**Name** : Aditya Sumbaraju  
**Date** : 02/20/21  
**Title** : Predicting Hospital Readmission of Diabetic patients

**Section 1**

* **Introduction**

"Diabetes" is a chronic (long-lasting) health condition that affects our body turns food into energy. Diabetes is one of the top ten leading causes of death in the world and the most expensive chronic disease in the United States. Hospitalized patients with Diabetes are at higher risk of readmission than those without Diabetes. Therefore, reducing readmission rates for diabetic patients has a great potential to reduce medical costs significantly. Hence it is crucial to know if a patient will be readmitted to some hospital to change the treatment plan to avoid readmission. This project aims to analyze and predict the likelihood of a diabetic patient being readmitted.

**Research questions**

The significant variables contributing to the analysis are the Number of

* Inpatients
* Length of stay
* Number of Medications
* Number of Diagnoses
* Age

Based on these parameters, I would like to see which one of them affects the Readmission rate. If we can understand the thresholds and come up with a model for such parameters that can predict if the patient is more likely to get readmitted, it will help to take care of them at the right time.

* Is the Quality of care is the main factor to increase the readmission rate?
* Is it possible to identify diabetic patients who have a higher likelihood of being readmitted within 30 days?
* Is it possible to predict Diabetes beforehand?
* **Approach**

Unplanned readmission is the most useful type when evaluating a hospital's Quality of care as it highlights a practitioner's diagnosis or treatment error. Beyond being a core indicator of the Quality of care, unplanned readmissions also constitute a financial problem for nations. Therefore with a predictive model to assess unplanned readmission risk could optimize the Quality of hospital services and state Medicare.

Below are the pointers of my approach:

* Understanding Dataset
* Data Cleansing and Preprocessing need to be performed to remove noisy data and deal with missing values and data inconsistencies.
* Data Reduction – It is essential to optimize the data by reducing uniquely identified categorical variables.
* Remove Data outliers
* Create histograms to analyze the readmitted probability
* Perform Model comparison and its evaluation
* Perform Analytical techniques by splitting the dataset into training Data and Test Data
* **How your approach addresses (fully or partially) the problem.**

In my view, it partially addresses the problem due to the data at hand has a limited time range (1999-2008) and not for the whole world. There may be some influence due to specific local regions due to eating habits, stress levels, or other unknown parameters that are not captured in this dataset.

However, if we get more data from different regions of the world, we may predict the readmittance rate of diabetic patients better. Hence, the Model's accuracy is based on limited data.

I would anticipate that this project may help in further analyzing similar datasets.

* **Data**

The data set represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria.

* It is an inpatient encounter (a hospital admission).
* It is a diabetic encounter, that is, one during which any kind of Diabetes was entered into the system as a diagnosis.
* The length of stay was at least one day and at most 14 days.
* Laboratory tests were performed during the encounter.
* Medications were administered during the encounter.

The data contains such attributes as a patient number, race, Gender, age, admission type, time in the hospital, medical specialty of admitting physician, Number of lab test performed, HbA1c test result, diagnosis, Number of medication, diabetic medications, Number of outpatients, inpatient, and emergency visits in the year before the hospitalization, etc.

**Source Data:**

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

Below are the list of Variables I would be using in this project.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| ***Encounter ID*** | Unique identifier of an encounter |
| ***Patient number*** | Unique identifier of a patient |
| ***Race Values*** | Caucasian, Asian, African American, Hispanic, and other |
| ***Gender Values*** | male, female, and unknown/invalid |
| ***Age*** | Grouped in 10-year intervals 0, 10), 10, 20), …, 90, 100) |
| ***Weight*** | Weight in pounds |
| ***Admission type*** | Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available |
| ***Discharge disposition*** | An integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available |
| ***Admission source*** | Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital |
| ***Time in hospital*** | An integer number of days between admission and discharge |
| ***Payer code*** | An integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical |
| ***Medical specialty*** | Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon |
| ***Number of lab procedures*** | Number of lab tests performed during the encounter |
| ***Number of procedures*** | Numeric Number of procedures (other than lab tests) performed during the encounter |
| ***Number of medications*** | Number of distinct generic names administered during the encounter |
| ***Number of outpatient visits*** | Number of outpatient visits of the patient in the year preceding the encounter |
| ***Number of emergency visits*** | Number of emergency visits of the patient in the year preceding the encounter |
| ***Number of inpatient visits*** | Number of inpatient visits of the patient in the year preceding the encounter |
| ***Diagnosis 1*** | The primary diagnosis (coded as first three digits of ICD9); 848 distinct values |
| ***Diagnosis 2*** | Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values |
| ***Diagnosis 3*** | Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values |
| ***Number of diagnoses*** | Number of diagnoses entered into the system 0% |
| ***Glucose serum test*** | the result would indicate the range of the result or if the test was not taken. Values ">200,” ">300,” "normal," and "none" if not measured |
| ***A1c test result*** | Indicates the range of the result or if the test was not taken. Values ">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. |
| ***Change of medications*** | Indicates if there was a change in diabetic medications (either dosage or generic name). Values "change" and "no change." |
| ***Diabetes medications*** | Indicates if there was any diabetic medication prescribed. Values: "yes" and "no" 24 features for medications For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examine, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride- pioglitazone, metformin-rosiglitazone, and metformin- pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "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 |
| ***Readmitted*** | Days to inpatient readmission. Values: "<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 |

* **Required Packages**

In this project, I will be utilizing several libraries and packages, some provided with R and some additional for data cleansing, plotting the data and visualizations, comparing the models, etc.

Based on what plots and models, I may use the below list of packages and more.

1. dplyr
2. skimr
3. stringr
4. psych
5. ROSE
6. ggplot2
7. caret

* **Plots and Table Needs**

Below is the list of plots and tables that I am assuming will be needed to do the analysis and come up with a model

* Histogram
* Scatter plots for correlation
* Prediction Plots
* Accuracy Tables
* **Questions for future steps.**

1. Is this data accurate legitimate enough to proceed with predictive analytics?
2. Is Gender a predictor variable?