Northeastern University
CS6220 – Data Mining Techniques
Fall 2017, Harsha Jakkappanavar, Jenny Johns, Savan Patel, Chang Zhou

Project Report: Study of Dataset on Diabetes

Introduction/Background

Millions of people around the world are affected by Diabetes. It occurs when the blood glucose level is elevated over time. The dataset we chose spanned from the years 1999-2008, (about 10 years), and includes patient records of patient care over 130 hospitals in the United States. The dataset was comprised of 101,766 patient encounters that fulfilled the following criteria:

1. If it was an inpatient encounter.
2. If the patient was diagnosed with diabetes during their hospital stay.
3. If the patient stayed at the hospital from 1 - 14 days.
4. If the patient participated in laboratory tests during their stay.
5. If the patient was given medication during their stay.

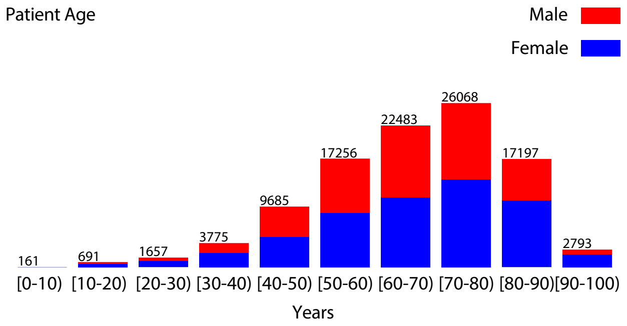
There were 50 different attributes that each record had including: race, gender, admission type, and time spent in the hospital. During our analysis of the dataset, some of these attributes were dropped because they did not contribute to our analysis or they were missing a significant portion of data. For example, encounter ID and patient number were unique numbers associated with each patient, so they would not have any value to our analysis, or with the weight attribute, it was missing 97% of the data, and therefore could not be included.

A paper based on this dataset, written by Strack, Gennings, et al., titled, 'Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records'. They came to the conclusion that there was a correlation between readmission and whether the HbA1c test was performed rather than the value of the test. They believed that measure the HbA1c would also show the effectiveness of the current treatment of care that the patient was receiving as well as lower the cost of care for those patients.

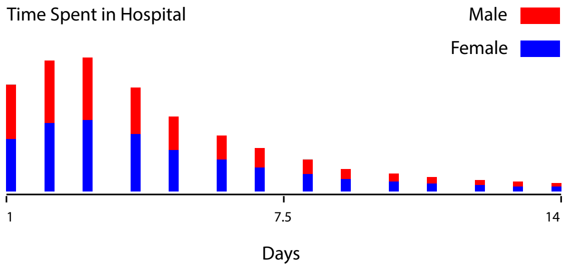
Exploratory Analysis

During the exploratory analysis of this dataset, we found that patients were categorized into one of five possible races, including, African American, Asian, Caucasian, Hispanic, or Other. A majority of the dataset is seen to be of Caucasian race with 76,099 out of 101,767 patient records. The second largest group were African Americans, third Hispanics, fourth Others, and finally Asians. Some of the patient records in the dataset were missing values for the feature of race.

Gender and age were also features of patient records. Gender was broken down into three categories: Male, Female, Unknown/Invalid. A majority of patients were recorded to be of the female gender with a count of 54,708 versus 47,055 recorded male patients. Age was a range from 0-100, broken down into 10 year intervals. The largest subset of the diabetic patients were in the age range of [75-80) years old with a count of 26,068 patients. One of the features of the data was weight, which was missing 97% of the data. Those patients who did have their weight recorded landed in the range of [75-100) pounds.

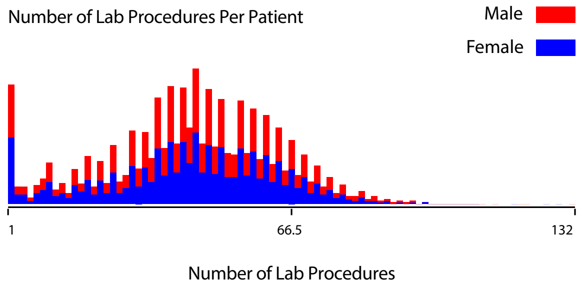


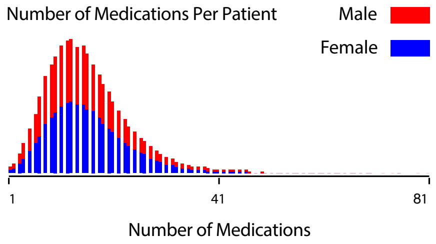
Each patient that was admitted to a hospital, was given a admission type. Within the dataset, these admission types were coded as number from 1 to 8. The different types included emergency admission (1), urgent admission (2), elective admission (3), newborn admission (4), admission type not available (5), NULL (6), admission to the trauma center (7), admission type not mapped (8). The most common admission type was an emergency admission to the facility (code: 1). The time patients spent in the hospital ranged from a minimum of 1 day to a maximum of 14 days, with the average time being about 4 days.



Patients also held a feature of payer code which corresponded to whether they used insurance to pay for their stay, or it was paid out of pocket. With a total of 23 possible different codes, including insurances such as Blue Cross/Blue Shield and Medicare, and missing 40% data, a majority of the instances (32,439 patients), were coded as using Medicare. This correlates with the fact that a majority of patients were over 65.

Each patient also had a record of the number of lab procedures performed during their hospital stay. Every patient had at least one lab procedure, the maximum number of lab procedures being 132, and the average being 43 lab procedures per patient. There was also an account of other procedures performed (excluding lab procedures), not every patient had one of these procedures. The maximum number of procedures was 6, with an average of about 1.34 procedures per patient. Each patient was administered at least one medication during their hospital stay, with the average number of medications administered being 16, and the maximum number being 81 medications in one patient encounter.





Each patient had three diagnosis features. The first being the primary diagnosis, the secondary diagnosis and the third being an additional secondary diagnosis. Each diagnosis is a code from the icd9, each mapping to a specific illness. The majority of the primary diagnoses were from ”heart failure” which corresponds to an icd9 code of 428, affecting 6,862 patients. From the secondary diagnosis, a majority of the diagnoses were from ”disorders of fluid, electrolyte, and acid-base balance” which corresponds to an icd9 code 276, affecting 6,752 patients. From the additional secondary diagnosis, the majority of diagnoses were ”diabetes mellitus”, which corresponds to an icd9 code of 250, affecting 11,555 patients. These diagnoses were a nominal value, and had many levels (each a different icd9 code), so the codes were then grouped based on which system of the body they affected, for example, code 428 (heart failure), would be grouped as circulatory. The different categories were circulatory, respiratory, digestive, diabetes, injury, musculoskeletal, genitourinary, neoplasms, and other. Circulatory was the most highly affected system for all three diagnoses, followed by other. The figures showing these three diagnoses can be found in the appendix.

In order to test for diabetes, the hospital could have performed a glucose serum test on the patient. A glucose serum test shows the level of glucose in the blood of the patient at that time. The different values that this feature can have are ’None’, ’>300’, ’Norm’, and ’>200’. A majority of values are ’None’ which indicates that the glucose test was not performed on this patient. The ’None’ value has a count of 96,420 patients, followed by 2,597 patients being tested and returning a test result of normal. There is a similar circumstance with the A1c results of patients, where a majority of patients (84,748) did not have the test performed. The different values of the A1c test can be ’None’, ’>7’, ’>8’, and ’Norm’. After the ’None’ value the highest count was 8,216 patients who were tested and had a result of ’>8’, which indicates that the test result shows that the A1c was greater than 8%. This means that at least 8% of that patient’s red blood cells have glucose attached to them, and indicates that the patient has probably had uncontrolled diabetes for quite a while.

There are 24 different features for diabetic medications that can be administered during a patient’s hospital stay. The four possible values for each of these features were ’up’, ’down’, ’steady’, or ’no’. For all the medications, the value with the highest percentage was ’no’ meaning that the medication was never prescribed. Along with these 24 medications, the dataset included a feature which indicated whether there was a change in the diabetic medications taken by that patient, the possible values being ’change’ or ’no change’. Records were also kept about a patient’s readmission to a hospital, values including ’¡30’, meaning that the patient was readmitted within less than 30 days of their previous hospital stay, '¿30', if the patient was readmitted after 30 days of their previous hospital stay, or 'no' is there was no record of their readmission.

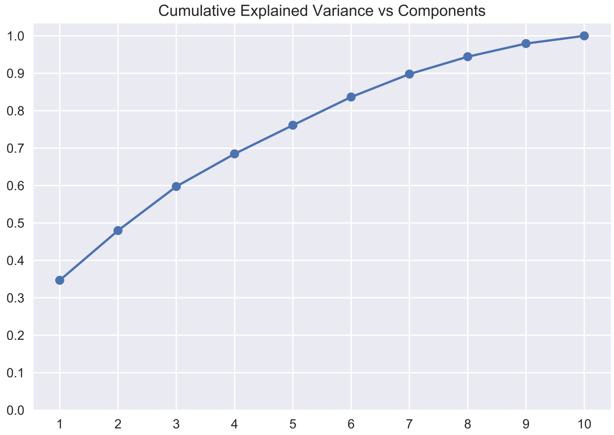
Data Mining Analysis

Readmission Prediction Model

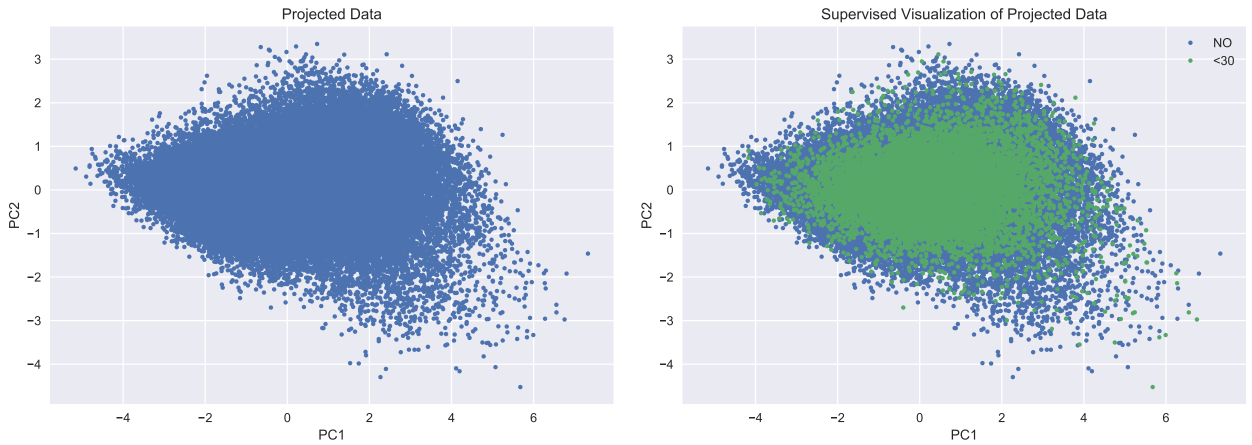
A logistic regression model was built in order to predict whether a patient has a high probability of readmission within less than 30 days of their last encounter. The attributes used were number of medications, time spent in the hospital, number of lab procedures, number of procedures, number of outpatient visits, number of emergency visits, number of inpatient visits, and number of diagnoses. After dropping the unnecessary columns, the data was grouped based on whether the patient was readmitted or not. After selecting training and testing data, using sklearn linear model's logistic regression, the model was able to predict with about 54.88% accuracy. PCA was then used to reduce the original dimensionality of the data to 2. After retraining the data, and following the previous steps with sklearn, the model delivered a slightly improved accuracy of 55.05%.

Singular-value Decomposition

Attributes used in SVD were age, admission type, discharge disposition, time spent in the hospital, number of lab procedures, number of medications, diagnosis 1, diagnoses 2, diagnosis 3 and readmitted. The three diagnoses were changed from the icd9 code to being grouped by which body system they affected. Then for age, the three diagnoses, and readmission, the categorical features were changed to numbers. Visualizations were produced for these attributes based on readmission values, 'NO' or '<30', meaning that the patient was either not readmitted or readmitted within less that 30 days after the last recorded hospital stay. Sklearn preprocessing was used to standardize the data in order to fit and then transform the data. The covariance matrix was then calculated for use with SVD. The cumulative explained variance per component that was computed is represented in the following figure:



Seven components were needed to explain 90% of the variance, which was used to apply PCA to the data. The eigenvectors produced by the PCA were multiplied by the standardized values previously calculated which producing the plot for the projected data, which was then overlayed with the readmission data to produced the supervised visualization.



TALK ABOUT INTERPRETING THE COMPONENTS

Classification Using Random Forests

FP-Growth

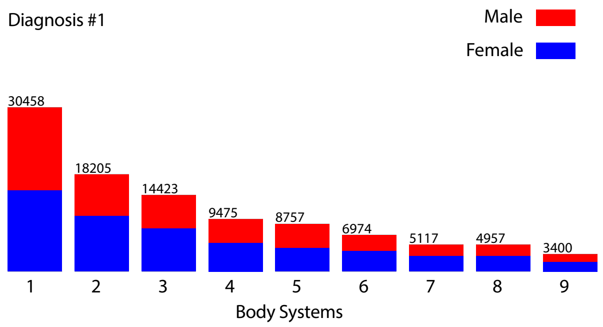
FP-Growth was used to find any associations between the 24 different types of medications. The value was transcribed as up if the dosage of that particular medication was increased during that patients hospital stay, down if it was decreased, steady of the dosage wasnt changed, and no if it was never prescribed. We specifically only looked at the first 18 features of medications because the last six are combinations of those medication features. After converting this data to the ARFF format, it was run through wekas FP Growth algorithm, which was able to produce two association rules:

$$[i1 = 1] : 19988 \longrightarrow [i18 = 1] : 10012 < conf : (0.5) > lift : (0.72) lev : (-0.05) conv : (0.61)$$
$$[i18 = 1] : 54383 \longrightarrow [i1 = 1] : 10012 < conf : (0.18) > lift : (0.72) lev : (-0.05) conv : (0.91)$$

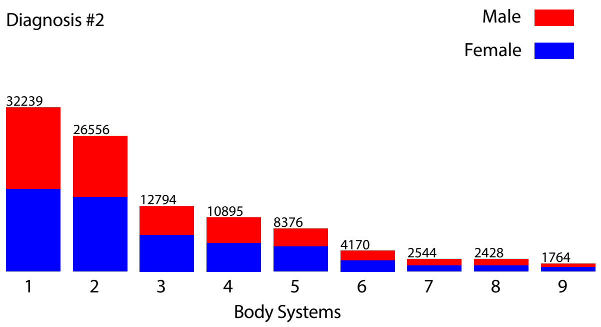
The rules are showing that the medication metformin, (i1), and insulin, (i18), have an association between them. From the first rule, we can see that the confidence of insulin occurring given metformin is 0.5 and from rule two, we can see that the confidence of metformin occurring given insulin is 0.18. These confidences are not high, which means that there is not really a strong association between these medications. And we can see that the lift is 0.72, and since in lift a 1 means they are independent, this shows that the two medications, metformin and insulin, are pretty close to occurring independently from each other. **ADD MORE ABOUT WHY**

Discussion

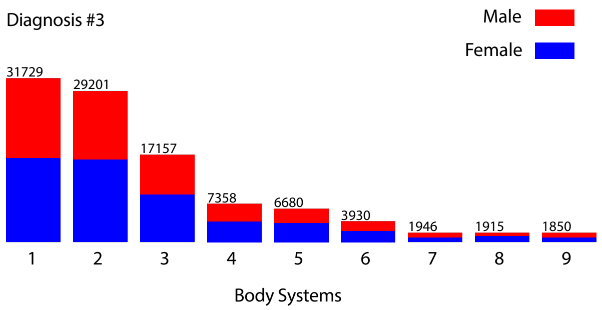
Appendix



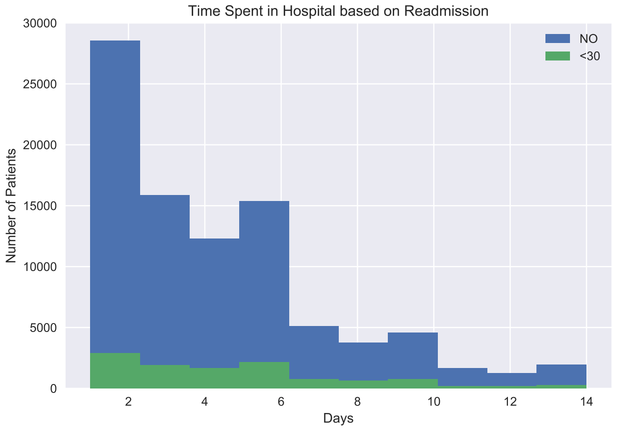
1. Circulatory
2. Other
3. Respiratory
4. Digestive
5. Diabetes
6. Injury
7. Genitourinary
8. Musculoskeletal
9. Neoplasms

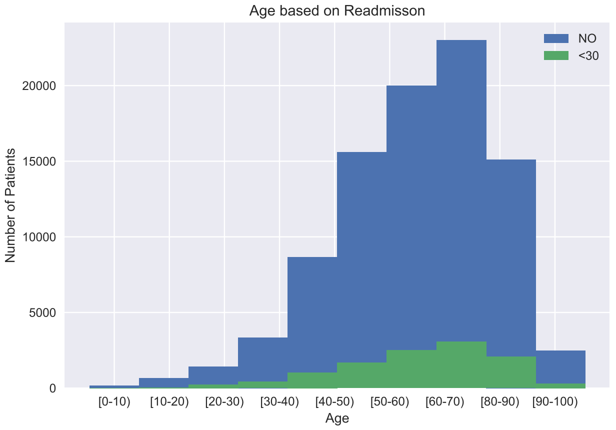


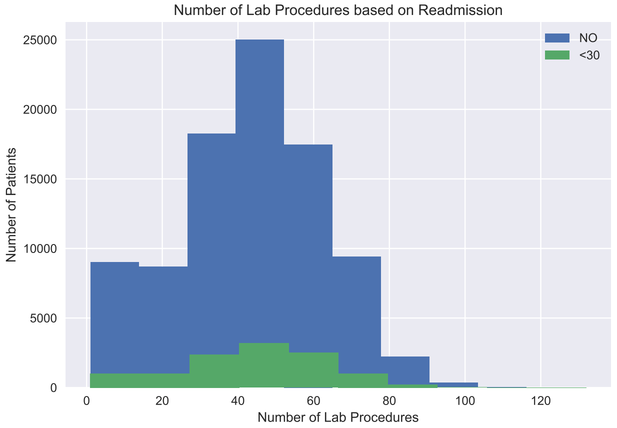
1. Circulatory
2. Other
3. Diabetes
4. Respiratory
5. Genitourinary
6. Digestive
7. Neoplasms
8. Injury
9. Musculoskeletal

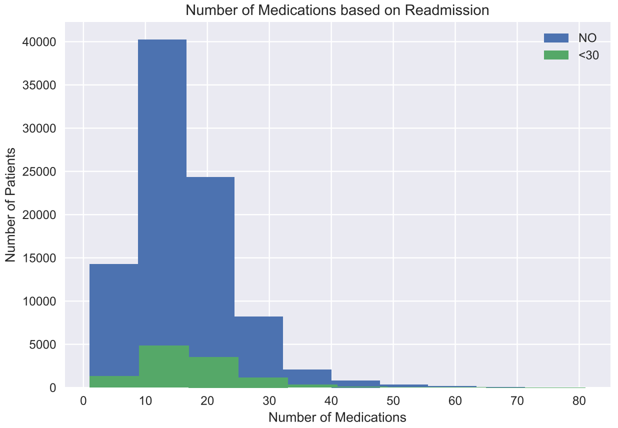


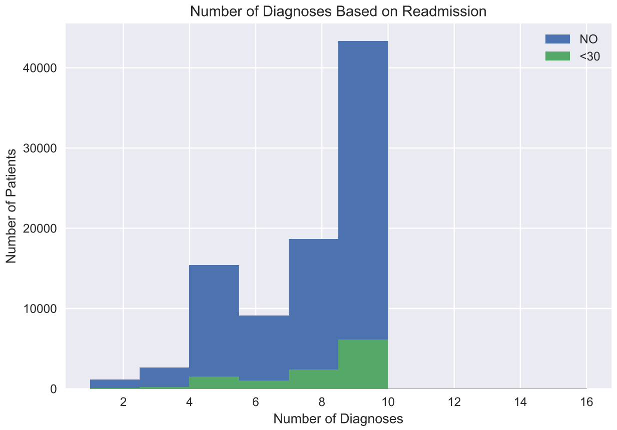
1. Circulatory
2. Other
3. Diabetes
4. Respiratory
5. Genitourinary
6. Digestive
7. Injury
8. Musculoskeletal
9. Neoplasms


Time Spent in Hospital based on Readmission
Number of Patients
Days
NO
<30
0
5000
10000
15000
20000
25000
30000
2
4
6
8
10
12
14


Age based on Readmisson
Number of Patients
Age
NO
<30
0
5000
10000
15000
20000
[0-10)
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[80-90)
[90-100)


Number of Lab Procedures based on Readmission
Number of Patients
Number of Lab Procedures
NO
<30
0
5000
10000
15000
20000
25000
0
20
40
60
80
100
120


Number of Medications based on Readmission
Number of Patients
Number of Medications
NO
<30
0
5000
10000
15000
20000
25000
30000
35000
40000
0
10
20
30
40
50
60
70
80


Number of Diagnoses Based on Readmission
Number of Patients
Number of Diagnoses
NO
<30
0
10000
20000
30000
40000
2
4
6
8
10
12
14
16