Prediction of Readmission of Diabetes Patients

GROUP 5

Participants:

Nisarga Kulkarni

Pranavaditya Rekha

Amrithaa.Y

Shreya Uppalapati

Lakshmi Reddy

R.V.Somnath Kumar

Under Supervision of:

Vibha Santhanam

greatlearning

Introduction:

Diabetes mellitus is a serious public health problem that has implications for individuals, communities, and health and human services. Diabetes comprises a complex of metabolic disorders associated with impaired insulin secretion and glucose metabolism . The importance of early detection and management of diabetes to prevent disease progression, poor health outcomes including early onset of complications, and increased use of health services is recognised and supported by policy and practice interventions to improve diabetes care. Yet diabetes remains a significant reason for preventable contact with the health system.

A number of studies have demonstrated that people with diabetes have hospital admission rates between 2 and 6 times higher than people without diabetes . People with diabetes also have excessive lengths of hospital stay compared to people without diabetes. These previous studies used hospital or practice-based populations . Study of hospital-based populations may represent people with severe diabetes including complications of diabetes and its associated morbidity. As a result, the associated risks as well as hospitalisation rates could be overestimated. There is a need to determine the risk of hospitalisation and impact of diabetes among a general community population.

Need of The Study:

It is important to know if a patient will be readmitted in some hospital. The reason is that you can change the treatment, in order to avoid a readmission. As the healthcare system moves toward value- based care, creating interventions to provide additional assistance to patients with increased risk of readmission increases the value of the Hospital.Diabetes is a medical condition that affects ap- proximately 1 in 10 patients in the United States. According to some surveys, patients with diabetes have almost double the chance of being hospitalized than the general population. Therefore, in this project, we will focus on predicting hospital readmission for patients with diabetes.

Proposed Model:

As part of this study the below models are explored,

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- KNN Classifier

- AdaBoost Classifier
- Gradient Boosting Classifier
- XGBoost Classifier

Problem Statement:

With this project, the aim is to find the best model for readmission prediction and the factors which most likely affect the readmission.

Scope:

The stakeholder of this project will be the hospital officials who can use the results to figure out which patients have higher readmission chances. This will help save hospitals millions of dollars and also, improve the healthcare quality.

The Data:

The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated de- livery networks. It includes over 50 features representing patient and hospital outcomes and has 101766 rows . Information was extracted from the database for encounters that satisfied the following criteria.

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
3. The length of stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, A1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency vis- its in the year before the hospitalization, etc.

Feature Description:

Feature Name	Type	Description
Encounter ID	Numeric	Unique identifier of an encounter
Patient number	Numeric	Unique identifier of a patient

Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other
Gender	Nominal	Values: male, female, and unknown/invalid
Age	Nominal	Grouped in 10-year intervals: (0, 10), (10, 20), …,(90, 100)
Weight	Numeric	Weight in pounds.
Admission Type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
Discharge Disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
Time in hospital	Numeric	Integer number of days between admission and discharge
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay
Medical specialty	Nominal	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	Numeric	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	Numeric	Number of distinct generic names administered during the encounter
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter

Number of emergency visits	Numeric	Number of emergency visits of the patient the year preceding the encounter
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter
Diagnosis 1,2,3	Nominal	The diagnosis (coded as first three digits of ICD9);
Number of diagnoses	Numeric	Number of diagnoses entered to the system
Glucose serum test result	Nominal	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	Nominal	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	Nominal	Indicates if there was a change in diabetic medications. Values: "change" and "no change"
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"
23 features for medications	Nominal	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,. 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	Nominal	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.

Data Cleaning:

The following changes have been done for better analysis, visualization and model building. The changes done for the required columns are as below:

1. Missing Values:

We can start analysis by looking at the percentage of missing values in each column.
We found that seven columns have missing values namely
weight,medical_speciailty,payer_code,race,diag_3,diag_2 and diag_1.

	Missing Values	% of Total Values
weight	98569	96.9
medical_specialty	49949	49.1
payer_code	40256	39.6
race	2273	2.2
diag_3	1423	1.4
diag_2	358	0.4
diag_1	21	0.0

We should be careful when dropping columns so as to not discard the available information
Thus if a column has a high percentage of missing values, then it probably will not be of much use.

What columns to retain may be a little arbitrary, but for this project, we will remove any columns with more than 30% missing values. We hence drop weight,medical_speialty and payer_code

We are dropping all rows containing null values as the number of nulls present are less than 5% of the entire dataset.

2.Convert Data to Correct Types

There are a number of columns with numbers that have been recorded as object datatypes. These will have to be converted to numeric datatype before we can do any numerical analysis.

We convert the columns with numbers into numeric data types by replacing the strings which can be interpreted as floats. Then we will convert the columns that contain numeric values into numeric data types.

Values of Diagnosis:
International Classification of Diseases (ICD-9) - The International Classification of Diseases (ICD) is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics.For codes and related diseases, please refer to the following link
https://www2.gov.bc.ca/gov/content/health/practitionerprofessionalresources/msp/physicians/diagno stic-code-descriptions-icd-9.

Diabetes
Circulatory
Respiratory
Digestive
Injury
Musculoskeletal
Genitourinary
Neoplasms
Other

Values of Age:
The values of the age column are given as ranges and is of type string. We will convert them into 3 Categories and create a new feature 'age_cat' .
We will convert age into 3 groups:

0 - 30 as 1
30 - 60 as 2
Greater than 60 as 3

Values of Medications:
For the generic names: **metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, aceto- hexamide, 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**.**

No:0
Down:1
Steady:2
Up:3

Values for Max_glu_serum:

Blood sugar level - The blood sugar level, blood sugar concentration, or blood glucose level is the concentration of glucose present in the blood of humans and other animals. Glucose is a simple sugar and approximately 4 grams of glucose are present in the blood of a 70-kilogram (150 lb) human at all times. The body tightly regulates blood glucose levels as a part of metabolic homeostasis. Glucose is stored in skeletal muscle and liver cells in the form of glycogen; in fasted individuals, blood glucose is maintained at a constant level at the expense of glycogen stores in the liver and skeletal muscle.

For various levels, please refer to the below link -

https://www.sciencedirect.com/topics/immunology-and-microbiology/glucose-level

Values for A1C test result:

If you have diabetes, you should have an A1C test at least twice each year to find out your long-term blood glucose control. The A1C test measures your average blood glucose during the previous 2-3 months, but especially during the previous month.

For various levels, please refer to

thebelow link -

https://clinical.diabetesjournals.org/conten

t/24/1/9

Values for readmitted:

We will proceed based on the theory that either the person was not readmitted or Readmitted .

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.

NO:0
>30:1
<30:1

Exploratory Data Analysis:

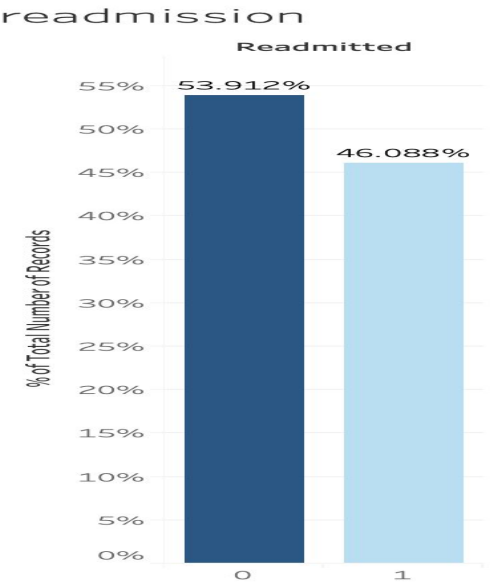
Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

The EDA approach is precisely that--an approach--not a set of techniques, but an attitude/philosophy about how a data analysis should be carried out.

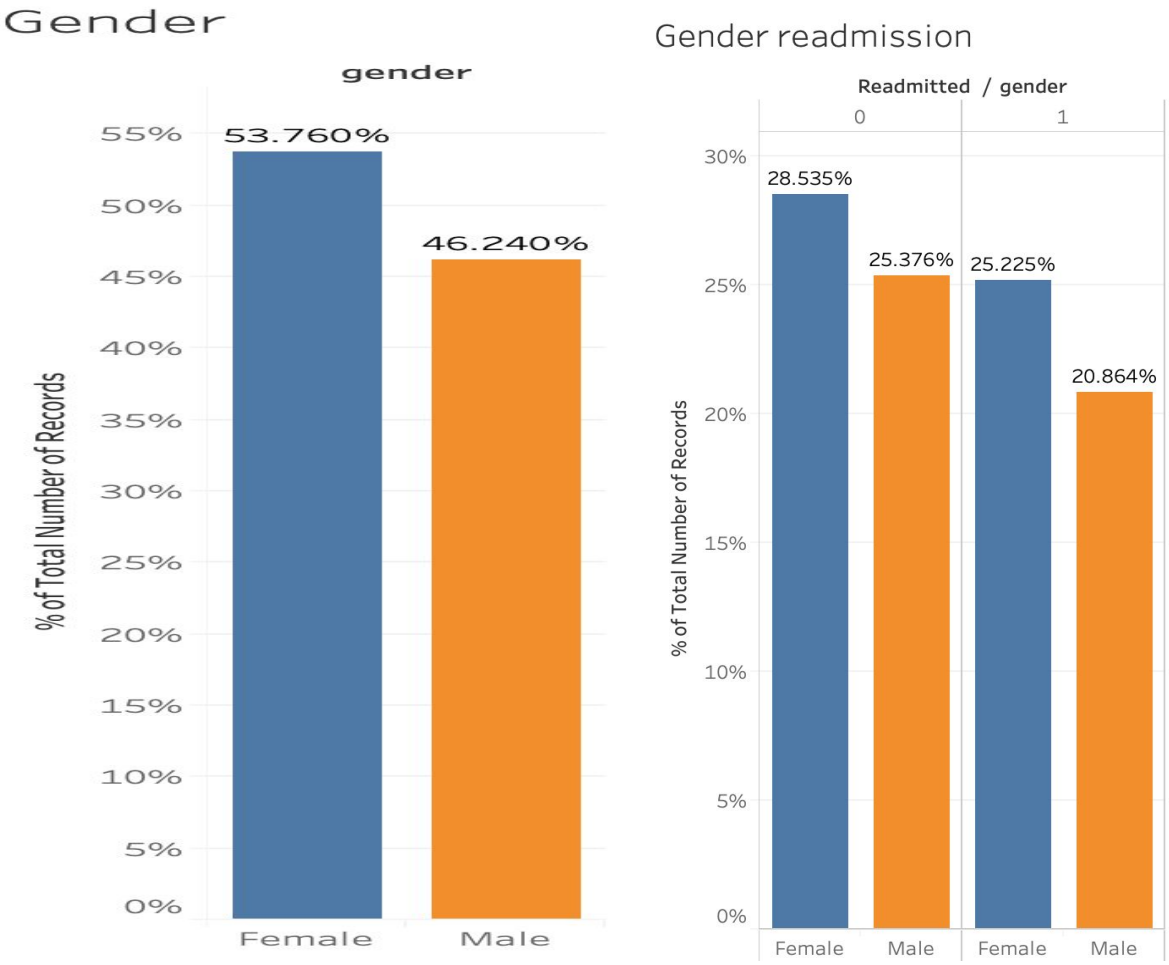
1.READMISSION:

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. we consider (**No as 0** and **<30 and >30 as 1.)**



The number of patients appearing for readmission in the hospitals is slightly less than the ones not appearing for the same.The target variable is fairly balanced.

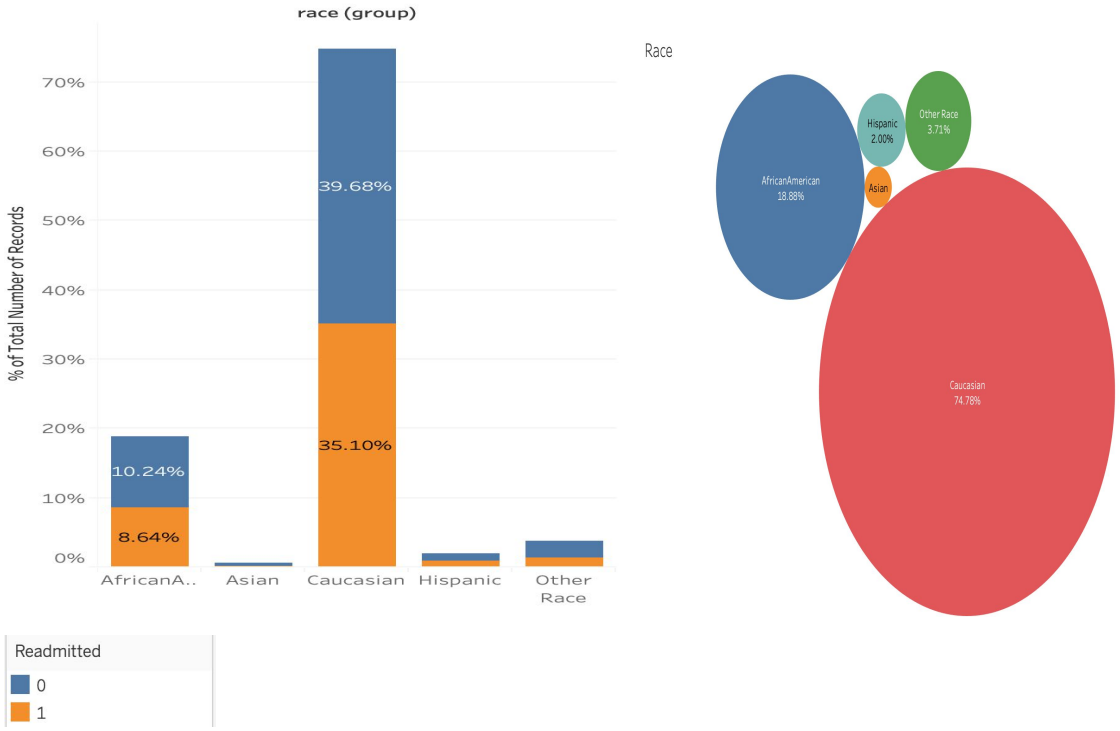
2.GENDER: Values: Male and Female



The number of females is more than the number of males.
Percentage of males getting readmitted is approximately 45.67% of total males.
Percentage of females getting readmitted is approximately 47.44 of total females

3.RACE:

Values: Caucasian, Asian, African American, Hispanic, and other.

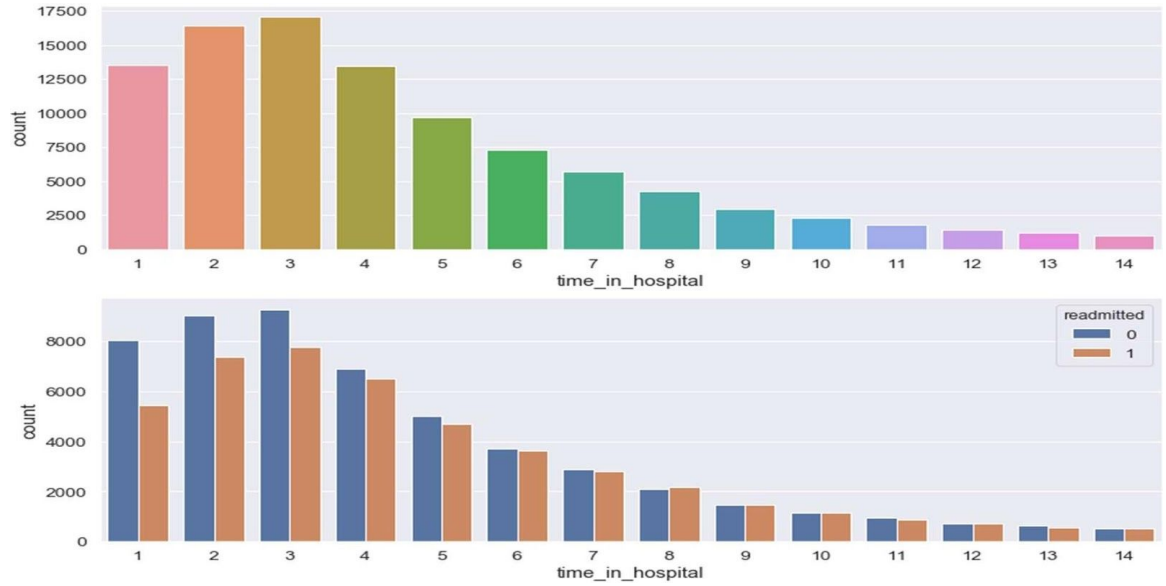


The relative percentage for Asian, Hispanic, and other are very less, so together we consider all the three races as Other.

The total number of Caucasians are significantly more than any other race.The race of Asians are least in number.47.13% of the Caucasians get readmitted (highest) whereas 35.84% of Asians get readmitted (lowest).

4.TIME_IN_HOSPITAL:

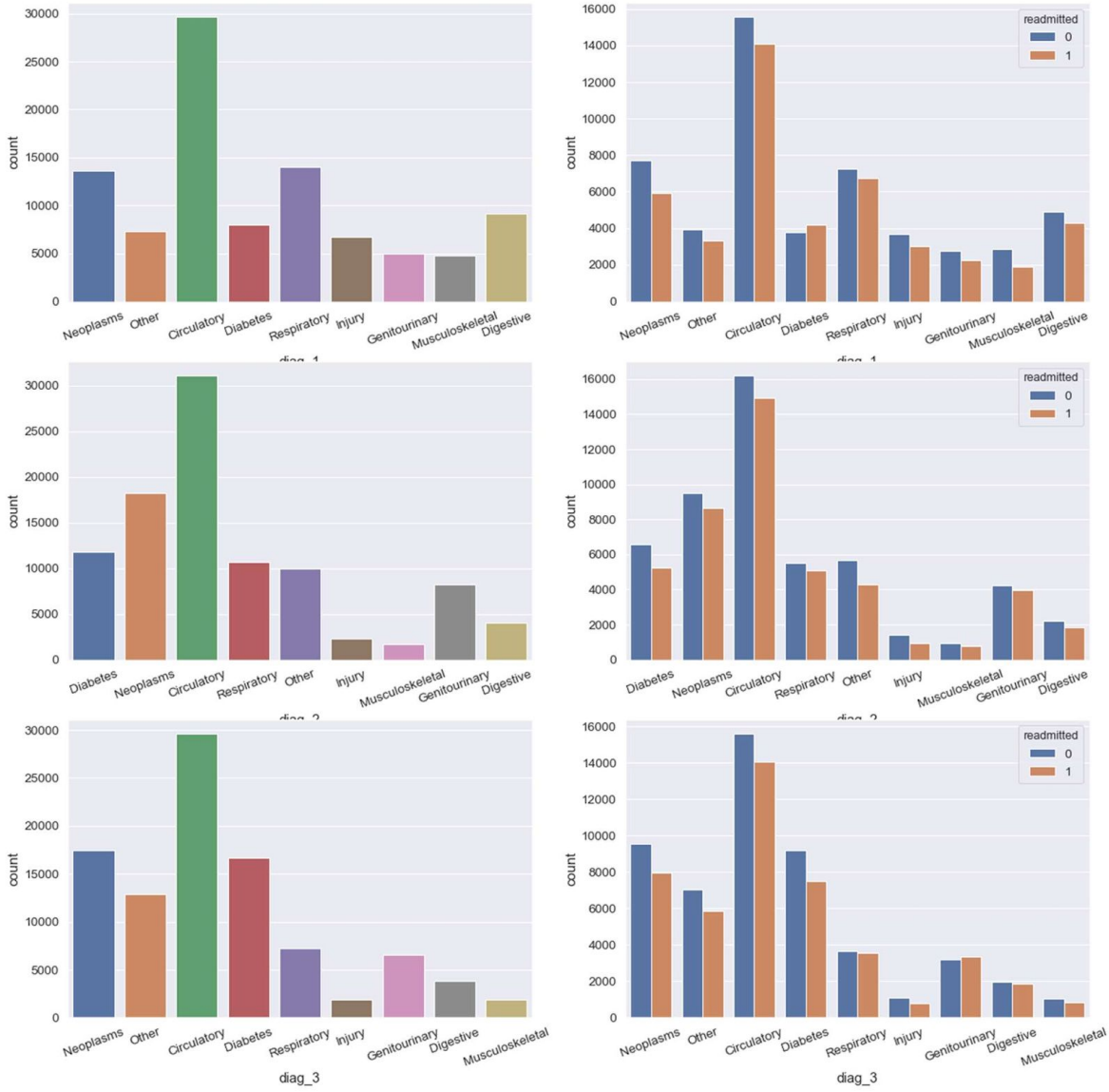
Integer number of days between admission and discharge



The average number of days of admission is around 4 days.
Most patients spend time ranging from 1 to 4 days in the hospital.
Approximately 50-51% of the patients get readmitted under the time of 8 to 10 days (highest) whereas 40.31% in case of patients coming for a single day (lowest).

5.DIAGNOSIS:

The diagnosis (coded as first three digits of ICD9);
International Classification of Diseases (ICD-9) - The International Classification of Diseases (ICD) is designed to promote international comparability in the collection, processing, classification, and presentation of mortality statistics.

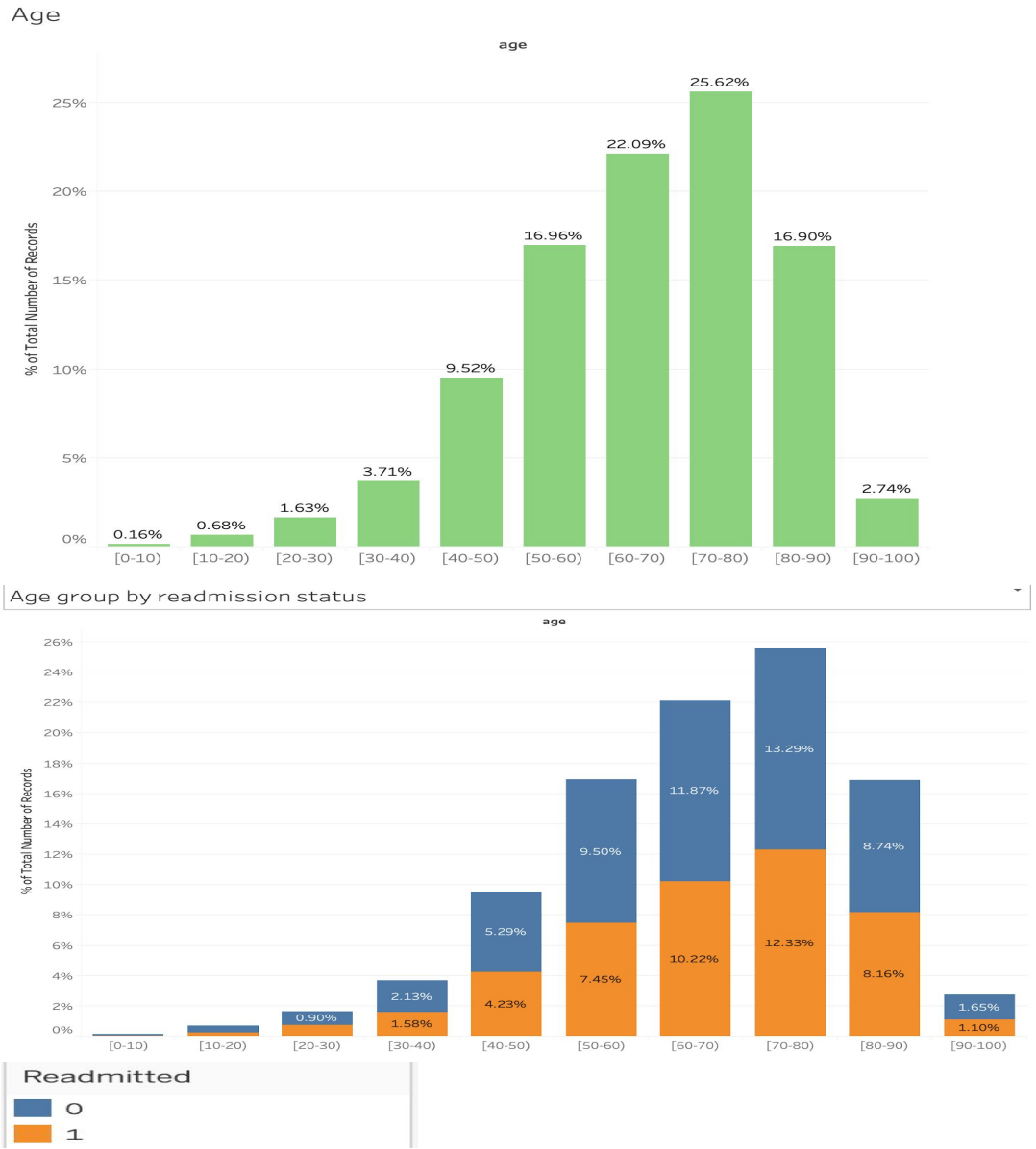


The total number of patients under the primary diagnosis category of 'Circulatory' is maximum.

Diabetes has the maximum number of readmissions under primary diagnosis (52.74%) whereas Musculoskeletal has the lowest with 39.82%.
Genitourinary and Respiratory has the maximum number of readmissions under secondary diagnosis (approx. 48%) whereas Injury has the lowest with 39.99%.
Genitourinary has the maximum number of readmissions under tertiary diagnosis (51.16%) whereas Injury has the lowest with 41.59%.

6.AGE_DISTRIBUTION:

Grouped in 10-year intervals: (0, 10), (10, 20), …,(90, 100) We have nine different categories.

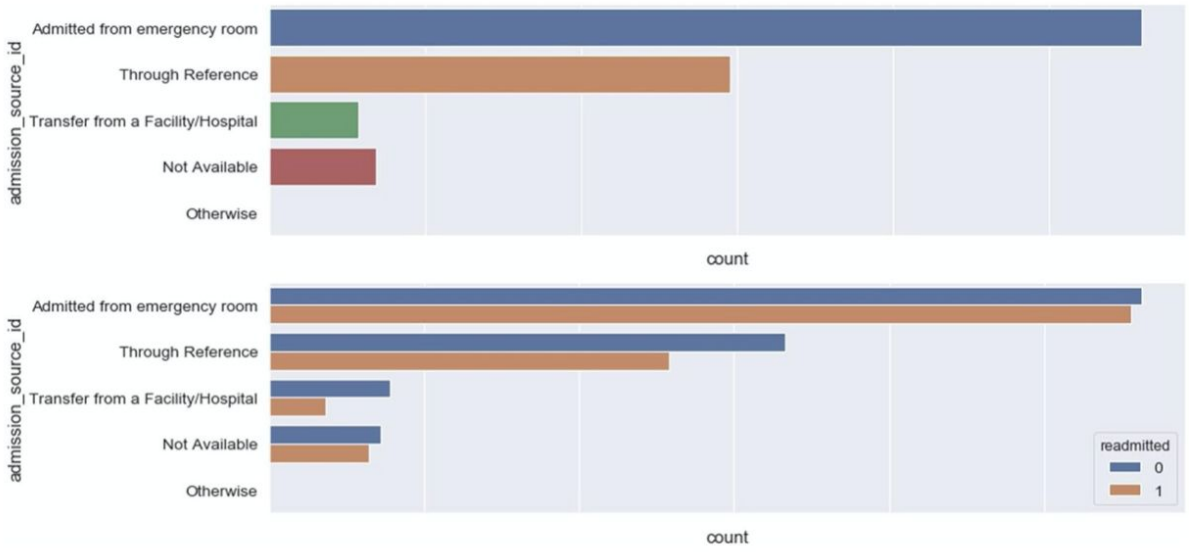


Patients with age ranging from 70-80 years are most in number followed by 60-70. Around 48.5% of patients under the age groups of 70-80 and 80-90 gets readmitted which is the most in number while the age group of 0-10 has the least with 21.54%.

7.ADMISSION SOURCE_ID:

Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital. considering all the different descriptions in admission_source we have categorized them into five different categories. They are:

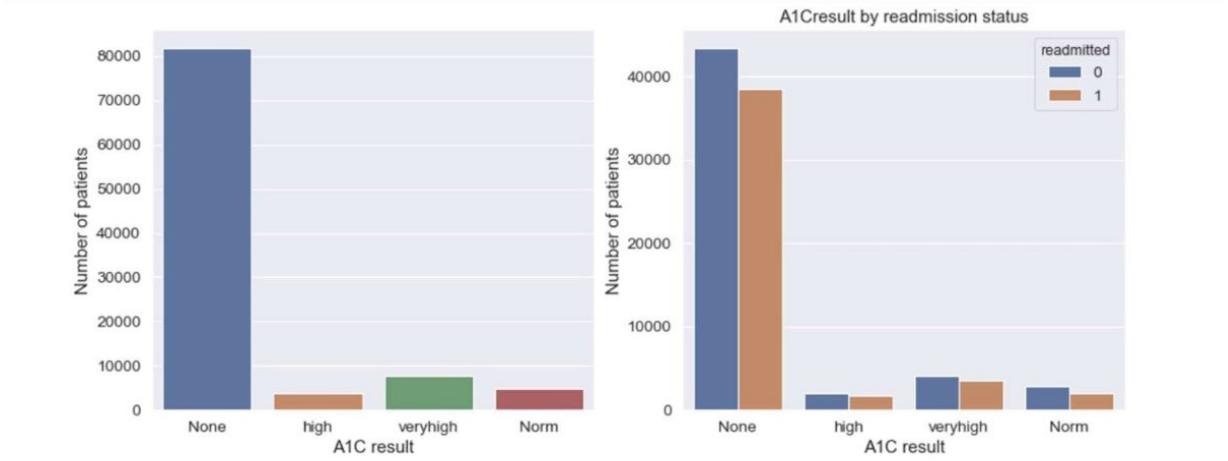
admitted from the emergency from,through reference,transfer from a facility/hospital,not available,otherwise.



Patients having an admission source of emergency rooms are most in number followed by references.Patients admitted from emergency rooms are having the highest rate of readmission (49.69%).Patients transferred from other undefined sources are having the lowest rate of readmission (30%).

8.A1CRESULT:

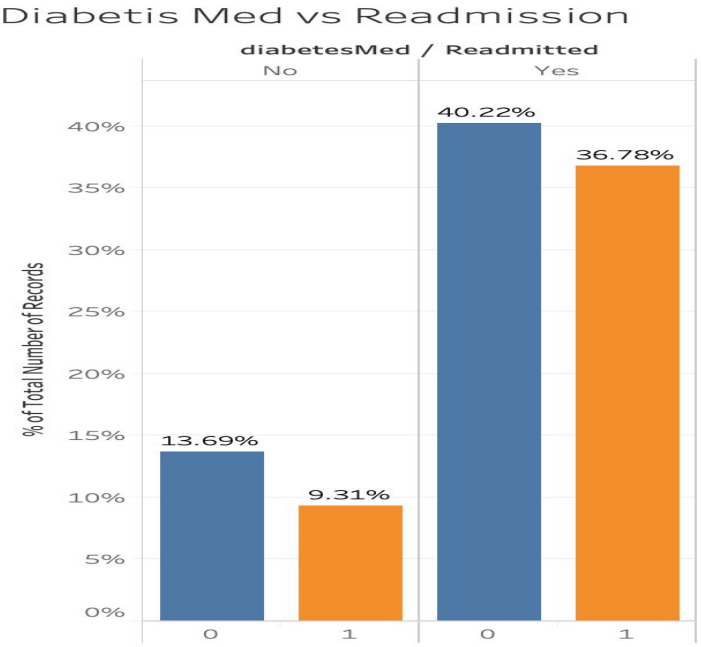
Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.



Number of patients who did not take the A1C test are higher.Approximately 47% of patients are getting readmitted who either did not take the test or had a very high test result (highest) while people whose results were normal are lowest in number (42.03%)

9.DiabetesMed:

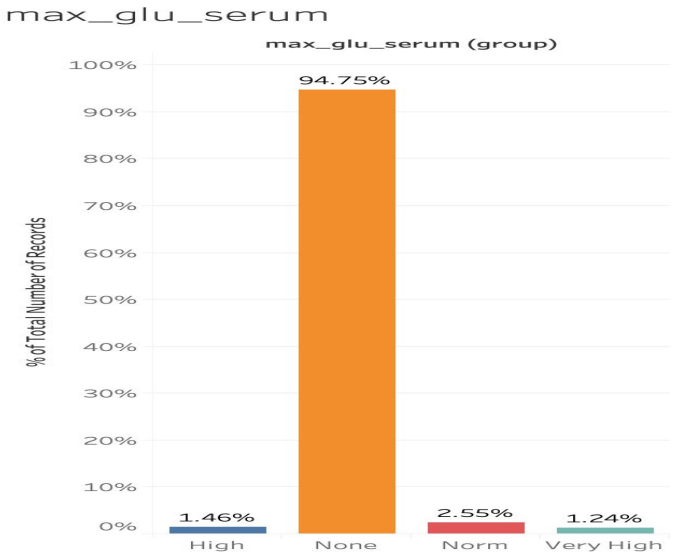
Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”

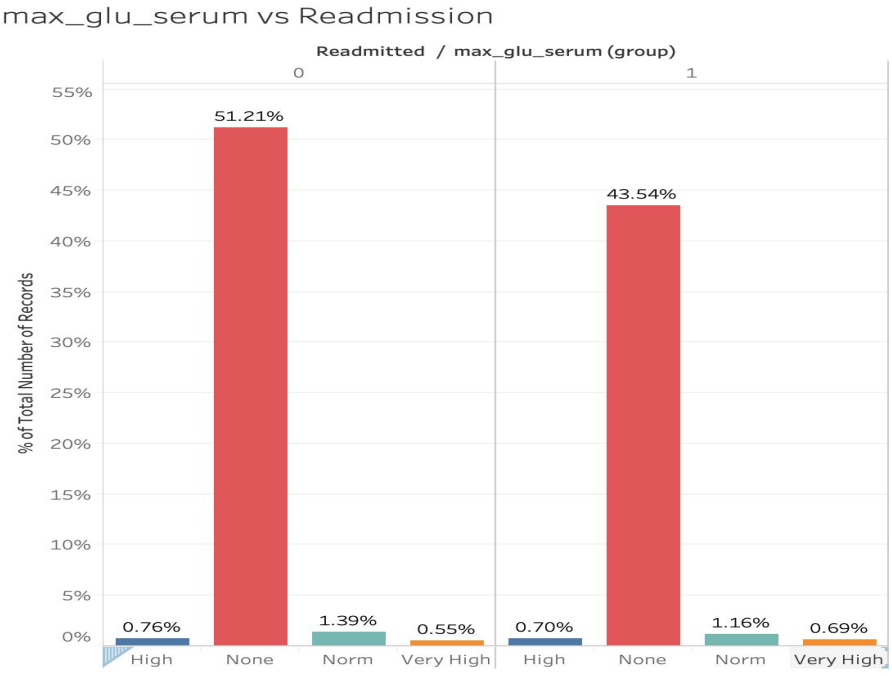


Patients who are prescribed with diabetes related medication are having a higher rate of re- admission (48.33%).

10.Max_glu_serum:

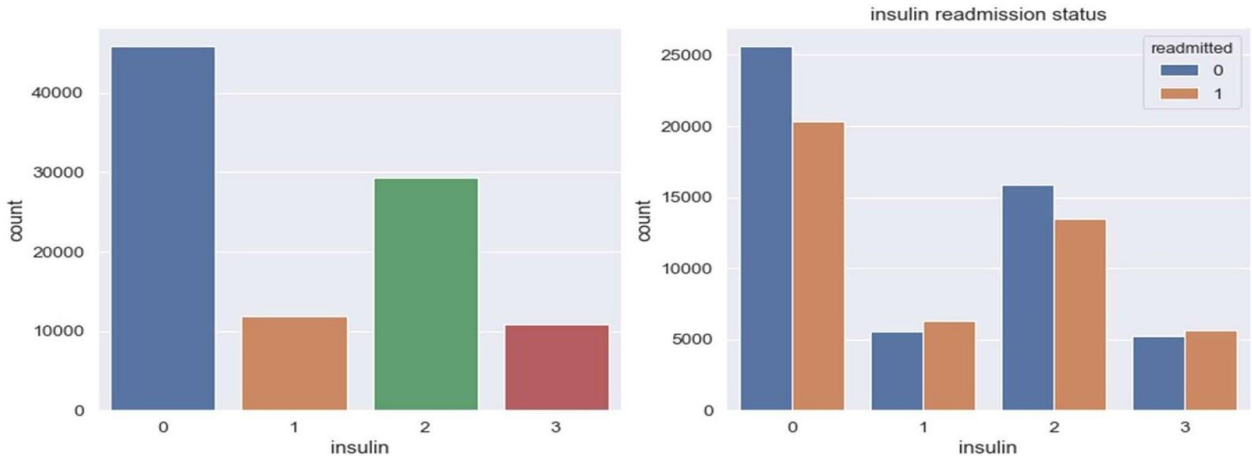
Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured.





- Approximately 95% of patients are not taking max_glu_serum test
- Approximately 56.07% of patients whose results are very high were readmitted (highest) whereas 45.66% of patients under the category of people whose results are normal were re- admitted (lowest).

11.INSULIN



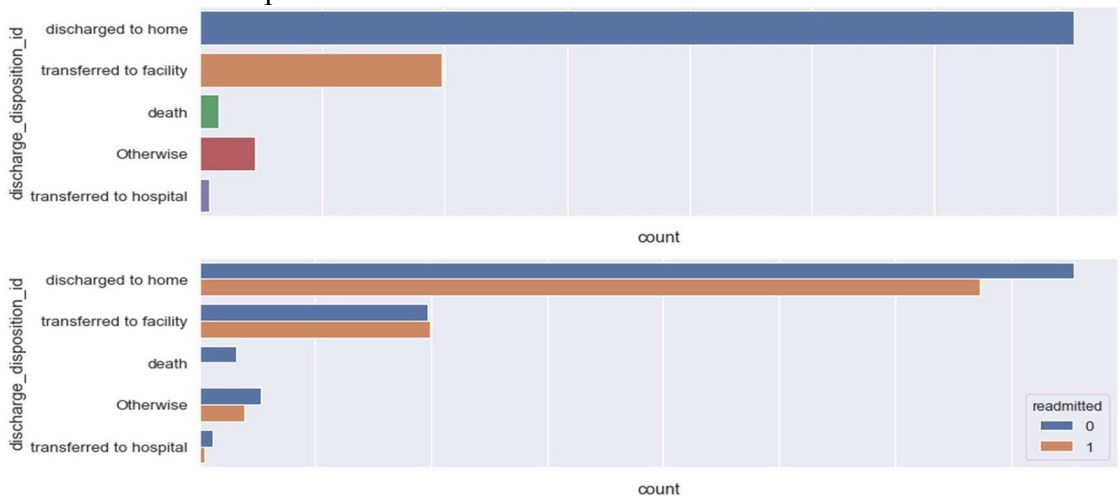
- Insulin was not prescribed for 46% of the patients.
- There is no change in insulin dosage for approximately 30%.
- Readmission rate is more for the patients whose insulin dosage is either increase or decrease.

Note : except insulin all other medications dosage was steady.

12.DISCHARGE DISPOSITION:

Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available. considering all the different descriptions we have categorized them into five different categories. They are:

- discharged to home,
- transferred to facility,
- death,
- otherwise,
- transferred to hospital.

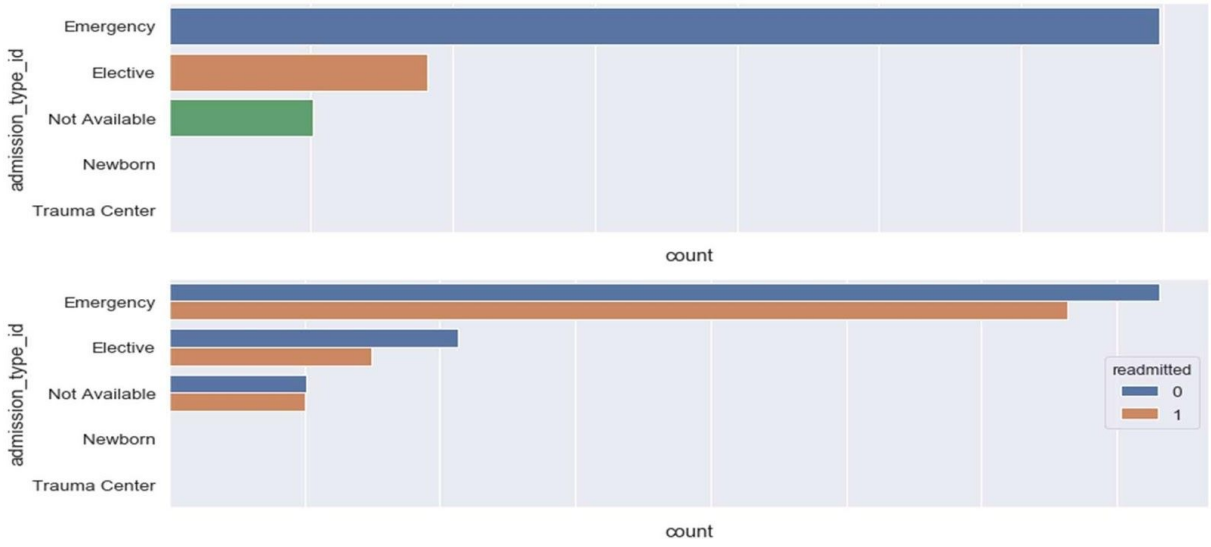


Most of the patients are discharged to their homes.50.29% of patients who were transferred to another facility got readmitted, having the highest rate.Patients who were transferred to other hospitals had the least rate of readmission (27.01%).

13.ADMISSION TYPE:

considering all the different descriptions in admission_type we have categorized them into five different categories. they are:

- emergency,
- elective ,
- not available ,
- newborn ,
- trauma centre.

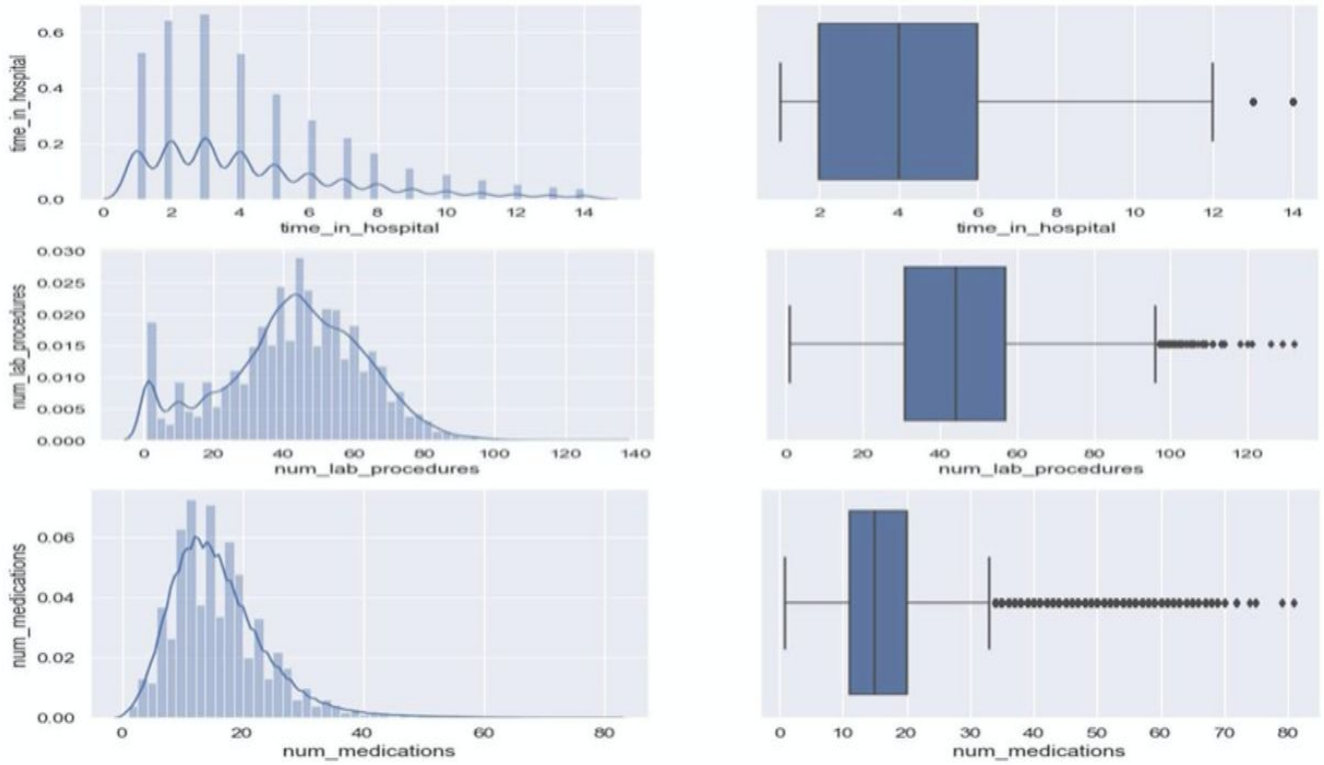


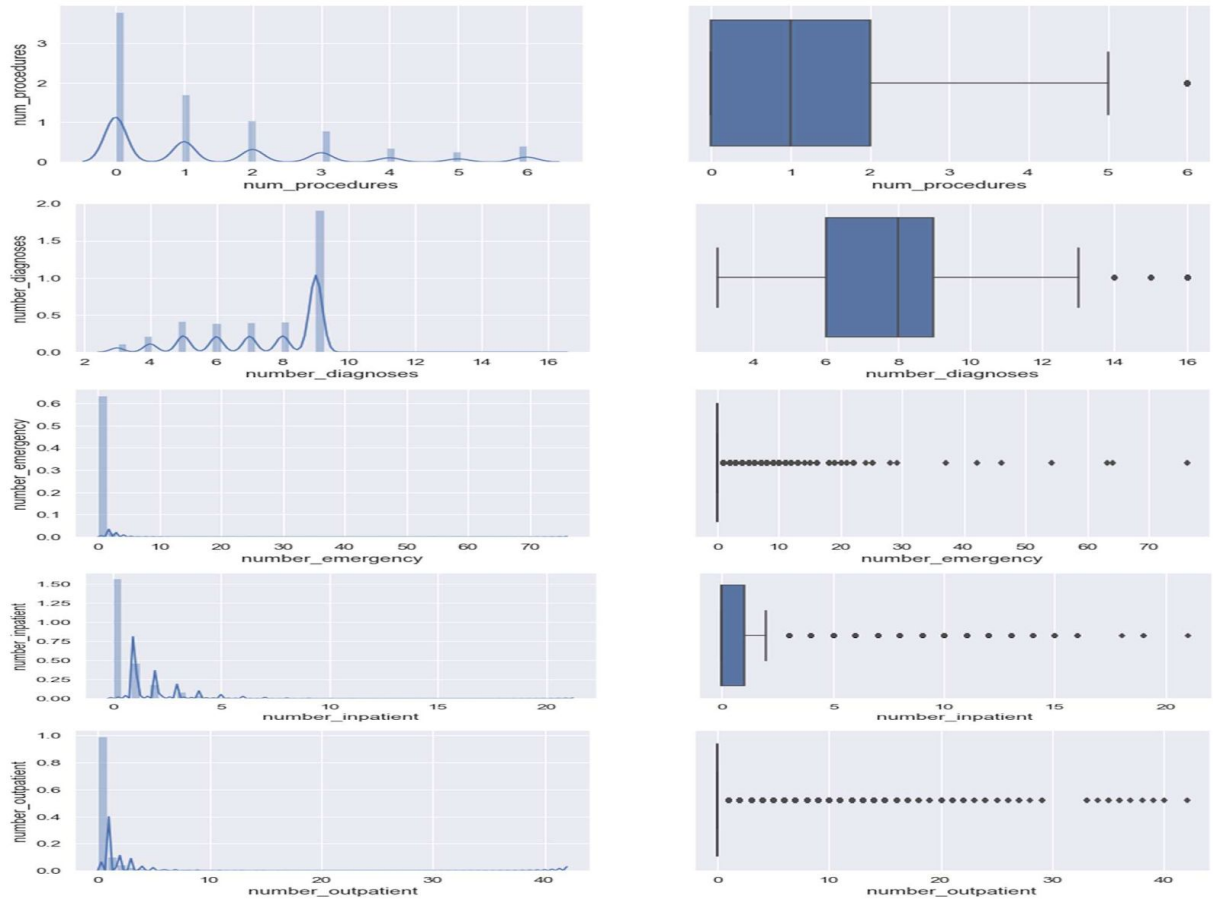
- Admissions of emergency are most in number followed by elective types.
- Admissions of undefined types are having the highest rate of readmissions (49.79%).
- Trauma Centre type admissions are having the lowest rate (0%).

14.CHECKING OUTLIERS FOR NUMERICAL VARIABLES:

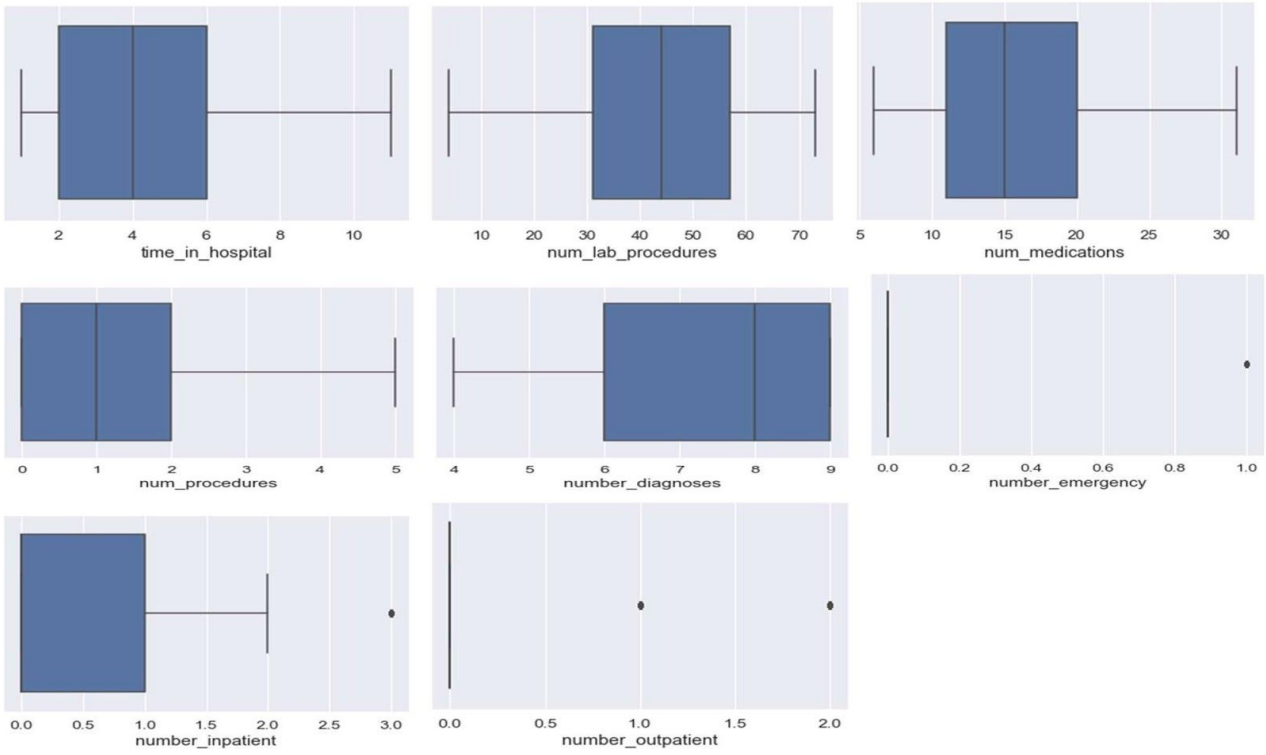
Outliers can occur for many reasons: typos, malfunctions in measuring devices, incorrect units, or they can be legitimate but extreme values. Outliers can throw off a model because they are not indicative of the actual distribution of data.

When we remove or cap outliers, we want to be careful that we are not throwing away measurements just because they look strange. They may be the result of actual phenomena that we should further investigate.
When capping outliers, we tried to be as conservative as possible, using the definition of an outlier based on winsorization :
https://en.wikipedia.org/wiki/Winsorizing





As we have outliers in many of the attributes we treat them using transformation.



Modelling:

1.Feature Selection:

Feature Selection: The process of choosing the most relevant features in your data. "Most relevant" can depend on many factors, but it might be something as simple as the highest correlation with the target, or the features with the most variance. In feature selection, we remove features that do not help our model learn the relationship between features and the target. This can help the model gener- alize better to new data and results in a more interpretable model. Generally, I think of feature selec- tion as subtracting features so we are left with only those that are most important. Feature selection is an iterative process that will usually require several attempts to get right. Often we will use the re- sults of modeling, such as the feature importance from a random forest, to go back and redo feature selection, or we might later discover relationships that necessitate creating new variables. Moreover, these processes usually incorporate a mixture of domain knowledge and statistical qualities of the data.

Feature selection often has the highest returns on time invested in a machine learning problem. It can take quite a while to get right, but is often more important than the exact algorithm and hyper param- eters used for the model. If we don't feed the model the correct data, then we are setting it up to fail and we should not expect it to learn.
For feature selection, we will do the following:

1. Remove columns related to ids.
2. Perform a Chi-square test of independence to select only significant features based on their p- values.
3. For the Numerical variables we performed ANOVA test.

After performing the statistical test (chi-square and ANOVA) we considered features with p-value (less than 0.05) as significant features.

Those Features are:

- **time_in_hospital**
- **num_lab_procedure**
- **num_procedures**
- **num_medications**
- **number outpatient**

- **number emergency**
- **number inpatient**
- **number diagnoses**
- **race**
- **gender**
- **diag_1**
- **diag_2**
- **diag_3**
- **max_glu_serum**
- **metformin**
- **repaglinide**
- **glipizide**
- **pioglitazone**
- **rosiglitazone**
- **acarbose**
- **insulin**
- **diabetesMed**
- **age_cat**
- **discharge_disposition**
- **admission_source**
- **admission_type**
- **change**
- **A1c Result**
- **Readmitted**

2. Scaling Features

The final step to take before we can build our models is to scale the features. This is necessary because features are in different units, and we want to normalize the features so the units do not affect the algorithm. Linear Regression and Random Forest do not require feature scaling, but other methods, such as support vector machines and k nearest neighbors, do require it because they take into account the Euclidean distance between observations. For this reason, it is a best practice to scale features when we are comparing multiple algorithms.

There are two ways to scale features:

- For each value, subtract the mean of the feature and divide by the standard deviation of the feature. This is known as standardization and results in each feature having a mean of 0 and a standard deviation of 1.

- For each value, subtract the minimum value of the feature and divide by the maximum minus the minimum for the feature (the range). This assures that all the values for a feature are be- tween 0 and 1 and is called scaling to a range or normalization.

When we train the scaling object, we want to use only the training set. When we transform features, we will transform both the training set and the testing set.We have used min_max scaler for scaling our features in this project.

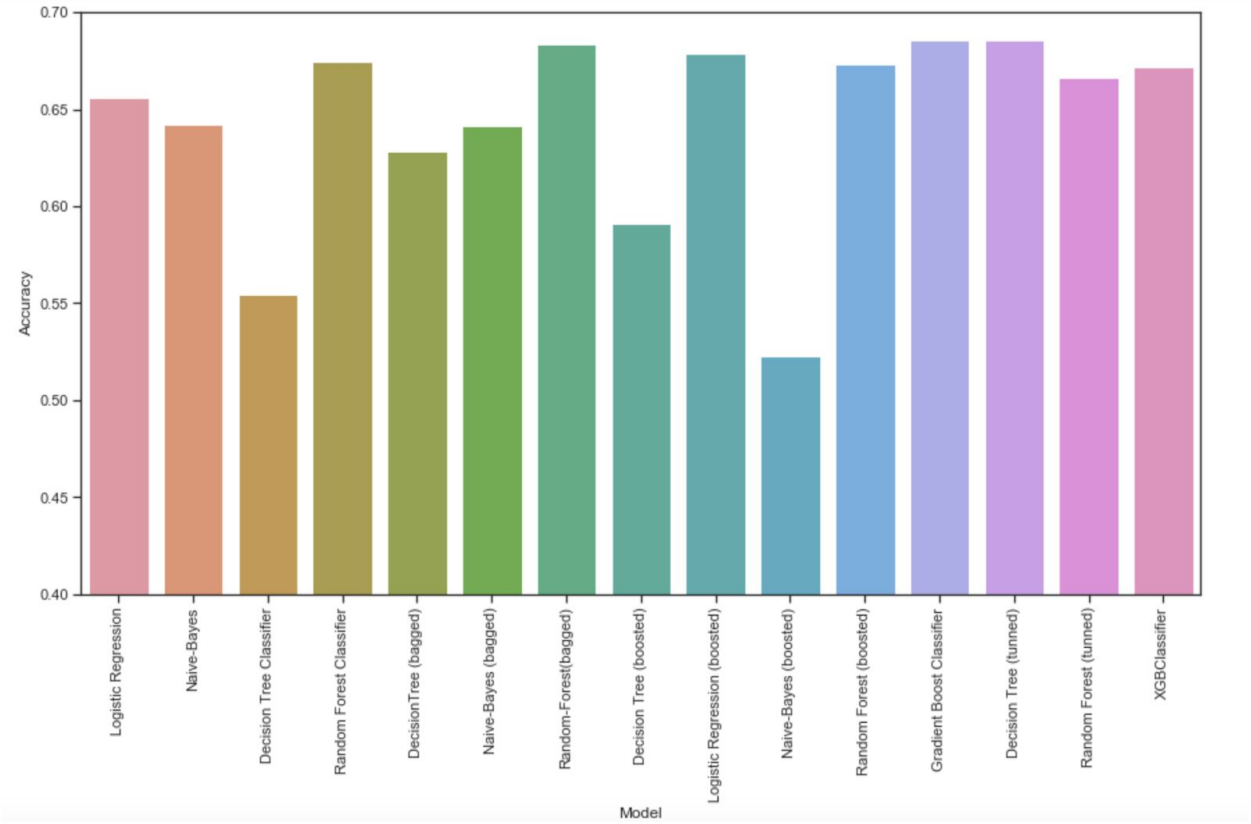
3.MODELS TO EVALUATE

We will compare different machine learning models using the great Scikit-Learn library:

1. Logistic Regression
2. Naive-Bayes
3. Decision Tree Classifier
4. Random Forest Classifier
5. Ensemble Techniques (Adaboost, Gradient Boosting, and Bagging)

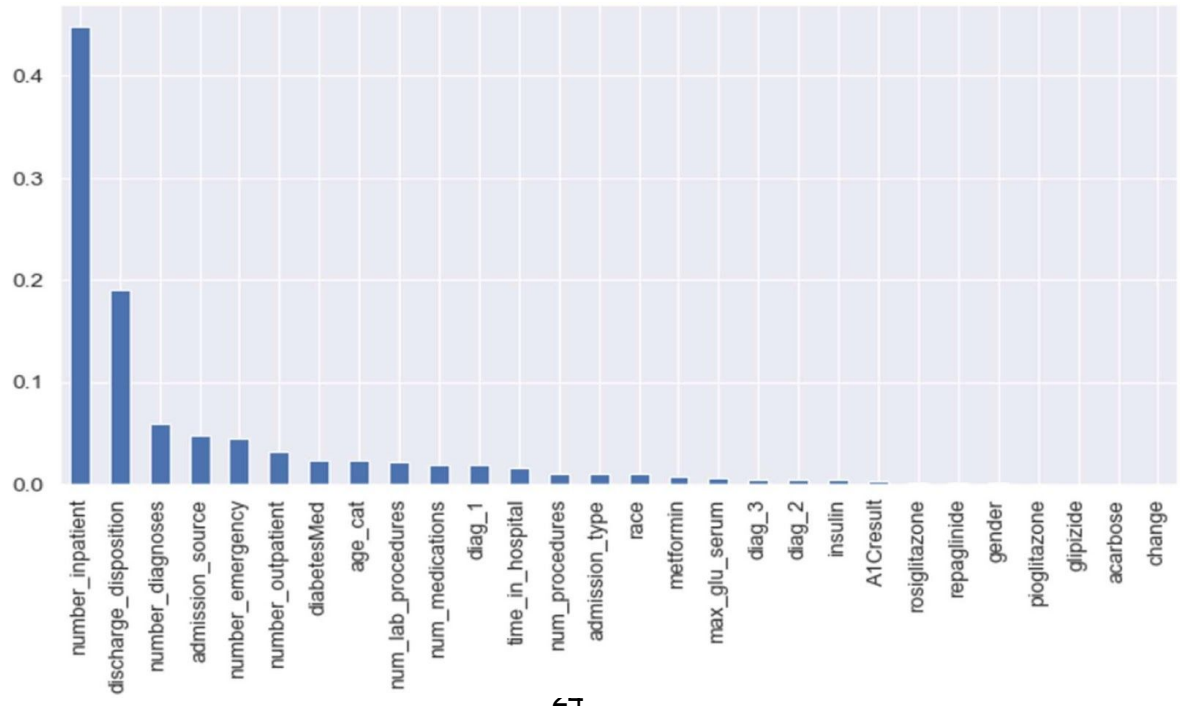
To compare the models, we are going to be mostly using the Scikit-Learn defaults for the model hyper parameters. Generally these will perform decently, but should be optimized before actually using a model. At first, we just want to determine the baseline performance of each model, and then we can select the best performing model for further optimization using hyper parameter tuning.

	model	train_accuracy	test_accuracy
0	Logistic Regression	0.655984	0.655708
1	Naive-Bayes	0.641281	0.641793
2	Decision Tree Classifier	1.000000	0.554627
3	Random Forest Classifier	1.000000	0.674653
4	DecisionTree (bagged)	0.998816	0.627994
5	Naive-Bayes (bagged)	0.641135	0.641673
6	Random-Forest(bagged)	0.998829	0.683691
7	Decision Tree (boosted)	0.594463	0.590826
8	Logistic Regression (boosted)	0.678737	0.678558
9	Naive-Bayes (boosted)	0.522961	0.522784
10	Random Forest (boosted)	1.000000	0.673077
11	Gradient Boost Classifier	0.691868	0.685490
12	Decision Tree (tunned)	0.691408	0.685461
13	Random Forest (tunned)	0.676380	0.666343
14	XGBClassifier	0.683134	0.671926



4.FEATURE IMPORTANCE:

Gradient Boost Classifier



5.CHALLENGES:

One of the major challenges was converting data to correct types and feature engineering as there are a large number of categorical features.

There was a lot missing data in some columns i.e., weight with 96.9%, medical_speciality with 49.1% and payer code with 39.6% of missing data. Although we were careful to not dis- card information and when dropping columns, as a column has a high percentage of missing values, it probably will not be of much use. What columns to retain may be a little arbitrary, but for this project, we removed columns with more than 30% missing values.

As we have 50 features in the data, feature selection by statistical tests played a major role in obtaining the optimal features as we were unable to get the correlation between the features due to a large number of categorical features.

6.RECOMMENDATIONS:

- Develop risk stratification and predictive analytics capabilities
- Leverage patient engagement technology
- Utilize clinical decision support tools
- Enhance care coordination, communication

7.CONCLUSION:

Readmission rate is a quality evaluation metric customarily used to extrapolate the quality of life in- dex of the patient population and the quality of healthcare delivery. Irrespective of the developments in biomedical and healthcare research practices, hospital quality control offices still use traditional pre- defined sets of variables to infer the probability of patient readmission. However, predictive analytics could provide evidence to improve the quality of healthcare delivery.

In this Project, we implemented a predictive approach to identify patients prone to readmission and thus, systematically reduce the number of avoidable readmissions mainly caused by patient non- compliances to medication instruction or early discharge from hospital. Our project recommended a gradient boosting method for identifying patients most likely to get readmitted. The model was able to catch 68.62% of the readmissions and is about 1.5 times better than just randomly picking patients.

REFERENCES:

1. https://www2.gov.bc.ca/gov/content/health/practitioner-professional-resources/msp/physicians/diagnostic-code-descriptions-icd-9
2. https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/diagnosis-treatment/drc-20351199
3. https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html
4. https://new.hindawi.com/journals/bmri/2014/781670/
5. https://www.datacareer.de/blog/parameter-tuning-in-gradient-boosting-gbm/