Project Report

Aim:

To examine the association of in-hospital diabetes patients with subsequent 30-day risk for unplanned readmission/emergency department admission.

Dataset Source:

Diabetes 130-US hospitals for years 1999-2008 Data Set

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

Dataset Information:

Data Set Characteristics: Multivariate
Attribute Characteristics: Integer
Number of Instances: 101767
Number of Attributes: 50
Missing Values: Yes

Introduction:

A hospital readmission is when a patient who is discharged from the hospital, gets re-admitted again within a certain period of time. Hospital readmission rates for certain conditions are now considered an indicator of hospital quality, and also affect the cost of care adversely. For this reason, Centre's for Medicare & Medicaid Services established the Hospital Readmissions Reduction Program which aims to improve quality of care for patients and reduce healthcare spending by applying payment penalties to hospitals that have more than expected readmission rates for certain conditions. Being able to determine factors that lead to higher readmission in such patients, and correspondingly being able to predict which patients will get readmitted can help hospitals save millions of dollars while improving quality of care.

Column Description:

Feature Name	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, new-born, and not available
Discharge Disposition	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	Integer number of days between admission and discharge
Payer code	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
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	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 to the system
Glucose serum test result	Indicates 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, examide, 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.

Knowing the Dataset:

1. We started our dataset with finding the number of columns and number of rows.

```
nrow(diabetic)
101766
ncol(diabetic)
50
```

2. Now we structured the dataset and find the type of the variables.

```
$ encounter_id            : int  2278392 149190 64410 500364 16680 35754 5
5842 63768 12522 15738 ...
 $ patient_nbr             : int  8222157 55629189 86047875 82442376 42519
267 82637451 84259809 114882984 48330783 63555939 ...
 $ race                    : Factor w/ 5 levels "AfricanAmerican",..: 3 3
1 3 3 3 3 3 3 3 ...
 $ gender                  : Factor w/ 3 levels "Female","Male",..: 1 1 1
2 2 2 2 2 1 1 ...
 $ age                     : Factor w/ 10 levels "[0-10)","[10-20)",..: 1
2 3 4 5 6 7 8 9 10 ...
 $ weight                  : Factor w/ 9 levels "[0-25)","[100-125)",..: N
A NA NA NA NA NA NA NA NA NA ...
 $ admission_type_id       : int  6 1 1 1 1 2 3 1 2 3 ...
 $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
 $ admission_source_id     : int  1 7 7 7 7 2 2 7 4 4 ...
 $ time_in_hospital        : int  1 3 2 2 1 3 4 5 13 12 ...
 $ payer_code              : Factor w/ 17 levels "BC","CH","CM",..: NA NA
………….
```

3. We also concluded the X-Variables and Y-Variable from the dataset.

Pre-processing of Data:

1. Dealing with missing values:

a) First, we have to see how many missing values are (which were coded as "?" for most variables in the data)

```
race 2273
weight 98569
payer_code 40256
medical_specialty 49949
diag_1 21
diag_2 358
diag_3 1423
gender 3
```

- Weight is missing in over 98% records. Owing to the poor interpretability of missing values and little predictive generalizability to other patients, best thing is to just drop it.
- Payer code and Medical Specialty of treating physician also have 40–50% missing values. We decided to drop these.

b) Also, one more cleaning step that depends on understanding the data, since we are trying to predict readmissions, those patients who died during this hospital admission, have zero probability of readmission. So we should remove those records (discharge_disposition = 11, 19, 20).

c) We also noticed that for two variables (drugs named citoglipton and examide), all records have the same value. So essentially these cannot provide any interpretive or discriminatory information for predicting readmission, and we dropped these columns as well. Technically, this isn’t a missing value problem but rather a missing information problem.

2. Exploratory Data Analysis:

a) Fixing of missing values:

1. Race (Column):

Checked the total number of NULL present in the race column.

```
race
2273
```

We converted NULL values to “Others” in the race column as the data was not available for region, hence we considered it in the “Others” region to eliminate the data loss.

```
Other
3725
```

b) Dropping Columns:

1. encounter_id
It is unique identifier of an encounter, it will not be required in any of the analysis.

2. patient_nbr
It is Unique identity of a patient, hence will not be required.

3. weight
Weight is missing in over 98% records. Owing to the poor interpretability of missing values and little predictive generalizability to other patients, best thing is to just drop it.

4. payer_code
Payer code of treating physician also have 40–50% missing values. We decided to drop these.

5. medical_specialty
Medical Specialty of treating physician also have 40–50% missing values. We decided to drop these.

6. citoglipton
All records have the same value. Incorrect data, hence we dropped this column.

7. examide
All records have the same value. Incorrect data, hence we dropped this column.

8. discharge_disposition_id = 11
We are trying to predict readmissions, those patients who died during this hospital admission, have zero probability of readmission. So, we should remove those records.

Graphical Representation:
a) Variable Distributions

1. Age Distribution

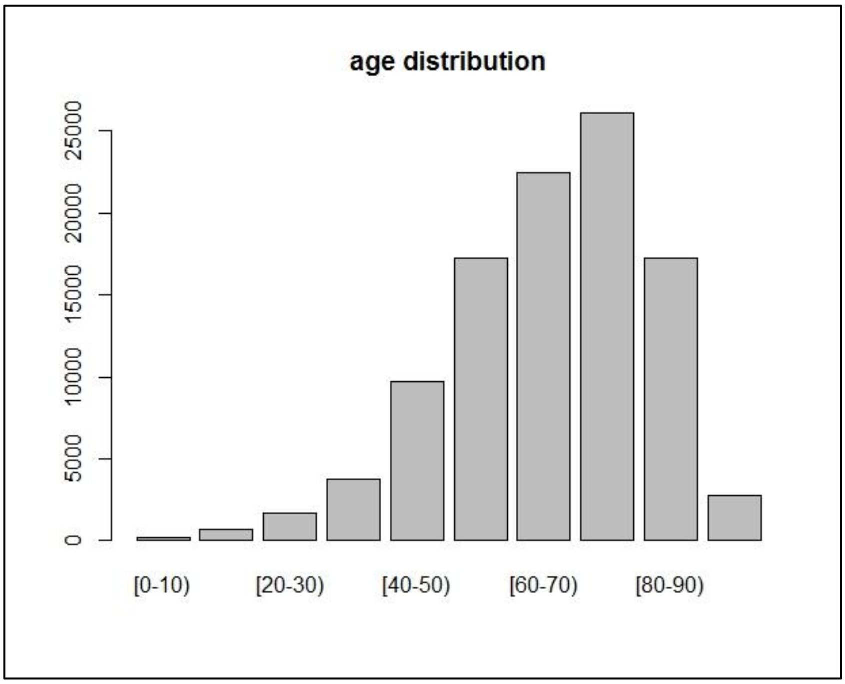


Fig 1: Age Distribution

We can see that from age group 70-80 there are maximum number of patients, followed by the age group 60-70. Age groups 40-50 and 80-90 have almost same number of patients. We can also see the normal distribution is right skewed.

2. Gender Distribution

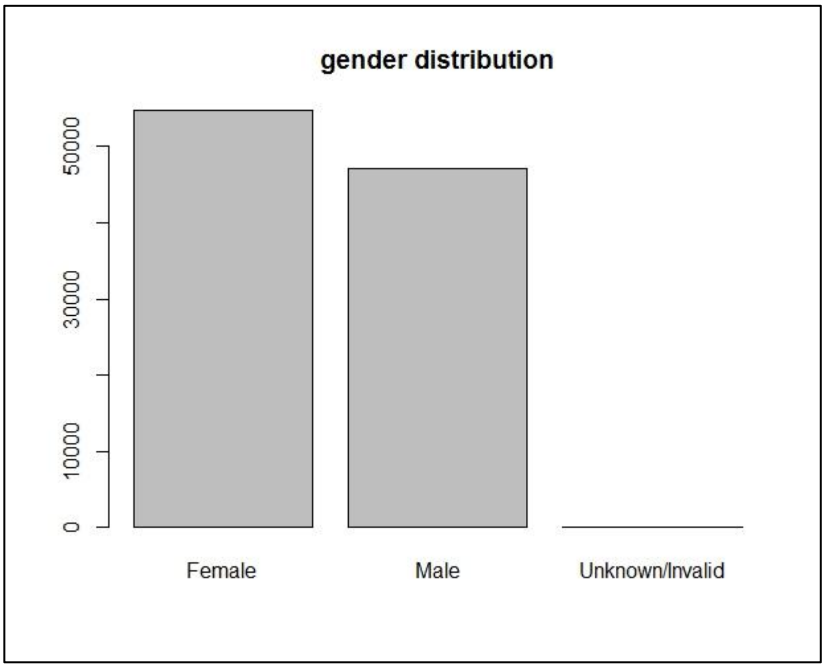


Fig 2: Gender Distribution

We checked the gender distribution, by getting the results gender distribution was 53% were females, 46% were males and only 0.002% were the unknown/invalid.

3. A1c Test Result Distribution

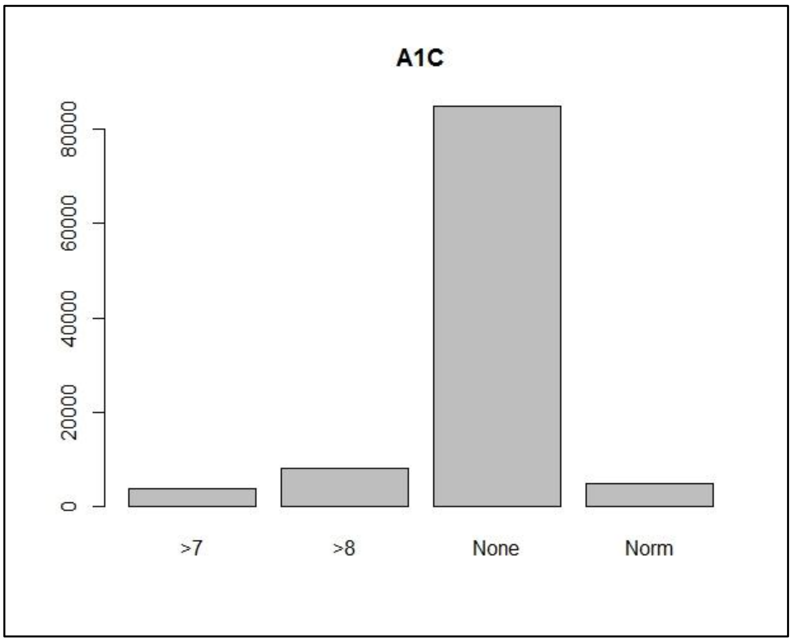


Fig 3: A1C Test Result

A1C test results indicates the range of the result or if the test was not taken. Graph shows 84% patients test was not taken.

4. Readmitted Distribution

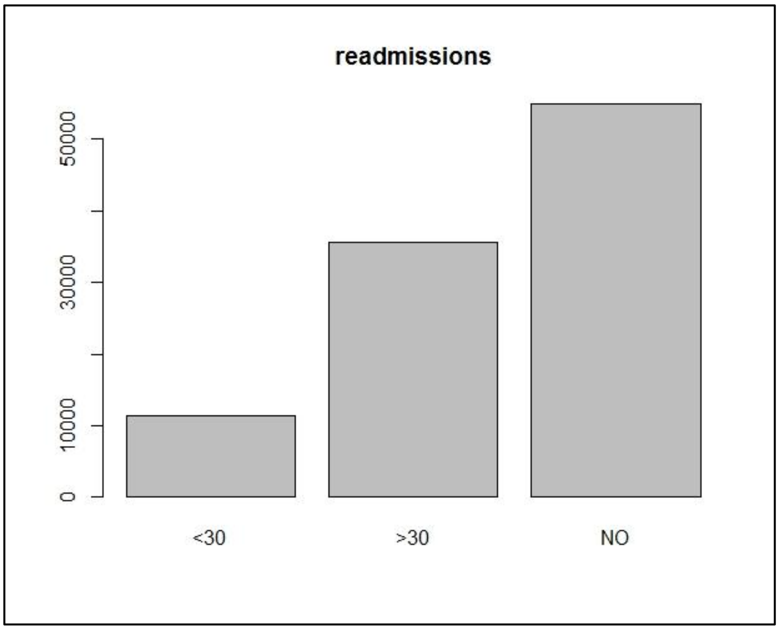


Fig 4: Readmission Distribution

The graph shows more than 53% of patients were not readmitted in the hospital. 34% of the patients where readmitted after 30 days, and 11% of patients where readmitted before 30 days.

b) Some more graphical representation

1. Time_in_hospital vs Age

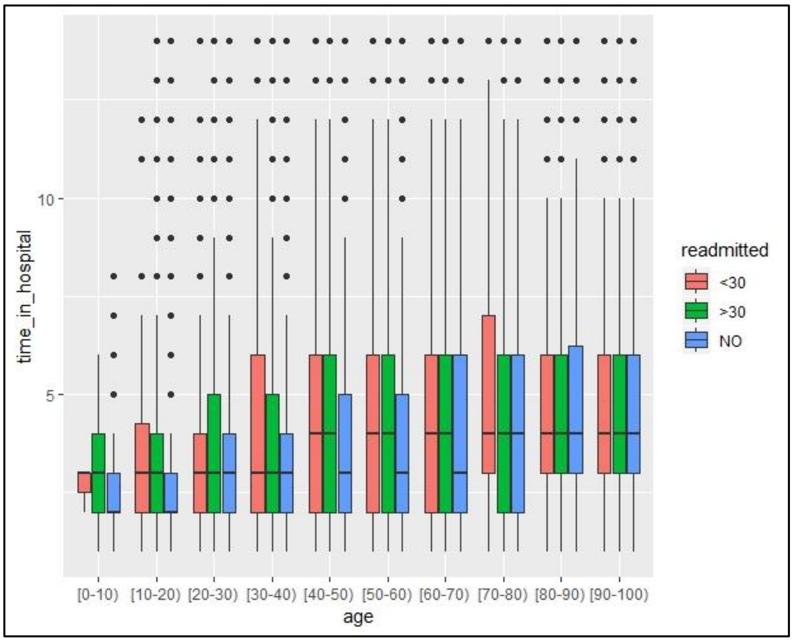


Fig 5: time_in_hospital vs age

From this graph we can see that patients with <30 days readmission from the age group 70-80 had spend longer time in the hospital than any patients else. Also, patients in the age group 30-40 <30 days readmission have spent longer time in the hospital. Age group 60-70 has spend almost same time in the hospital for >30, <30 and no readmission. The age group 20-30 has spend more time in the hospital after >30 readmission.

2. A1C Results vs Num_Medication

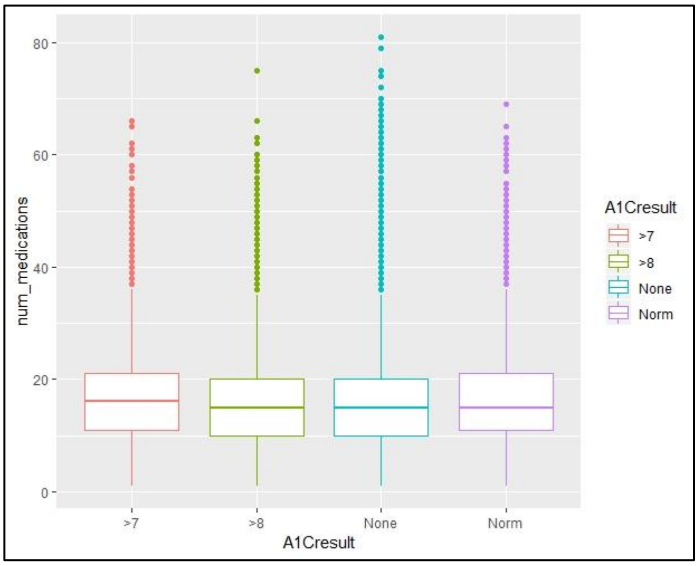


Fig 6: A1C Result vs num_medication

This graph shows the same similarities as the of A1C test results distribution.

3. A1C test result vs time_in_hospital

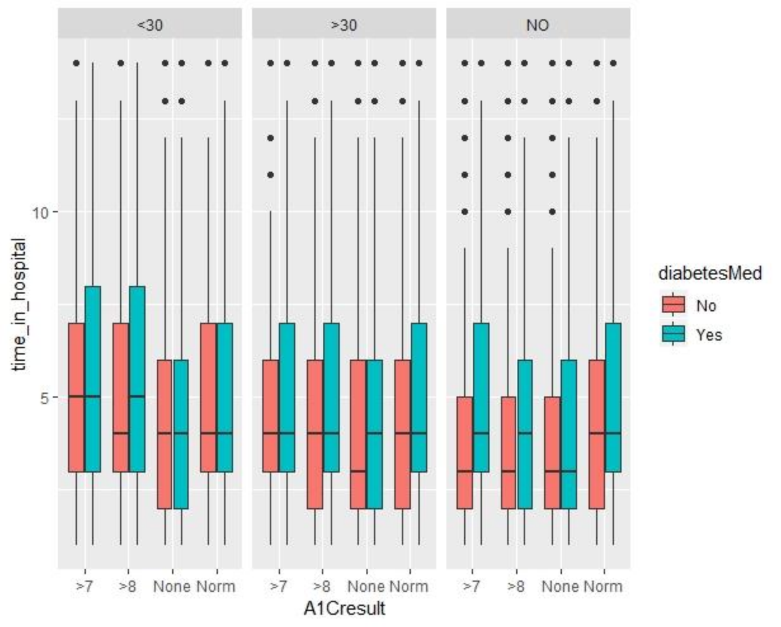


Fig 7: A1C test result vs time_in_hospital

This graph shows that patients with no readmission had generally less time in hospital. Patients with <30 readmission has generally more time in the hospital.

4. Age vs. Num_medication

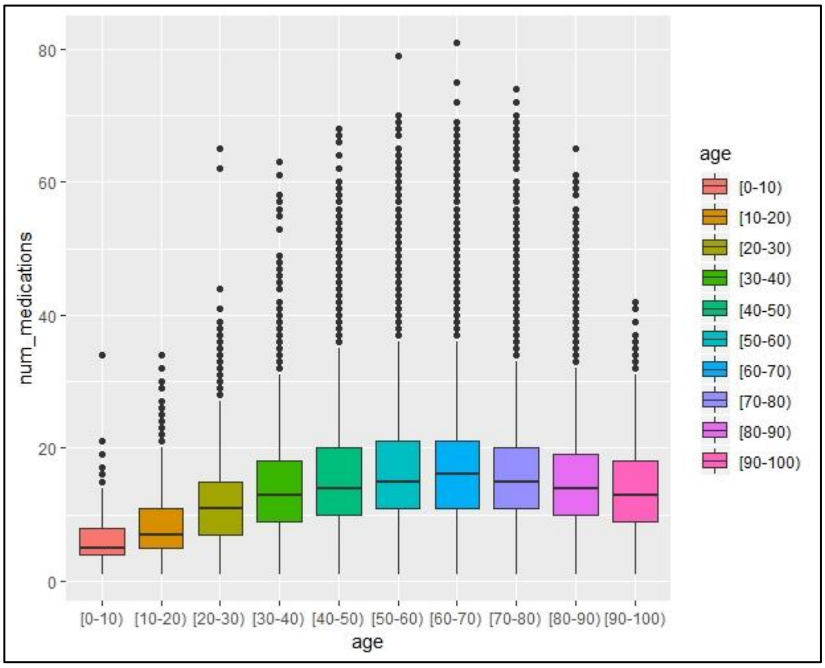


Fig 8: Age vs Num_medication

This graph shows patients in the age group 60-70 has highest number of medications followed by the age group 50-60 yrs patients. Age group 40-50, 70-80 and 80-90 has shown same number of medications taken.

5. Diag2 vs time_in_hospital

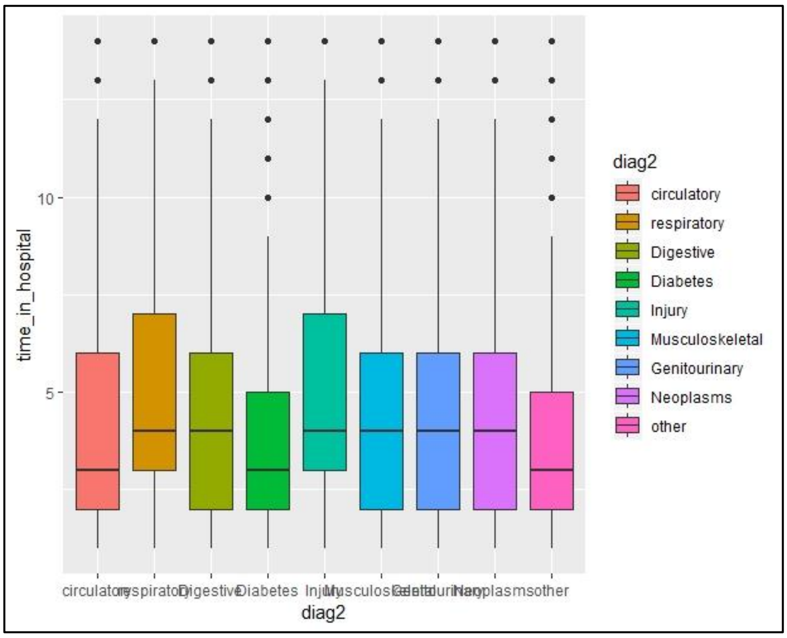


Fig 9: Diag2 vs time_in_hospital

This chart shows that respiratory and injury in diagnosis 2 stayed for the longer time in hospital. Also, the patients with Circulatory, Digestive, Musculoskeletal, Genitourinary and Neoplasms in diagnosis 2 has spend almost the same time in the hospital.

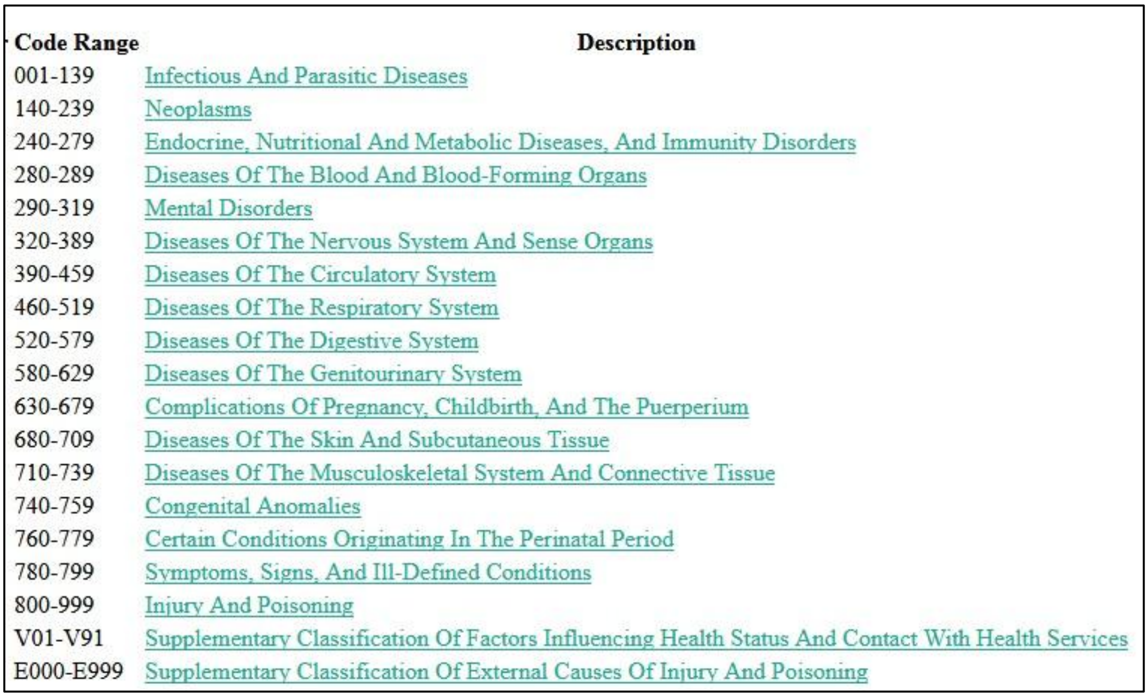
Feature Extraction:

1. Diag_Columns

The dataset contained 3 diagnoses for a given patient (diag_1, diag_2 and diag_3). However, each of these had 700–900 unique ICD codes and it is extremely difficult to include them in the model and interpret meaningfully. Therefore, we collapsed these diagnosis codes into 9 disease categories in an almost similar fashion to that done in the original publication using this dataset. These 9 categories include Circulatory, Respiratory, Digestive, Diabetes, Injury, Musculoskeletal, Genitourinary, Neoplasms, and Others. We referred the ICD-9 codes for the same.

```
data2$diagnosis_group <- factor( rep("other",nrow(data2)),ordered = F,
levels = c("circulatory","respiratory","Digestive","Diabetes","Injury",

"Musculoskeletal","Genitourinary","Neoplasms","other"))
```



Code Range	Description
001-139	Infectious And Parasitic Diseases
140-239	Neoplasms
240-279	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
280-289	Diseases Of The Blood And Blood-Forming Organs
290-319	Mental Disorders
320-389	Diseases Of The Nervous System And Sense Organs
390-459	Diseases Of The Circulatory System
460-519	Diseases Of The Respiratory System
520-579	Diseases Of The Digestive System
580-629	Diseases Of The Genitourinary System
630-679	Complications Of Pregnancy, Childbirth, And The Puerperium
680-709	Diseases Of The Skin And Subcutaneous Tissue
710-739	Diseases Of The Musculoskeletal System And Connective Tissue
740-759	Congenital Anomalies
760-779	Certain Conditions Originating In The Perinatal Period
780-799	Symptoms, Signs, And Ill-Defined Conditions
800-999	Injury And Poisoning
V01-V91	Supplementary Classification Of Factors Influencing Health Status And Contact With Health Services
E000-E999	Supplementary Classification Of External Causes Of Injury And Poisoning

2. admission_type_id

The dataset contains 8 levels of admission_type_id for each patient. So, by putting together the same admission_type_id predicting the variable might be easier. We put “Emergency”, “Urgent” and “Trauma Centre” as one variable because they all defines the same. Also, we have done on other variables to.

3. admission_source_id

The dataset contains 26 levels of admission_source_id for each patient. By mapping some of the same levels together can get the predictions much better. Levels like “Not Available” , “Null”, “Not Mapped” and “Unknow/Invalid” can be get together, which can help us to reduce the levels. This will help to predict much better by decreasing the levels as many as possible.

4. discharge_disposition_id

The dataset contains 29 levels of discharge_disposition_id for each patient. Some levels can also be combined together such as “Discharged/transferred to SNF”, “Discharged/transferred to ICF”, “Discharged/transferred to another type of inpatient care institution”, “Discharged/transferred to home with home health service” and so on can be combined to one level and make the prediction much easier. Also, levels such as “Expired”, “Expired at home. Medicaid only, hospice.” and “Expired in a medical facility. Medicaid only, hospice.” Can be dropped as they have no probability of readmission in the hospital.

Building Training and Testing Model

We randomized (to avoid any selection bias) and divided the clean data obtained into two parts: Training and Test Data, in a 70:30 ratio, which allowed us to train our models on 70% of the data and use the other 30% to assess the performance of our models.

```
sampledata = sample(2, nrow(diabetic),
                    replace = T,
                    prob = c(0.7,0.3))
train = diabetic2[sampledata==1,]
test = diabetic2[sampledata==2,]
```

Feature Selection of Model

While there are many possible combinations of features one could test for. We selected some of the feature which may be relevant for the predictions. By applying different feature combination, we could get different predictions.

```
age+discharged_to+time_in_hospital+num_lab_procedures+num_proc
edures+num_medications+number_outpatient+number_emergency+numb
er_inpatient+number_diagnoses+insulin+change+diabetesMed+diag_
1+diag_2+diag_3+A1Cresult
```

Selecting models

1. Decision Trees: By iteratively and hierarchically observing the level of certainty of predicting whether someone would be readmitted or not, we find the relative importance of different factors using a more human-like decision making strategy in establishing this determination.

2. **Random Forests**: By considering more than one decision tree and then doing a majority voting, random forests helped in being more robust predictive representations than trees as in the previous case. For both Decision Trees and Random Forests, we removed the interaction terms from the feature set since these are already accounted for in tree-based models.

3. Support Vector Machines: Support Vector Machines can help model linearly inseparable data, thus allowing us to explain complex non-linear relationships. However, because of high-dimensional structure and complexity, they are limited by their interpretability to gain insights on how different features are weighted/assigned importance.

4. K-nearest Neighbors: While K-nearest neighbors provide decent predictions, they cannot help in deciding the features that contribute to this decision the most, since features are weighted equally (assuming we normalize them) based on simply their contribution to the proximity/distance function