Machine Learning Project

Jagtar Singh

N01476715

1 Abstract

The way hyperglycemia is managed in hospitalised patients has a big impact on how things turn out in terms of morbidity and mortality. There aren't many national studies on diabetes care during hospitalisation, though, so they can't be used as a benchmark for improvement. To give such an assessment and identify potential future approaches that could enhance patient safety, this analysis of a sizable clinical database (74 million unique encounters corresponding to 17 million unique patients) was carried out. There were around 70,000 inpatient diabetes visits that could be analysed. While controlling for factors such demographics, the severity and type of the disease, and the kind of admission, multivariable logistic regression was used to fit the association between the measurement of HbA1c and early readmission.

According to the findings, only occasionally (18.4%) did inpatient settings measure HbA1c. According to the statistical model, the primary diagnosis affects how the HbA1c test and likelihood of readmission are related. The statistics also imply that better patient outcomes and lower inpatient care costs may result from the increased focus on diabetes that is reflected in HbA1c determination.

2 Introduction

The management of hyperglycemia in hospitalised patients is now widely

acknowledged to have a major impact on outcomes in terms of morbidity and mortality. Due to this realisation, structured protocols with strict glucose targets have been developed in the intensive care unit (ICU) environment in many institutions. For most non-ICU inpatient hospitalizations, however, the same cannot be stated. Instead, anecdotal data suggests that when standard management procedures are used, inpatient management is arbitrary and frequently results in either no therapy at all or significant variations in glucose. Recent controlled trials have shown that protocol-driven inpatient techniques can be both successful and safe, despite the paucity of evidence. As a result, it is now advised to develop protocols in the hospital context. However, there aren't many national studies that may be used as a starting point for improvement in diabetes care for hospitalised patients. In order to analyse historical trends of diabetes care in patients with diabetes admitted to a US hospital and to inform future strategies that may improve patient safety, the present analysis of a sizable clinical database was conducted. When many people were identified as having a diagnosis of diabetes mellitus, we specifically looked at the use of HbA1c as a metric of attention to diabetes management. Our hypothesis is that measuring HbA1c is linked to a decline in readmission rates among hospital patients.

In terms of missing values, incomplete or inconsistent records, and high dimensionality—defined not only by the quantity of features but also by their complexity—clinical data databases contain valuable but heterogeneous and challenging data. Additionally, because one has no control over how and what kind of information was gathered, interpreting external data is more difficult than examining the outcomes of a carefully planned experiment or trial. However, it is crucial to make use of these vast amounts of data to discover new knowledge that may be unavailable elsewhere.

3 Methods

Note: for pre-processing I have used Python and for training and evaluation I have used Weka.

First, I preprocessed and analysed the data and dropped some of the attributes like ids which has no meaning to the prediction and evaluation.

Text

Description automatically generated

I also checked the total number of null values in the sample data and if there is any, I replaced it by their mean values. Also, I calculated outlier data points using interquartile range (IQR). The outlier data points when using the IQR are those that fall below Q1-1.5 IQR or above Q3 + 1.5 IQR. The dataset's Q1 and Q3 are its 25th and 75th percentiles, respectively, and the IQR is the interquartile range determined by subtracting Q3 from Q1 (Q3–Q1).

Using above-described approach, 101766 outliers detected in the sample. However, removing outliers is not a good approach always, because some of them could be very meaningful and could impact the outcome of the model. So, I replaced it using the mean values.

Moreover, I also checked if the sample data is skewed or not by counting the number of rows for each label in the target column:

Dataset[“readmitted”].value\_counts()

Graphical user interface, text, application

Description automatically generated

As you can see our sample data is highly skewed towards label ‘No’ and in order to make it I balanced it and took out 12000 sample rows from each, make a new balanced dataset with ratio of 12000:12000:11357.

At last, I standardised the data using StandardScaler from sklearn.preprocessing library and for splitting and training I have used a data mining tool called Weka.



In order to get better result, I did feature selection using PCA and Random Forest, and then again tried the Random Forest, Logistic Regression and KNN models to check the accuracy and test which model works better with the sample data.

4 Results

1 Random Forest

First, we used random forest and the results I got are:

Table

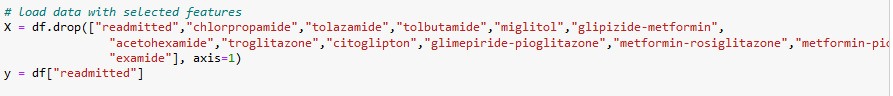
Description automatically generated

We can notice that accuracy is 58% however the recall and fscore is low. So, I tried using the selection features to improve accuracy, fscore and recall using Random Forest and compare the results with the previous one.

Chart, histogram

Description automatically generated

I dropped the features having less than 0.04 score.



However, there is so such significant improvement in the fscore and recall can be seen and also accuracy had been decreased with 1%.

Table

Description automatically generated

2 Knn

Other model I tried is KNN. We first tried with random k=1 and we got better results than Random Forest in the terms of recall and fscore. However accuracy is still far less than Random Forest.

A screenshot of a computer

Description automatically generated with medium confidence

To predic the perfect K, I used the plot Error rate vs Kvalue and check at which error rate is less and I selected 22 as K value.

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

With K =22 we can see that the accuracy had been improved to 40 and also recall and fscore and precision is better.

**Table

Description automatically generated**

3Logistic Regression

At last but not least, I have tried the third model using Logistic Regression and it outperformed other two models with better accuracy and better fscore, precision and recall.

Table

Description automatically generated

**4. Discussion**

**We can conclude that logistic regression worked better with the sample data as we compare it to Random Forest and Knn. We tried to increase the performed using feature selection using Random forest but it comes out PCA worked better.**

**Also, we observed that with this data set where accuracy is important and also with fscore, Logistic regression outperformed other models and is the most suitable for this type of datasets.**

**Dataset Link:** **https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008**