**project report on diabetes readmission project**

**Abstract**

As diabetes patient readmission rate is becoming one of the major concerns for many national hospitals in the U.S. It is of great significance to study the possible causes and risks of diabetes patients’ readmission. In this project, the relationship between diabetes and the various patient attributes are examined from the Health Facts data which is an extract representing 10 years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks throughout the United States. It includes over 50 features representing patient and hospital outcomes using the features available. Also drawing insights from the analysis to help the Hospitals understand the most important features that help the Hospitals understand the importance of each feature compared to others that will lead the readmission of diabetes patients.

The data-set contains 101718 rows and 50 columns to explore. Exploratory analysis is done after the data cleansing by changing the variables to correct format. Univariate analysis for every feature followed by bi-variate analysis to see how each feature is related, affecting the readmission status to find insights and understand the dependency.The insights that are expected and found are also discussed. We found outliers in the continuous features whose effect was reduced by using transformation. Categorical variables are encoded using label encoding.

Even though we use classification techniques to predict the readmission of patients, inference on how the readmission status varies with the features numerically has a higher importance than very high accuracy itself. Logistic regression being a simpler model, gives better inference, so a Logistic regression is fit to see the values of Accuracy score. This model yielded an Accuracy score of 0.6203. Further variables looking VIF values all are less than 5. Further drop in the features resulted in p-values in chi-square test for categorical and ANOVA for numerical variables less than 0.05. Other classification techniques like Decision Trees and ensemble techniques like boosting and bagging are also used to find them achieving slightly better accuracy score values.

• Techniques used: Exploratory analysis

• Tools: Python

• Domain: Medical

Table of Contents

**I.INTRODUCTION.................................................................................................................................. 1**

1. Problem statement:................................................................................................................................. 1

2.Dataset:.....................................................................................................................................................1

3.Shape:..................................................................................................................................................…1

**II.LITERATURE....................................................................................................................................... 1**

Columns:..................................................................................................................................................1

**III.DATACLEANSING..............................................................................................................................3**

1.Dataformats:..........................................................................................................................................3

2. Dealing with Categorical Data:............................................................................................................. 3

**IV. EXPLORATORY DATA ANALYSIS...................................................................................................4**

**VI. MODELLING:.....................................................................................................................................30**

1. Base Model.....:........................................................................................................................... 30

2. Feature Engineering..............................................................................................................................31

3. Other models:...................................................................................................................................... 31

**VII. RECOMMENDATIONS......................................................................................................................32**

**I. INTRODUCTION**

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,Readmissions are especially serious - they represent a failure of the health system to provide adequate support to the patient and are extremely costly to the system. one solution is to create interventions to provide additional assistance to patients with increased risk of readmission. We can identify these patients using predictive modeling from data science to help prioritize patients.

One patient population that is at increased risk of hospitalization and readmission is that of diabetes. Diabetes is a medical condition that affects approximately 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.

Databases of clinical data contain valuable but heterogeneous and difficult data in terms of missing values, incomplete or inconsistent records, and high dimensionality understood not only by number of features but also their complexity. Additionally, analyzing external data is more challenging than analysis of results of a carefully designed experiment or trial, because one has no impact on how and what type of information was collected. Nonetheless, it is important to utilize these huge amounts of data to find new information or knowledge that is possibly not available anywhere.

**1. Problem statement:**

The aim here is to Predict if a patient with diabetes will be readmitted to the hospital.

**2. Data-set:**

The data set represents 10 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.

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 visits in the year before the hospitalization, etc.

**3. Shape:**

101766 rows and 50 columns.

**II.LITERATURE:**

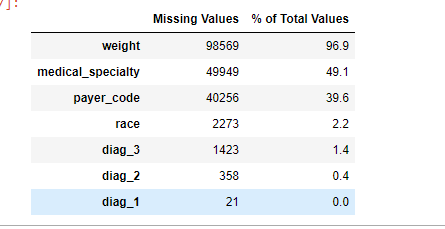
|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Type** | **Description** |
| Encounter ID | Numeric | Unique identifier of an encounter |
| Patient number | Numeric | Unique identifier of a patient |
| Race | Nominal | Values:Caucasian,Asian,African American, Hispanic, and other |
| Gender | Nominal | 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 in 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 (either dosage or generic name). Values: “change” and “no change” |
| Diabetes medications | Nominal | Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” |
| 24 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, 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 | 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. |

**III. DATA CLEANSING**

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. Missing values are fine when we do Exploratory Data Analysis, but they will have to be filled in for machine learning methods.



Although we want to be careful to not discard information and should be careful when dropping columns, 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 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**

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/practitioner-professional resources/msp/physicians/diagnostic-code-descriptions-icd-9](https://www2.gov.bc.ca/gov/content/health/practitioner-professional%20resources/msp/physicians/diagnostic-code-descriptions-icd-9).

**Values of Age:**

The values of the age column are given as ranges and is of type string. We converted them into 3 Categories and created a new feature 'age\_cat'. Considering

* Age<=40: 0
* Age>41 and age<=60: 1
* Age>60: 2

**Values of 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**.**

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

**Values for A1C test result and change (HbA1c)**:

The A1C test is sometimes called the haemoglobin A1C, HbA1c, glycated haemoglobin, or glycohemoglobin test. Haemoglobin is the part of a red blood cell that carries oxygen to the cells. Glucose attaches to or binds with haemoglobin in your blood cells, and the A1C test is based on this attachment of glucose to haemoglobin.The higher the glucose level in your bloodstream, the more glucose will attach to the haemoglobin. The A1C test measures the amount of haemoglobin with attached glucose and reflects your average blood glucose levels over the past 3 months.

The HbA1C test result is reported as a percentage. The higher the percentage, the higher your blood glucose levels have been.

Haemoglobin A1c is an important measure of glucose control, which is widely applied to measure performance of diabetes care. The measurement of HbA1c at the time of hospital admission offers a unique opportunity to assess the efficacy of current therapy and to make changes in that therapy if indicated. By a “change of medication” we understand any dosage change (increase or reduction) as well as change to a drug with a different generic name, for example, a change of the type of insulin or an introduction of a new drug.

We considered four groups of encounters:

1.no HbA1c test performed

2. HbA1c performed and in normal range

3. HbA1c performed and the result is greater than 8% with no change in diabetic medications

4. HbA1c performed, result is greater than 8%, and diabetic medication was changed.

**LABELENCODING:**

In the dataset there are multiple categorical variables. We have Performed Label EncodingTechnique to deal that variables.(**race,diag\_1,diag\_2, diag\_3,HbA1c,gender,max\_glu\_serum,discharge\_disposition,admission\_source,admission\_type,readmitted,diabetesMed**).

**Race:**

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

Label encoded values are:

* Caucasian:
* AfricanAmerican:
* Other:

**diag\_1, diag\_2, diag\_3:**

Label encoded values are:

* Circulatory:0
* Diabetes:1
* Digestive:2
* Genitourinary:3
* Injury:4
* Musculoskeletal: 5
* Neoplasms:6
* Other:7
* Respiratory: 8

**HbA1c:**

Label encoded values are:

* No test was performed: 0
* Result was high and the diabetic medication was changed: 1
* Normal result of the test: 2
* Result was high but the diabetic medication was not changed: 3

**gender:**

Label encoded values are:

* Female:0
* Male:1

**max\_glu\_serum:**

Label encoded values are:

* None:0
* Norm:1
* High:2
* Veryhigh:3

**discharge\_disposition:**

Label encoded values are:

* discharged to home: 0
* Otherwise:1

**admission\_source:**

Label encoded values are:

* Admitted because of physician/clinic referral: 0
* Admitted from emergency room: 1
* Otherwise: 2

**admission\_type:**

Label encoded values are:

* Elective: 0
* Emergency: 1
* NULL: 2
* Newborn: 3
* Not Available: 4
* Not Mapped: 5
* Trauma Center: 6
* Urgent: 7

**diabetesMed:**

Label encoded values are:

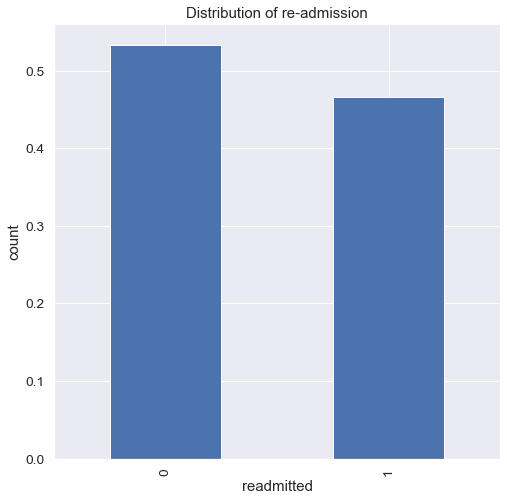
* No: 0
* Yes: 1

**IV. EXPLORATORY DATA ANALYSIS**

Exploratory Data Analysis (EDA) is an open-ended process where we make plots and calculate statistics in order to explore our data. The purpose is to to find anomalies, patterns, trends, or relationships. These may be interesting by themselves (for example finding a correlation between two variables) or they can be used to inform modeling decisions such as which features to use. In short, the goal of EDA is to determine what our data can tell us! EDA generally starts out with a high-level overview, and then narrows in to specific parts of the dataset once as we find interesting areas to examine.

**1.READMISSION:**

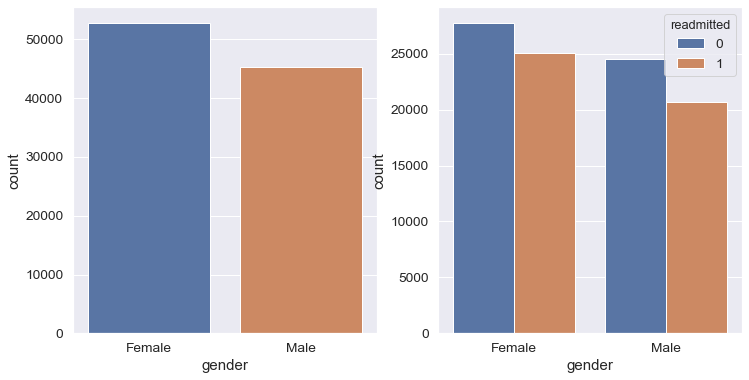
Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission. we consider **No as 0** and**<30 ,>30 as 1.**



* 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

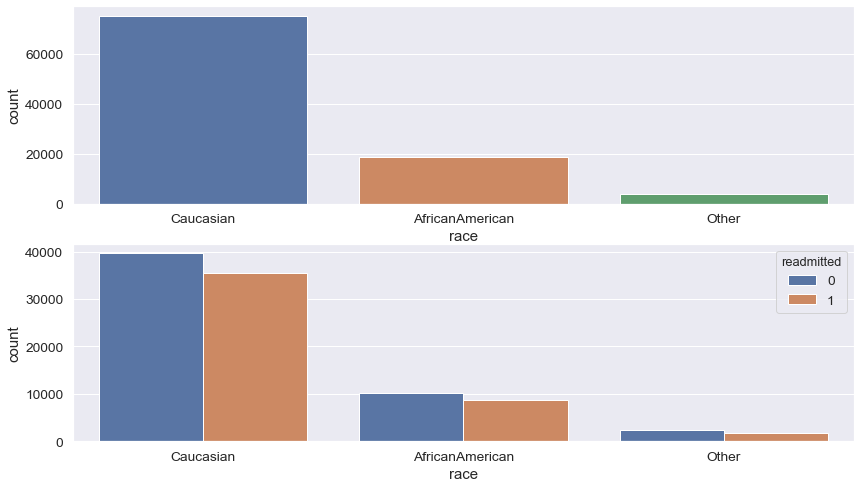


* The number of females are more than the number of males.
* Percentage of males getting readmitted is approximately 45% of total males.
* Percentage of females getting readmitted is approximately 47% of total females.
* Percentage of females getting readmitted are more than that of males.

**3.RACE:**

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

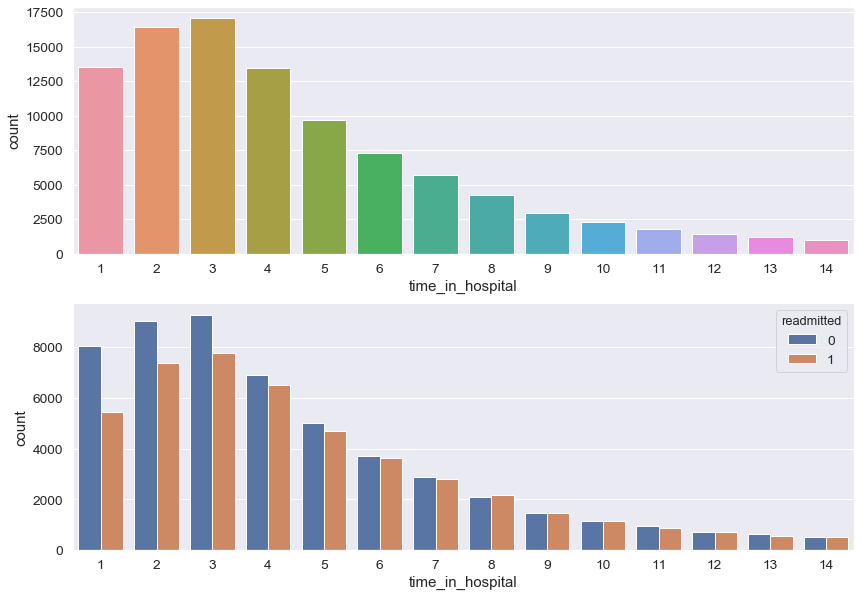
The relative percentage for Asian, Hispanic, and other are very less ,so together we consider all the three races as Other.



* The total number of Caucasians are signiicantly more than all other race.
* The percentage of readmissions as well as non-readmissions are more for Caucasians than any other race.
* The race of Asians are least in number.
* Approximately 90% of the Caucasians gets readmitted whereas 83% of the AfricanAmerican gets readmitting.

**4.TIME\_IN\_HOSPITAL:**

Integer number of days between admission and discharge

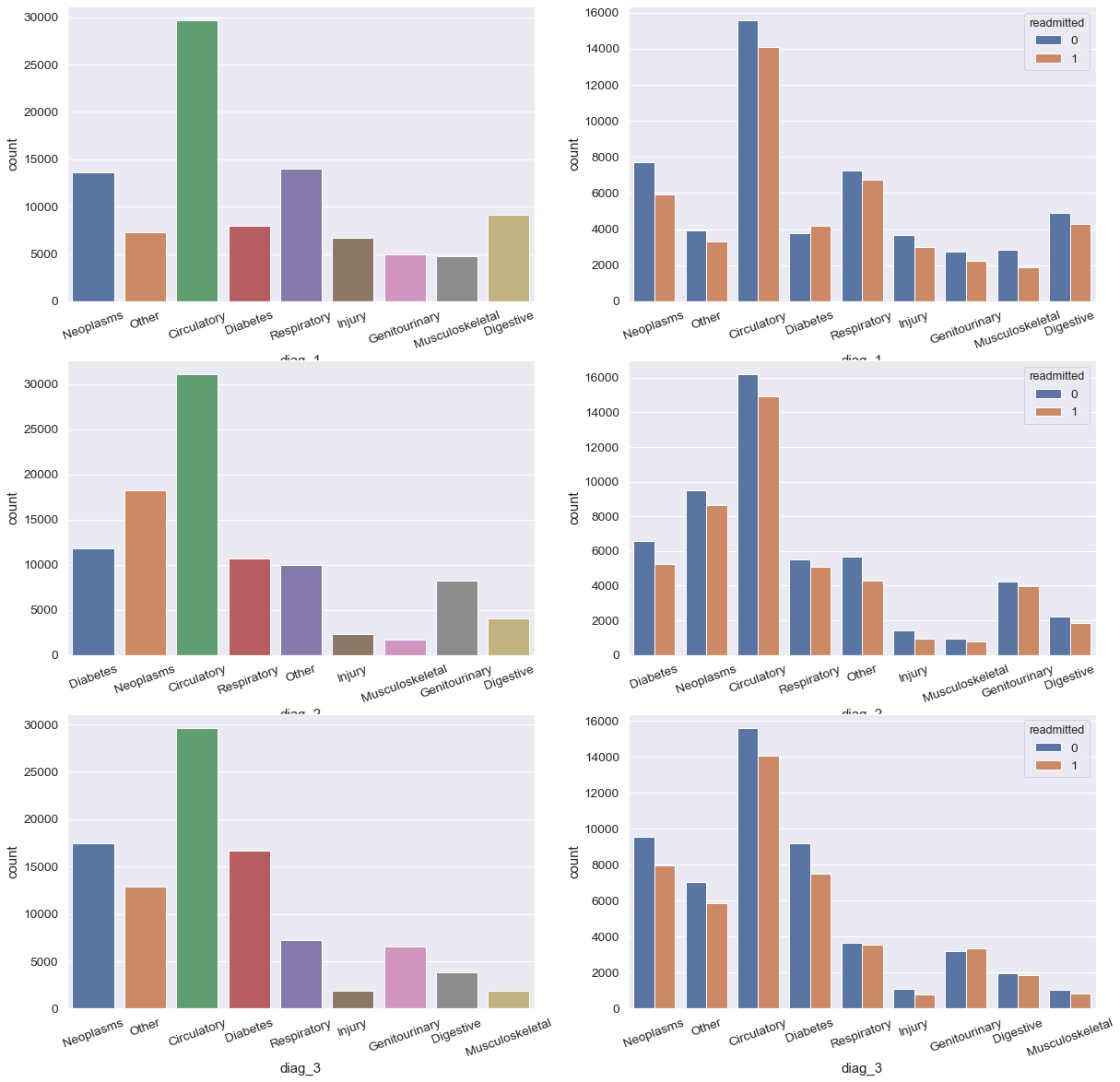


* The average number of days of admission is around 4 days.
* It is most likely that people readmitting those who were admitted for more than 7days in their first admission.
* Percentage of patients readmitting who have been admitted for 1day is very less compare to other patients.

**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



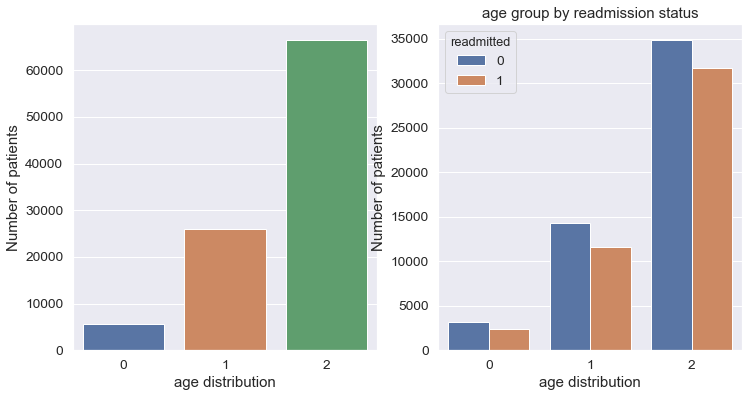
* The total number of patients under the primary diagnosis category of 'Circulatory' is maximum.
* The total number of patients under the primary diagnosis categories of 'Neoplasms' and 'Respiratory' are almost equal.
* The total number of readmissions under 'Diabetes' are more than the number of non-readmissions.
* Musculoskeletal has the least number of readmissions.

**6.AGE\_DISTRIBUTION:**

Grouped in 10-year intervals: (0, 10), (10, 20), …,( 90, 100)

**We had nine different categories earlier.We converted them into three categories**

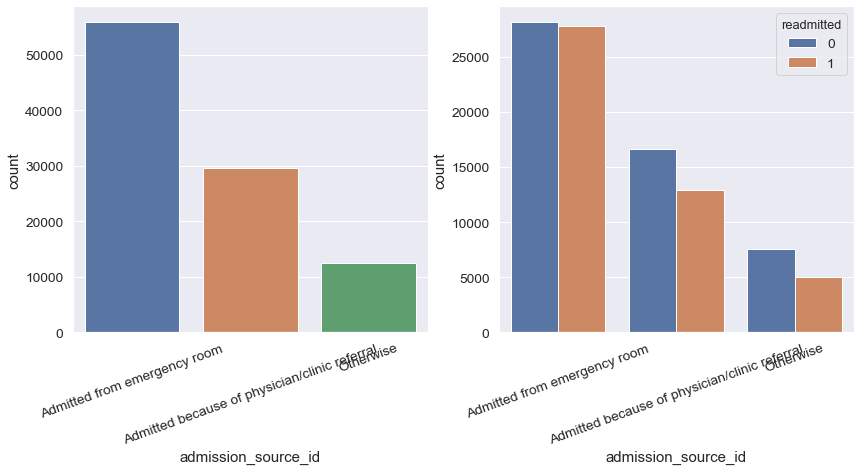
* **0-35**
* **45-60**
* **>60**



* Patients with age more than 60 years are more in number.
* Around 69% of the patients above 60 years are readmitting.

**7.ADMISSION SOURCE\_ID:**

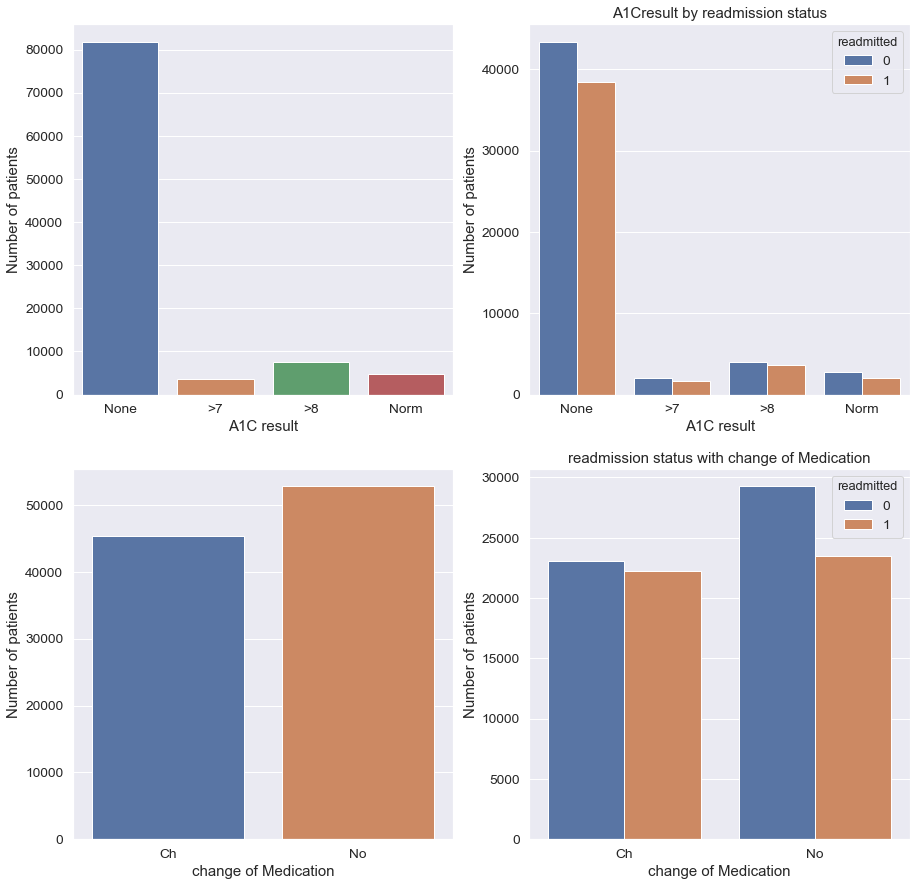
Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital



* Patients Admitted from emergency room are more in number than those admitted from any other admission source.
* Over 98.6% of the patients readmitted are those who admitted from emergency room.

**7.A1CRESULT OR CHANGE:**

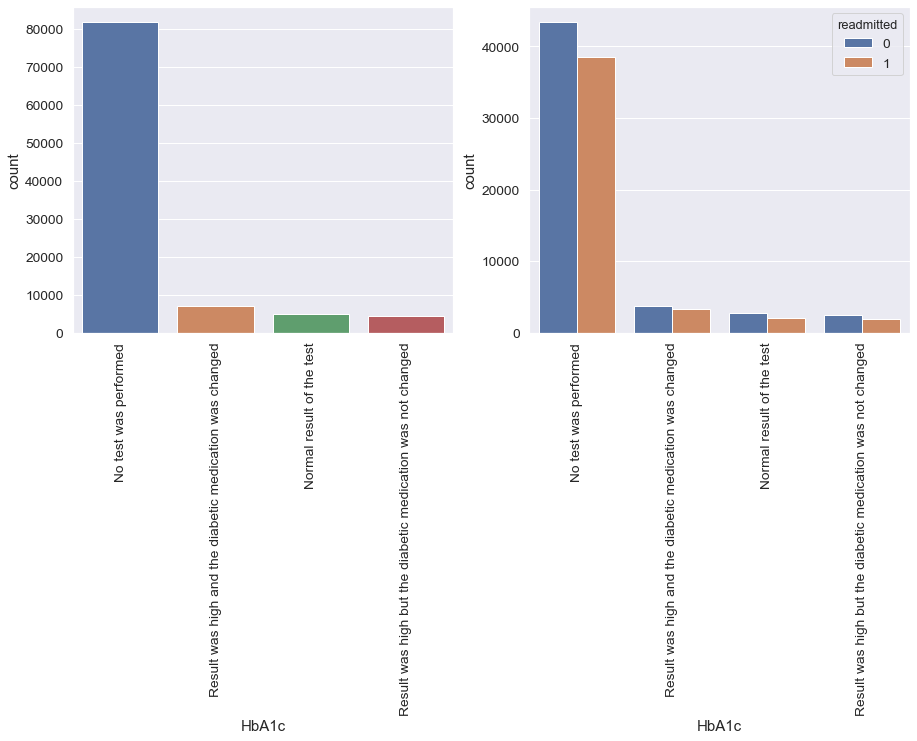
Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.



* Number of patients those who did not take A1C test are higher.
* Approximately 88.5% of patients those who did not take A1C test and 80% of patients who took A1C test readmitted.
* 42.02% of patients readmitted whose A1C test result was Normal.

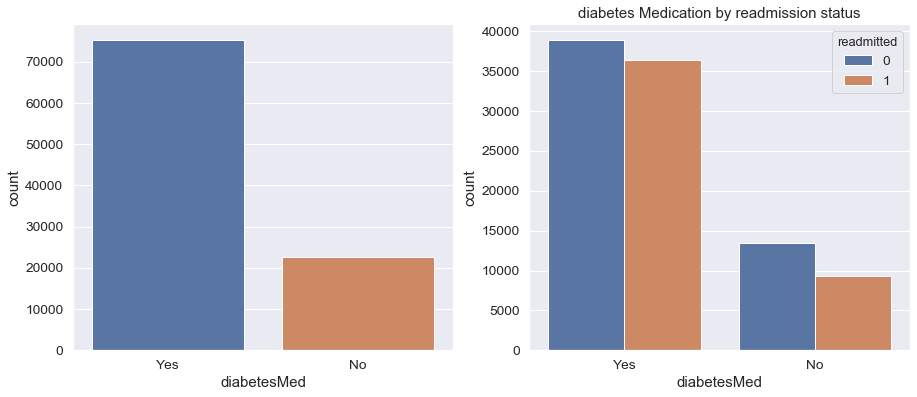
**HBA1C:**

We clubbed A1C result and change of medication together as HBA1C. We considered four groups of encounters: (1) no HbA1c test performed, (2) HbA1c performed and in normal range, (3) HbA1c performed and the result is greater than 8% with no change in diabetic medications, and (4) HbA1c performed, result is greater than 8%, and diabetic medication was changed



**8.DiabetesMed:**

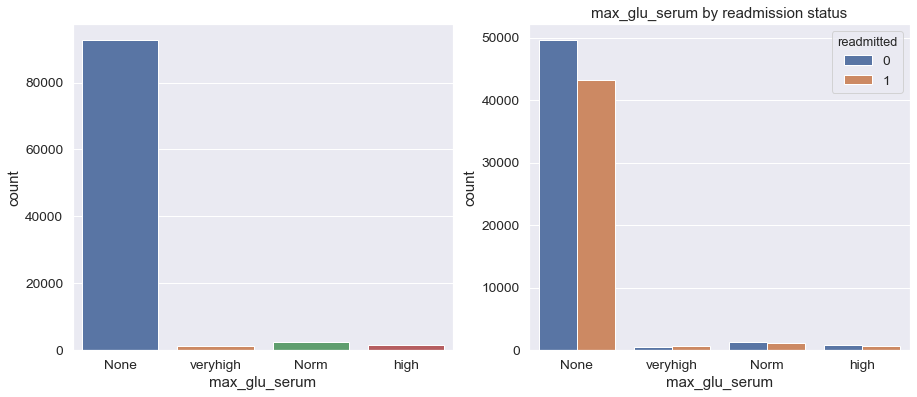
Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”



* From the total data around 76% are took medications and in that 95% are readmitted.
* Those who are didn't took medications around 67% are readmitted

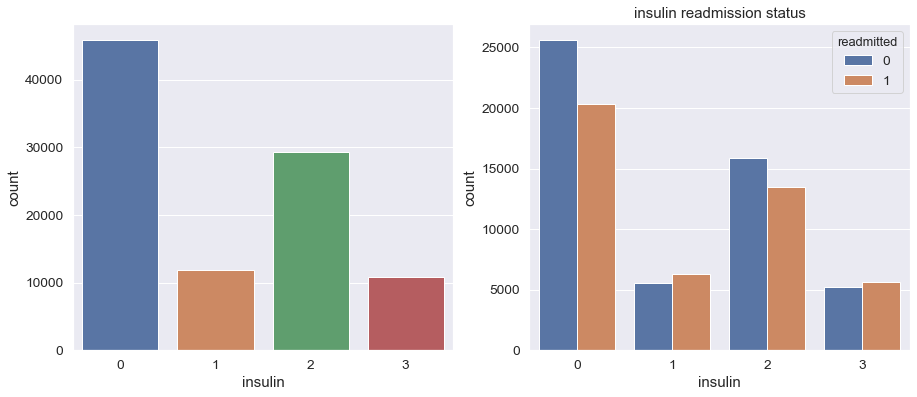
**9.Max\_glu\_serum:**

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



* Approx.,**95%** of patients are not taking max\_glu\_serum test
* Approx.,**46%** of patients who didnt take test readmitted
* Approx.,**56%** of the patients whose max\_glu\_serum values greater than 300 are readmitting

**10.INSULIN**



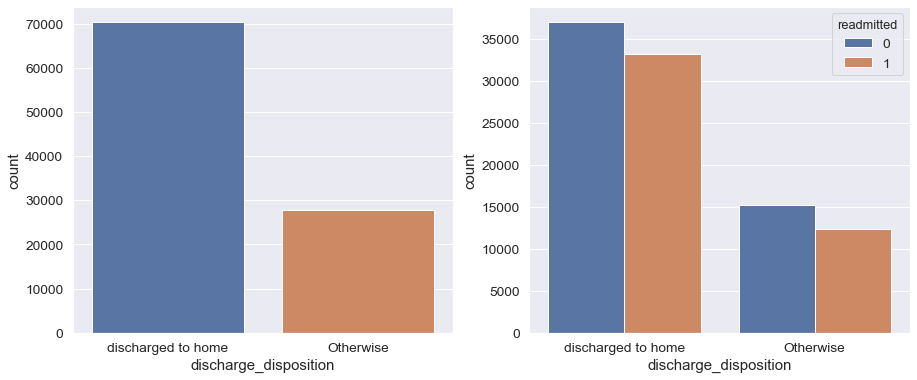
* insulin was not prescribed for 46% of the patiensts.
* there is no change in insulin dosage for Aprox..30%.
* readmission rate is more for the patients whose insulin dosage is either increase or decrease.

Note : except insulin all other medications dosage was steady.

**10.DISCHARGE DISPOSITION:**

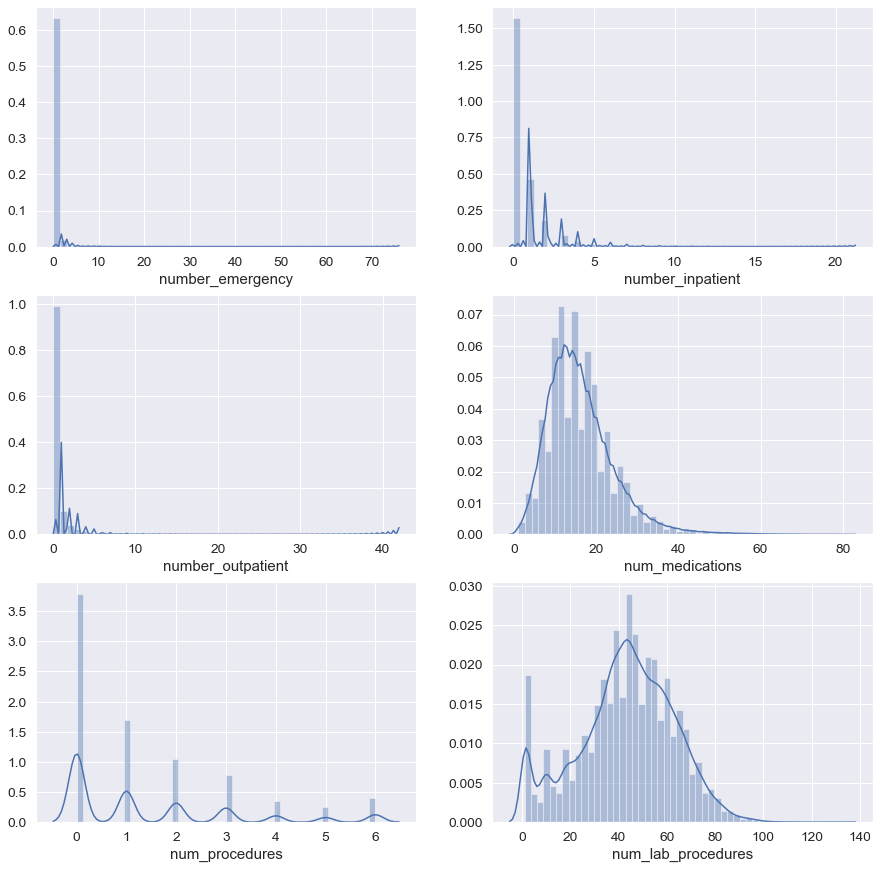
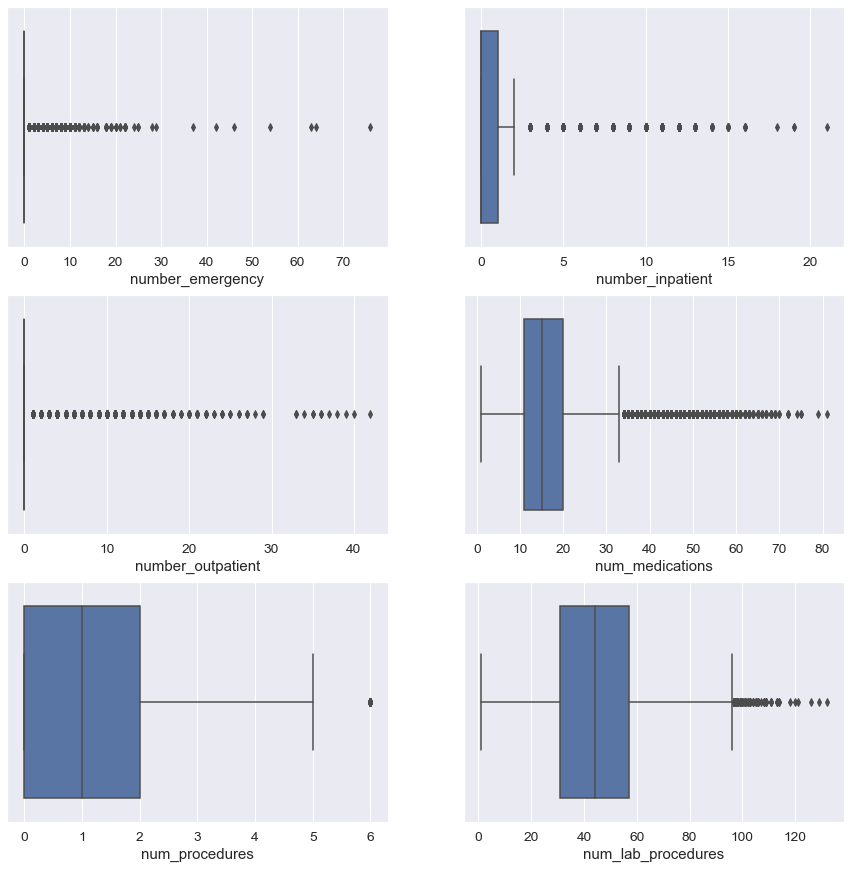
Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available. all these distinct values are converted into two categories

* Discharged to home.
* Otherwise.



* approx..72% of patients discharged to home .
* approx.. 73% of patients who discharged to home are readmitted.

**11.CHECKING OUTLIERS FOR NUMERICAL VARIABLES:**



* As we have outliers in many of the attributes we treat them using square root transformation and applied scaling on the features before running the model.

**VI. MODELLING:(Basic Model)**

Logistic Regression using all the variables except Id's ('encounter\_id','patient\_nbr'),we got Training score = 0.620,Test score = 0.619.

**Checking For Assumptions Of Regression:**

**1.Dependent binary variable**

* The dataset has only two values in its dependent variable

**2.No Repeated Observations**

* repeated observations are very less.

3.**No Multi-co linearity**

* From the VIF, all variables VIF value are less than 5.So,there is no multi-collinearity

**4. Large Sample size**

* good

From above we can notice that all the required assumptions are fulfilled.

# Feature Selection:

Now that we have explored the trends and relationships within the data, we can work on selecting a set of features for our models. In particular, we learned the following from EDA which can help us in engineering/selecting features:

**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 generalize better to new data and results in a more interpretable model. Generally, I think of feature selection 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 results of modeling, such as the feature importances 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 parameters 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 0.001) as significant features.

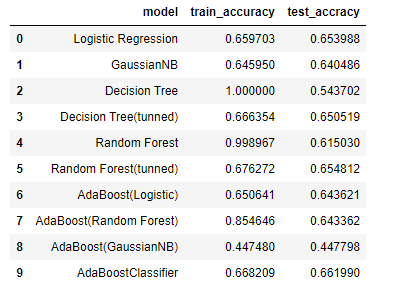
Those Features are :

* **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**
* **HbA1c**
* **Readmitted**
* **Models to Evaluate**

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

1. Logistic Regression
2. Naive-Bayes
3. K-Nearest Neighbors Classifier
4. Decision Tree Classifier
5. Random Forest Classifier

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.



Challenges:

* One of the major challenge was convertingdata to correct types and feature engineering as there are large number of categorical features.
* There was a lot of 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 discard 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 large number of categorical features.