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, which occurs when the blood glucose level is elevated over time. The dataset we chose spans from the years 1999-2008, (about 10 years), and includes patient records from over 130 hospitals in the United States. The dataset is comprised of 101,766 patient encounters that fulfill 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.

From a set of 55 different attributes, each record included: 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. Similarly, the weight attribute was dropped, as it was missing 97% of the data.

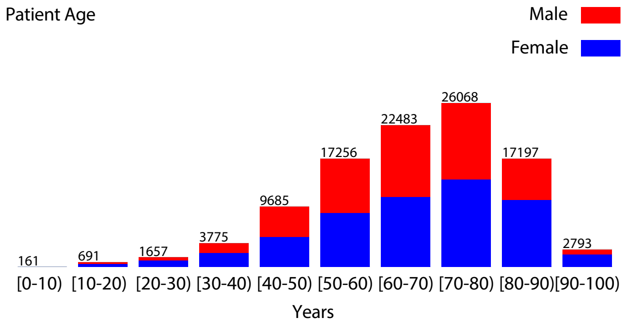
In a paper based on this dataset, written by Strack, Gennings, et al., titled, 'Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records', the authors found a correlation between readmission and whether the HbA1c test was performed rather than the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

The questions we hoped to answer were, if given certain attributes from a patient record, could we predict whether that patient would be readmitted, and were there any association rules between 18 medications administered to patients during their hospital stay?

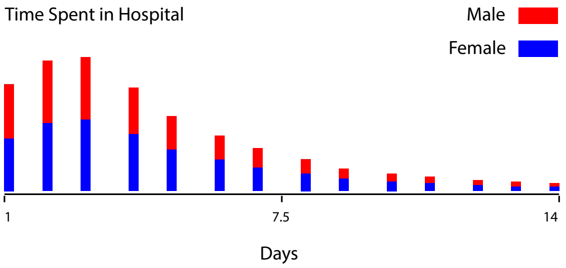
Exploratory Analysis

During the exploratory analysis of this dataset, we found that patients were categorized into one of five possible races: African American, Asian, Caucasian, Hispanic, or Other. A majority of the dataset, 76,099 out of 101,767 patient records were Caucasian. 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 as female (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 diabetic patients was in the age range of [75-80) years old with a count of 26,068 patients. Although weight was largely missing in the dataset, those patients who did have their weight recorded landed in the range of [75-100) pounds.

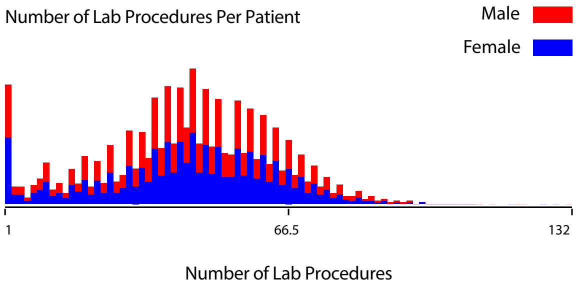


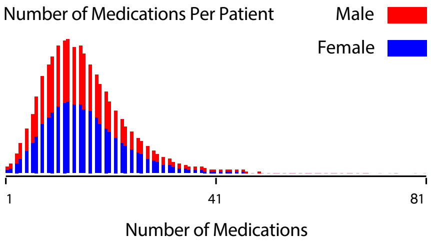
Each patient admitted to a hospital, was given an admission type. Within the dataset, these admission types were coded as numbers from 1 to 8. The different types included (1) emergency admission, (2) urgent admission, (3) elective admission, (4) newborn admission, (5) admission type not available, (6) NULL, (7) admission to the trauma center, (8) admission type not mapped. The most common admission type was 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 approximately 4 days.



Patients were also given a payer code which corresponded to whether they used insurance to pay for their stay, or if 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 remaining 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, with 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), and not every patient had one of these other 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 being 81 medications in one patient encounter.





Three features were include in every patient diagnosis. The first of these was a primary diagnosis, followed by a secondary diagnosis, and, finally, an additional secondary diagnosis. Each diagnosis is represented by a code from the icd9, that maps to a specific illness. The majority of primary diagnoses were from ”heart failure” which corresponds to an icd9 code of 428, affecting 6,862 patients. For secondary diagnoses, a majority of were ”disorders of fluid, electrolyte, and acid-base balance” which corresponds to icd9 code 276, affecting 6,752 patients. For additional secondary diagnoses, the majority of these 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. An example of this is code 428 (heart failure), would be grouped as circulatory. The different code categories are as follows: circulatory, respiratory, digestive, diabetes, injury, musculoskeletal, genitourinary, neoplasms, and other. Circulatory was the most commonly affected system across the three diagnoses, with the ’other’ being the second most common. The figures showing these three diagnoses can be found in the appendix of this report.

One method used by hospitals to test for diabetes is the glucose serum test. When performed on the patient, the glucose serum test shows the level of glucose in the blood of the patient at that time. In hospital records, the different values recorded for this test are ’None’, ’>300’, ’Norm’, and ’>200’. A majority of patient records in the dataset for this report (96,420) indicated ’None’, showing that the glucose serum test was not frequently administered to patients. A majority of the remaining patients who did receive the glucose serum test returned normal results. There is a similar circumstance with the A1c test results, 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 next highest results category was ’>8’,representing 8,216 patients and indicating the test result showing 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, indicating 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, with the possible values being ’change’ or ’no change’. Records were also kept about a patient’s readmission to a hospital, where a value '<30' indicates that the patient was readmitted within less than 30 days of their previous hospital stay. If the patient was readmitted after 30 days of their previous hospital stay, a value of '>30' was recorded. Otherwise, 'no' was recorded, indicating that no records exist of that patients readmission.

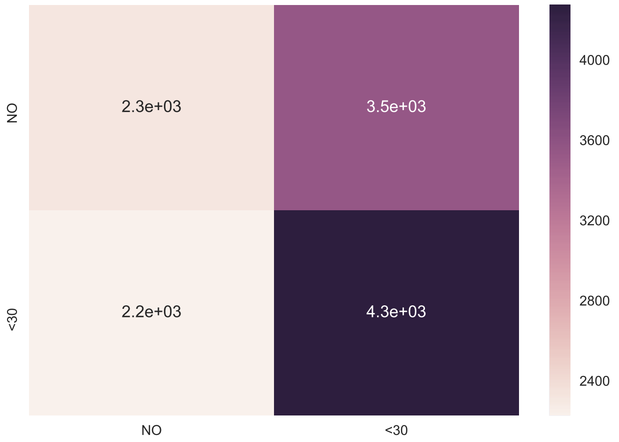
Data Mining Analysis

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 the sklearn linear model's logistic regression, the model was able to predict with about 50.41% 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 56.75%.

Classification Using Random Forests

Classification using random forests was also used to predict whether a patient had a probability of being readmitted. After preprocessing the data, it was then split into 'positive' and 'negative' readmission values, with positive being readmission within less that thirty days of the last hospital stay, and negative being no recorded readmission. There were approximately 12,000 positive values, and consequently the same number were randomly sampled from the negative values as well. The data was then combined to form the training and testing datasets. The sklearn ensemble was used to perform Random Forest Classifier on the training dataset, after which the accuracy score from sklearn metrics was used and produced that the model had an approximate accuracy of 51.90%. A confusion matrix was produced to show the accuracy of the algorithm.



The diagonals of the confusion matrix show the true positives. For about 2,300 predictions of 'No' the model was correct, and for about 4,300 predictions of '¡30' readmission, the model was correct.

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' if the dosage was not 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 weka's FP Growth algorithm using a confidence level of 0.1, 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 medications 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. Another reason these two could be associated is the fact that metformin is a diabetic medication that can be used to increase insulin sensitivity and could then be used in conjunction with insulin to treat Type 2 diabetes (Sloane, 2014). The fact that we see patients taking metformin with insulin can enhance the effect of glycemic control and reduce the dosage of insulin (Wulffel, Kooy, Lehert, et al., 2002). The other association rule with insulin and metformin, with a confidence level of 0.18, may occur because insulin can be used for Type 1 and Type 2 diabetes. In Type 1 diabetes, the patients cannot produce insulin, so while they may still be sensitive, they need to replace it with the medication, and do not need the metformin to increase their sensitivity.

Discussion

The questions posed about this dataset were if there was a model that could predict readmission rates based on the given attributes and any association rules between the 18 distinct diabetic medications administered to the patients during their hospital stay. Two interesting association rules were found using weka's FP-Growth algorithm that showed an association with insulin and metform. Research showed these two medications to be administered together to enhance the effects of insulin by increasing a patient's sensitivity to it using metformin. Two models were used to predict readmission rates: logistic regression and random forests. Both showed an accuracy of a little more than 50%.

Working with the dataset, showed that it was geared more towards supervised learning, which is why the models were used for prediction, yet, unsupervised methods were used improve the accuracy of those models. For example, within the logistical regression model, PCA was used to reduce the number of dimensions which was able to increase the accuracy from around 50% to around 57%.

Given more time, we would like to try to apply clustering algorithms to this dataset which proved difficult because there was such a large number of nominal values. The paper which used this dataset stated that there might have been a correlation between readmission and whether the HbA1c test was performed, rather than the actual value of the test. We could try to incorporate more datasets to include the clustering algorithms and see if we could find similar results between the two datasets using the same models/algorithms on all of them.

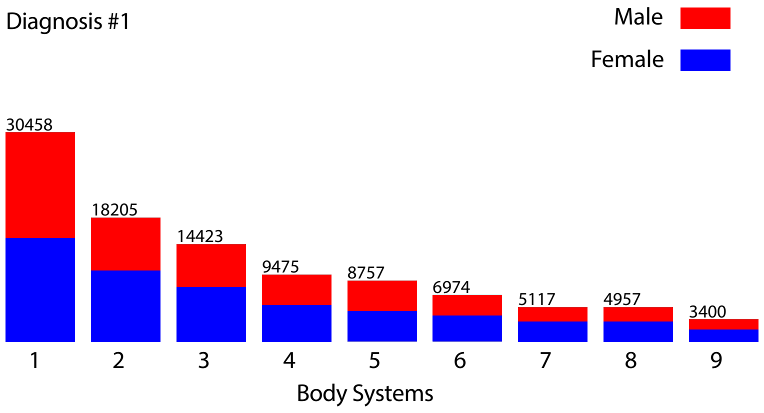
Bibliography

Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records, BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. doi:10.1155/2014/781670

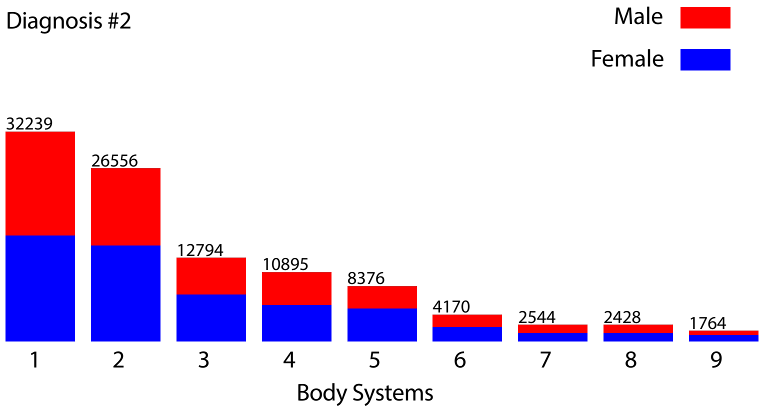
Michiel G. Wulffel, Adriaan Kooy, Philippe Lehert, et al., "Combination of Insulin and Metformin in the Treatment of Type 2 Diabetes," Diabetic Care, vol. 25, 7 pages, 2002, doi:10.2337/diacare.25.12.2133

Using Random Forests in Python with Scikit-Learn. (2017, July 27). Retrieved December 13, 2017, from http://www.blopig.com/blog/2017/07/using-random-forests-in-python-with-scikit-learn/

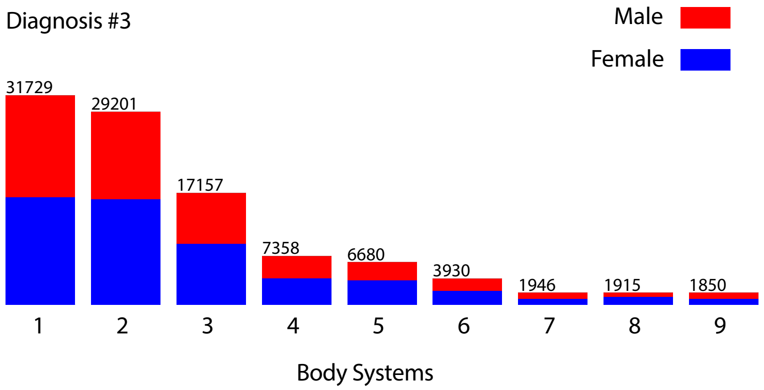
Appendix


Diagnosis #1

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


Diagnosis #2

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