant impact on the findings, and there are numerous strategies for resolving this issue. In general, if the number of missing values is modest, the sample with the missing values can be removed [3]. To handle the missing values the following samples were deleted in this process:

- In our analysis, we found that for the weight column 98569 values out of the 101766 were missing. Since more than 90 percent of the data were missing, the weight variable was removed. As there are too many unknown values, it will not produce any useful patterns.

- payer_code and medical_speciality also had too many unknown values, 40256 and 49949 respectively. So these variables were dropped

- Race, gender, diag_1, diag_2, and diag_3 had a lot less missing values than the other qualities we discarded, therefore we decided to drop only the rows with the missing values for these variables. 2273 samples had missing race, 3 missing in gender, 21 missing in diag_1, 358 missing in diag_2 and 1423 missing in diag_3.

Following the removal of missing values and other potential bias from the data, We removed data with the same values in all columns and reduced the number of unique values for categorical variables.

4.2. Data Aggregation

To optimise the features for data mining, the following steps were taken;

Some attributes had many possible values in the columns. We used a clustering approach to group similar variables together. This was done for 3 attributes, admission_type_id, discharge_disposition_id and admission_source_id.

admission_type_id: The terms "emergency" and "trauma centre" were combined to become "emergency". "Null" combined the terms "not available" and "not mapped."

discharge_disposition_id: This attribute had too many variables (25) so they might be grouped together even more to reduce the variables, the following were done:

- Wherever the patient was discharged or transferred out to a home, hospice, or care provider, about 4 categories were all mapped as discharged to home.

- Wherever the patient was transferred out to another health facility, hospital, inpatient care, or any other type of medical institution were categorised as discharged/transferred to another short term hospital. This condensed 10 variables into one.

- All those who were transferred within the hospital were mapped to discharged/transferred to SNF. Here 6 categories were condensed.

- Those who passed away are categorised into a category called expired. Here we managed to condense 3 categories into one.

- Rows with null and unknown/invalid values were mapped as NULL.

admission_source_id: This attribute could also be further condensed into fewer categories.

- Physician, clinical and HMO referrals were mapped to a category called referrals.

- A transfer from any type of health facility were mapped to a catergory called transfer.

- Emergency room, premature delivery, sick baby, and extramural birth were categorised as emergency.

- Not available, NULL, not mapped, unknown/invalid were mapped as not available.

- Normal delivery, born inside this hospital, and born outside this hospital were also mapped as one category.

4.3. Data Reduction

Data reduction condenses data into a smaller dimension while still keeping the integrity of the original data [4]. The attributes citoglipton and examide had the same value for every patient so these columns were dropped as they cannot provide any variation to the data or contribute any useful information. We also dropped those who were categorised as 'expired' as readmission will not occur for these patients.

4.4. Variable encoding and Categorization

- The "medication change" characteristic was encoded as 0 and 1 instead of "No" (no change) and "Ch" (changed).
- The Gender property in the dataset comprises values such as 'Male' and 'Female,' but because our model does not recognise string values, we changed them to numerical values, for example. Male – 1 and Female – 0 were encoded in the Gender attribute.
- No value was categorised as 0 in 23 prescribed medications, while steady, up, and down variables were classified as 1.
- We consolidated the readmission after 30 days and no readmission into a single category to limit our challenge to a binary classification. In the readmitted attribute, 0 was used to replace the values >30 and 0. whereas 1 was used to replace the readmission within 30 days (<30).

4.5. Grouping variables

We decided to group the values of several attributes in order to make the classification process easier.

- We chose to replace the A1Cresult attribute values >7 and >8, as well as the max_glu_serum attribute values >200 and >300 with 1. Whereas the value 'norm' in both attributes where replaced with 0.The values 'none' in the A1Cresult attribute and the max_glu_serum attribute were replaced with the value -99.
- We are not aware of the exact age of each patient because the data set only provides us age in 10-year increments. We presume that the patient's age is roughly in the middle of the age range. If the patient's age range is 20–30 years, we'll presume the patient is 25 years old. As a result, we translated age groups to midpoints, yielding a single number variable.
- All three diagnosis attributes, diag_1, diag_2, and diag_3, have been replaced by a numerical value that represents the value range to which they belong.The values in the diag_1, diag_2, and diag_3 characteristics that contain V and E were coded as 0. The attributes diag_1, diag_2, and diag_3 were classified into nine groups.

4.6. Duplicate Data

- Some of the patients in the database had many encounters. Then, as possible representations of numerous encounters, we looked at the first encounter and the last encounter independently. However, we decided to use first encounters of patients because last encounters provided severely skewed data for readmissions. The first entry was picked because it has the largest chance of being readmitted, which aids in data balancing.

5. Patterns and Findings

We did data visualization to identify hidden patterns in our data.

In previous sections readmitted is defined as the patient being readmitted within 30 days of discharge. There were 3 categories of readmission in the raw data. 'No' to indicate the patient was not readmitted again, '<30' indicating the patient was readmitted within 30 days and '>30' indicating the patient got readmitted more than 30 days after the discharge. Since we are only focusing on those who were readmitted within 30 days, we decided to categorise 'No' and '>30' into one class. We decided to classify No readmission within 30 days as '0' and readmission within 30 days as '1'. The pie chart in figure 1 below shows that 11.29% of the patients were readmitted within 30 days. As we can observe from the representation of our target variable's distribution, it is significantly asymmetrical, our prediction model will neglect the minority class and favour the majority. Patients who have been readmitted are far fewer than those who have not been readmitted.

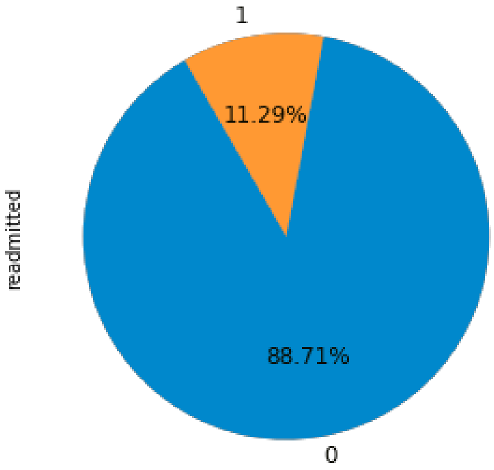

Figure 1: Total readmission statistics

We then decided to visualize the data further to see how the readmission rates are, in comparison to other attributes.

There were more female patients than male patients so even though the readmission rates are higher in females, there was no big difference in the ratio. Figure 5 represents this in a bar chart.We also compared the age and ethnicity statistics to see if there are any trends in readmission.

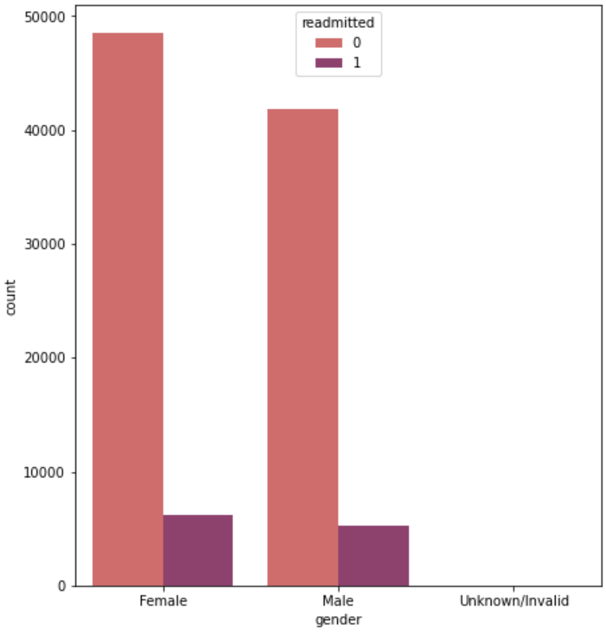

Figure 2: gender vs readmission

A1c result showed that if A1c is not measured there is a higher chance of readmission, figure 5.

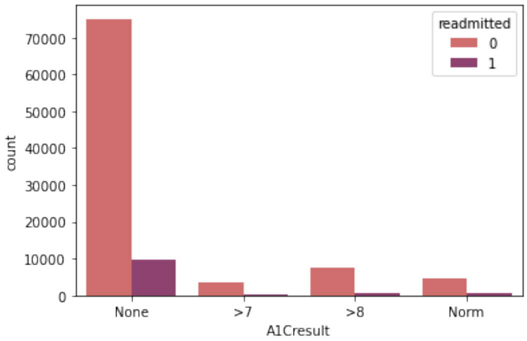

Figure 3: A1Cresult vs readmission

The number of lab procedures appeared to be similar for both readmitted and non readmitted patients so it does not create any bias in the mining. In figure 5 blue indicates the readmission frequency and green represents non readmission.

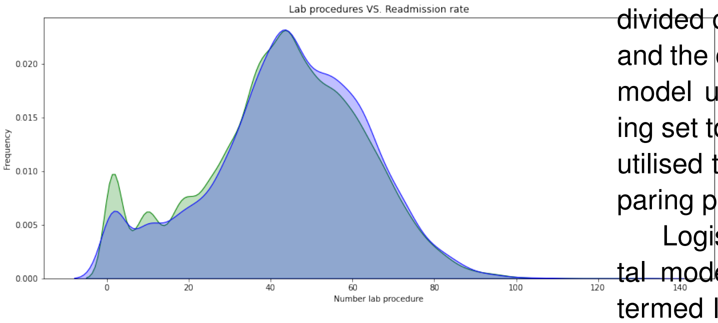

Figure 4: Lab procedure vs readmission

We then further analysed the effects of changing medication and the prescription of diabetes medication to the readmission rates. As it can be observed in figure 5 readmission rate is higher for those patients who were prescribed diabetic medication.

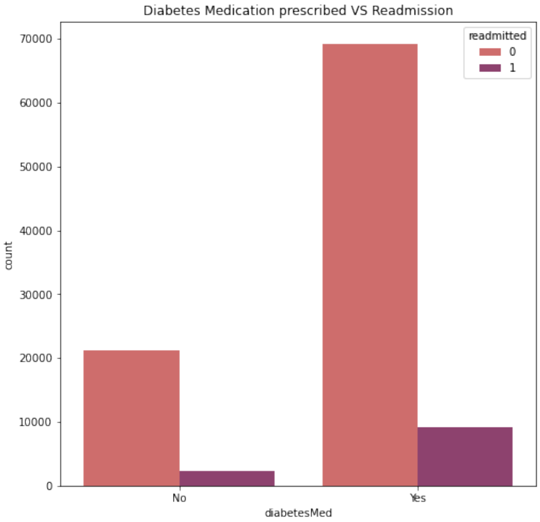

Figure 5: Diabetes medication vs readmission

6. discussions and comparison

Once data preprocessing was completed, we employed data mining techniques such as Logistic Regression, Decision Tree, and Random Forest to estimate the number of readmitted patients and compare which model performed better. Our final model was chosen because it produced the best outcomes.

Data split: We cannot use the complete dataset to train our model.If we train our model on the full dataset, we will run into the problem of overfitting, and our model will make incorrect assumptions for fresh observations.To evaluate our model's performance and dependability, we divided our dataset into two sets, one for training and the other for testing, in the ratio of 80:20.The model uses 80percent of the data in the training set to learn, while the remaining 20percent is utilised to test the model's performance by comparing projected and actual values.

Logistic Regression: It is the most fundamental model for classifying data and objects.It is termed logistic after the Logistic function, which is utilised at its core. Sigmoid or logistic function.Instead of predicting something continuous like weight, logistic regression predicts whether something is true or untrue.As a result, logistic regression predicts whether the patient will be readmitted to the hospital or not. We got the results observed in Table 1 after applying Logistic regression to the data set. Logistic regression can be used to determine whether a variable's effect on prediction is statistically different from zero; if it isn't, the variable isn't assisting in the prediction, and it can be removed.

Accuracy	Precision	Recall
62%	64%	55%

Table 1: Results of Logistic regression

Decision tree- The Decision Tree is the most extensively used model because it is easy to understand, which allows non experts to grasp how the model makes decisions.Using knowledge acquisition, the decision tree selects its root node and sub nodes until it reaches the goal node/leaf node.Our parent node will contain the most information, while sub nodes/child nodes will contain the least information.This process repeats until the target node has gained no information. Table 2 shows the results we obtained from using the decision tree algorithm on the selected data base.

Accuracy	Precision	Recall
90%	93%	91%

Table 2: Results of Decision Tree

Because decision trees outperformed logistic regression, we can conclude that tree-based models will perform better for our problem.

Random forest -Random forest is an ensemble learning approach that combines multiple weak models into a single powerful model.Because random forest learns from numerous trees, it performs well on very large volumes of data with high dimensionality. Therefore we used the random forest method on our data set and obtained the results shown in Table 3.

Accuracy	Precision	Recall
94%	98%	90%

Table 3: Results of Random forest

Among all of the models we tested, Random Forest fared the best, proving that our premise regarding tree-based models was right.

7. Conclusion & Possible Future Works

In this paper, we discussed the importance of reducing the risk of readmission to hospitals within thirty days of the initial discharge. This could benefit both the hospital and patient in terms of the costs incurred. This study proposes using data mining methods to predict the risk of readmission as a business plan. A database on hospital readmission that is available publicly of between the years of 1999 and 2008. This data set was used for further analysis of the relationship between patient health history and patient readmission. We began by preprocessing the data to reduce the chance of errors, bias and inconsistency. Three data mining techniques were employed; logistic regression, random forest and decision tree. We found that in comparison to the others, random forest produced the best results with 94% accuracy, 98% precision and 90% recall.

In terms of characteristics, the dataset we used to predict hospital readmission was very limited and outdated. More than 90% of data was missing in some attributes which could have been a crucial feature for prediction. Our current model only works for diabetic patients. In the future, with access to better data sets, the model can be trained to give better results. The current model only predicts readmission for diabetes patients. We could also tweak the current model to be able to make predictions for any disease in the future.

References

[1] C. J. Clark, A. Coe, N. F. Fino, and R. Pawa, "Endoscopic retrograde cholangiopancreatography in octogenarians: a population-based study using the nationwide inpatient sample," *Endoscopy International Open*, vol. 4, no. 06, pp. E624–E630, 2016.

[2] S. Jiang, K.-S. Chin, G. Qu, and K. L. Tsui, "An integrated machine learning framework for hospital readmission prediction," *Knowledge-Based Systems*, vol. 146, pp. 73–90, 2018.

[3] S. Cui, D. Wang, Y. Wang, P.-W. Yu, and Y. Jin, *An improved support vector machine-based diabetic readmission prediction*. Elsevier, 2018, vol. 166.

[4] T. Goudjerkan and M. Jayabalan, *Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron*. SAI Organization, 2019, vol. 10, no. 2.