A PROJECT REPORT ON

Predicting Readmission of Diabetic Patients and Feature Analysis

Submitted in partial fulfilment of the requirements for the post graduate program
in
Data Science & Engineering

by

Anusha N (SSID: Y447DOT2SX)
Aravind Raj C K (SSID: EO90W59L7A)
Debashis Das (SSID: K28IGYT5GP)
Ishan Chowdhury (SSID: 0ETV7RAXY8)
Nikhil Kumar Singh (SSID: GGASINRV0M)

Under the guidance of
Ms. Anjana Agrawal

GREAT LEARNING
Learning for life
HSR LAYOUT :: BANGALORE

ACKNOWLEDGEMENT

We welcome this opportunity to express our heartfelt gratitude and regards to our project guide Ms. Anjana Agrawal, for his unconditional guidance. We would also like to extend our gratitude to Ms. Shambhavi Shukla, Senior Data Scientist, Great Learning, Bangalore, for encouraging us to undertake the project and help us with the initial stages. We greatly admire and acknowledge the constant support we had from my friends and team members for all the effort and hard work that they have put into completing this project.

Anusha N (SSID: Y447DOT2SX)
Aravind Raj C K (SSID: EO90W59L7A)
Debashis Das (SSID: K28IGYT5GP)
Ishan Chowdhury (SSID: 0ETV7RAXY8)
Nikhil Kumar Singh (SSID: GGASINRV0M)

ABSTRACT

Hospital readmission is considered a key metric in order to assess health center performances. Indeed, readmissions involve different consequences such as the patient's health condition, hospital operational efficiency but also cost burden from a wider perspective. Prediction of 30-day readmission for diabetes patients is therefore of prime importance. The existing models are characterized by their limited prediction power, generalizability and pre-processing. This study proposes a comprehensive pre-processing framework in order to improve the model's performance while exploring and selecting prominent features for unplanned readmissions among diabetic patients. The pre-processing technique used includes comprehensive data cleaning, data reduction, and transformation. Random Forest algorithm for feature selection and SMOTE algorithm for data balancing are some example of methods used in the proposed pre-processing framework. The performance of the designed model was found, in this regard, particularly balanced across different metrics of interest with accuracy (f1-micro) and Area under the Curve (AUC) of 94% and close to the optimal recall of 92%.

CONTENTS

INTRODUCTION

What is Readmission?

Readmission is the process of being admitted to a place or organization again. Tracking the number of patients who experience readmissions to a hospital after a previous hospital stay is one category of data used to evaluate the quality of hospital care. The standard benchmark used by the Centers for Medicare & Medicaid Services (CMS) is the 30-day readmission rate. Rates at the 80th percentile or lower are considered optimal by CMS. 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 Ostling et al, patients with diabetes have almost double the chance of being hospitalized than the general population. Therefore, in this project we attempt to predict the hospital readmission for the patients with diabetes.

Industry Review

A survey conducted by the Agency for Healthcare Research and Quality (AHRQ) found that in the year 2011 more than 3.3 million patients were readmitted in the United States within 30 days of being discharged. Over $250 million was spent on a treatment of readmitted diabetic patients in 2011 (Hines et al., 2014). Current practice to identify at-risk diabetic patients are subjective: a clinician will assess the patient and decide what the appropriate care plan is for that individual. Research has shown that these subjective methods for determining readmission are slightly better than random guessing (Allaudeen et al., 2011). However, there are tools to objectively score readmission risk, such as LACE (van Walraven et al., 2010). These objective tools are seen to be useful because end-users can make these calculations manually and offer improved accuracy over subjective techniques. Machine learning models can be used to create objective models which then can be used to measure risk (Mingle, 2015). These models are more complex, but may be able to create more accurate risk predictions that should lead to improved diabetic patient outcomes. This study investigates the hypothesis that advanced machine learning techniques can make use of a wide set of clinical features to improve diabetic readmission risk prediction.

Background and Related Work

Many healthcare providers in the U.S. use LACE to identify at-risk patients. At its core LACE is a logistic regression model that makes use of a small set of features. LACE itself was derived from a set of 4812 patients, and validated on 1,000,000 patients using patient records from 2004 to 2008(van Walraven et al., 2010).

In addition, numerous previous studies have analyzed the risk factors that predict readmission rates of diabetic patients. However, much of the research is focused on subsets of diabetic populations and solutions are derived from a smaller sample size than this study. In some cases, the results were based on demographic and socioeconomic factors that influence readmission rates (Jiang et al., 2003). In some cases, the models are unspecific in target and focus on general readmission for all-cause (Hosseinzadeh, 2013). Our study considers data that covers demographic, clinical procedure-related and diagnostic-related features, as well as medication information for all ages to predict readmissions for diabetic patients within a 30-day window. Our goal does not analyse readmission cost as this is well documented by other researchers.

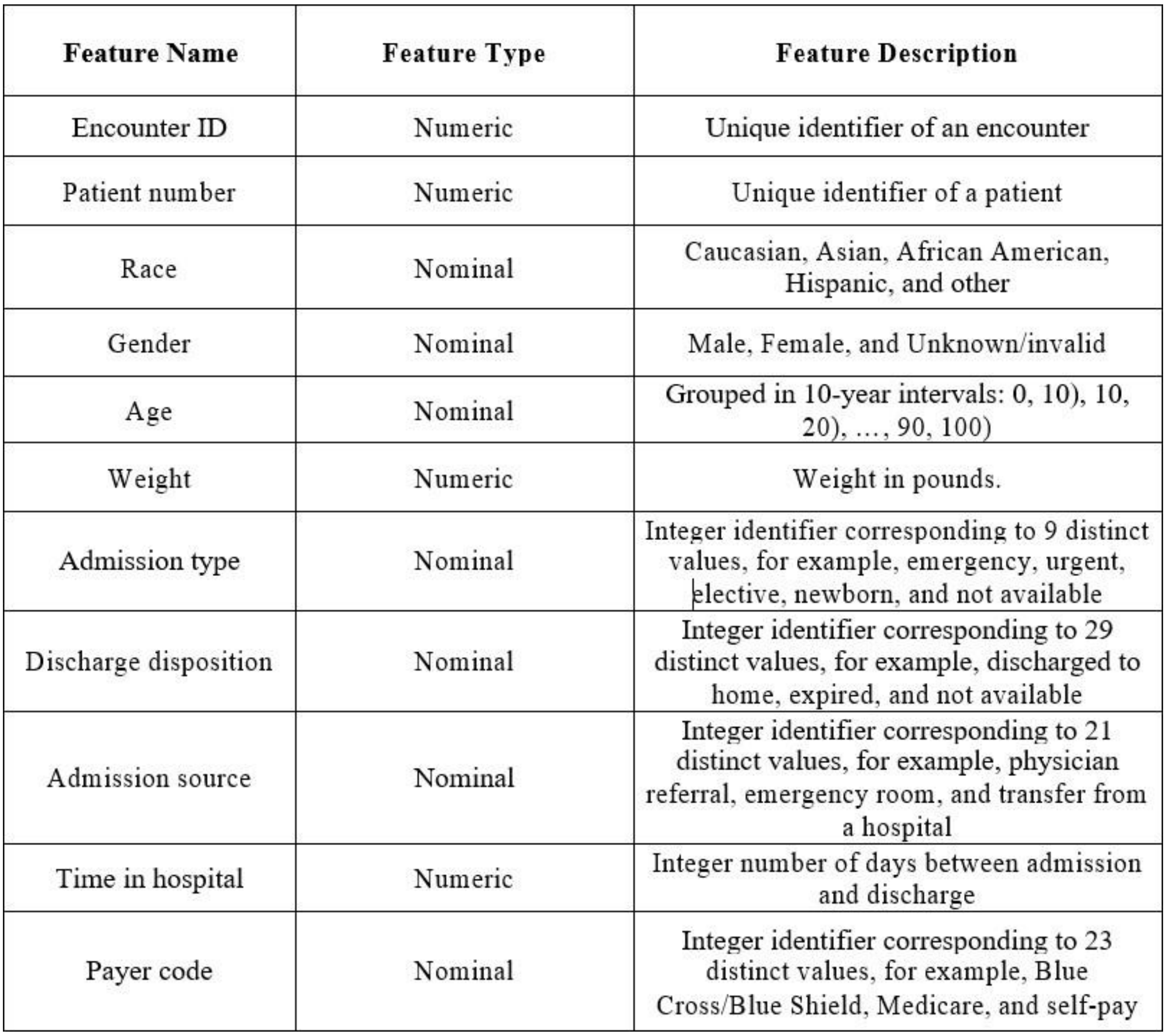
Problem Statement

Predicting whether a given patient having diabetes will get readmitted within 30 days or not given attributes such as race, gender, diagnosis, medication, glucose and changes in other drug prescriptions. The project will also trace out interdependence amongst the features and will aim at providing high level of interpretability as the readers should be able to comprehend the decisions made by the model given the nature of the domain this project is related with.

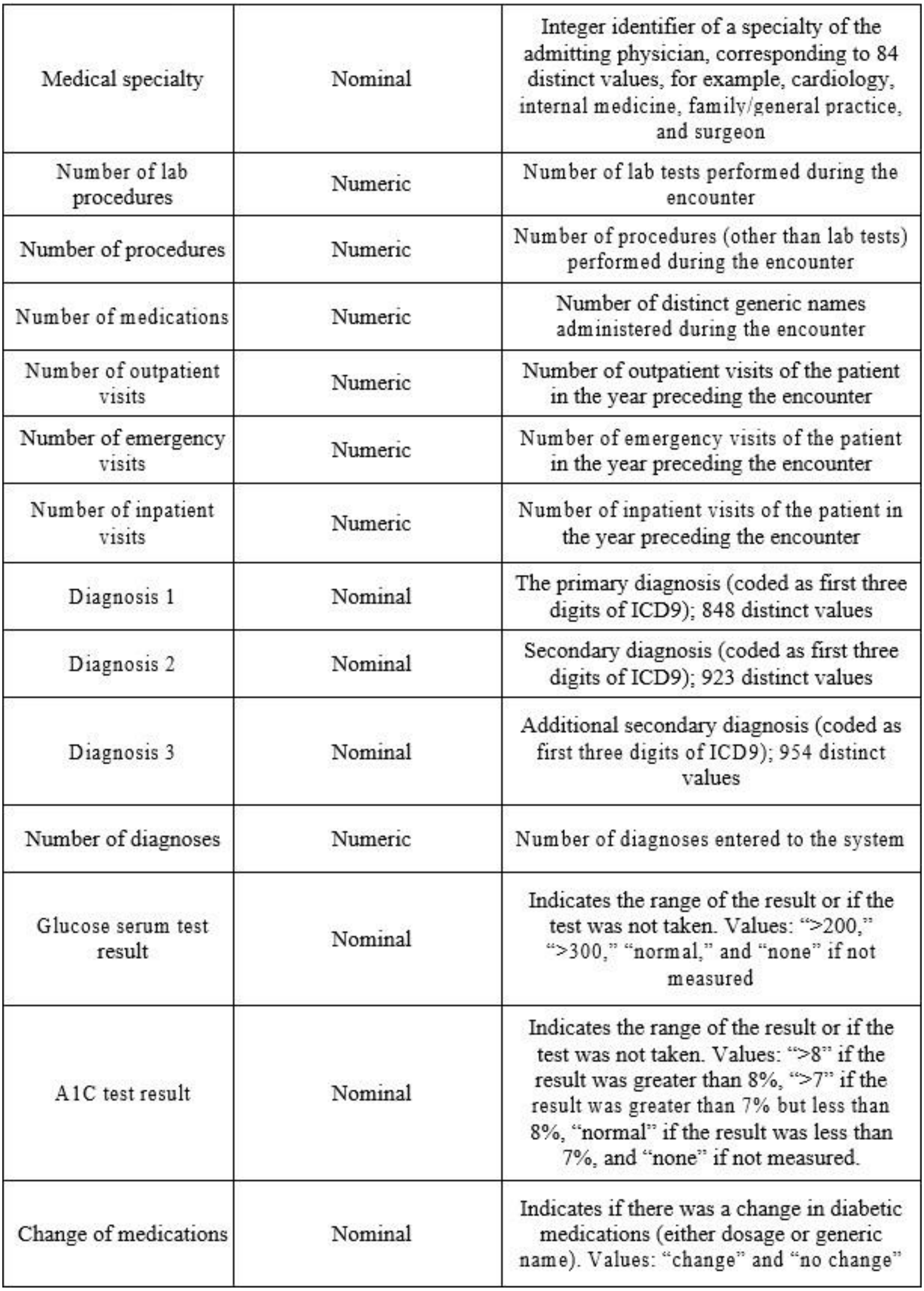
PRE-PROCESSING

Data Description

The dataset contains data collected on patients with diabetes from 130 hospitals over a period of 10 years. Different treatments and outcomes have been measured in the data. It consists of 50 features and records of 101766 patients. Data is available at http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

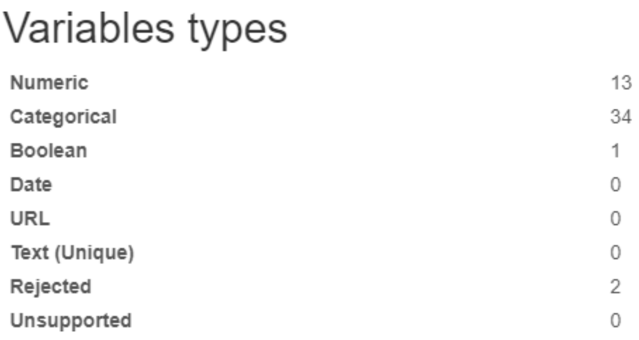


Feature Name	Feature Type	Feature Description
Encounter ID	Numeric	Unique identifier of an encounter
Patient number	Numeric	Unique identifier of a patient
Race	Nominal	Caucasian, Asian, African American, Hispanic, and other
Gender	Nominal	Male, Female, and Unknown/invalid
Age	Nominal	Grouped in 10-year 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



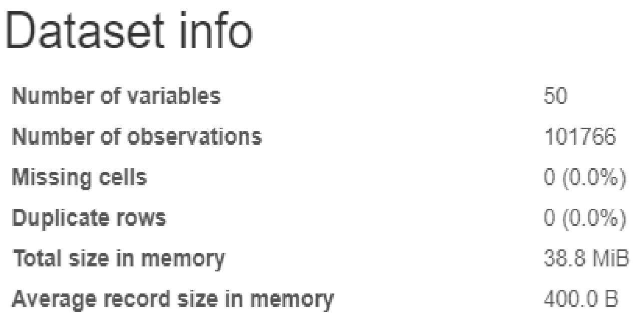
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon
Number of lab procedures	Numeric	Number of lab tests performed during the encounter
Number of procedures	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	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
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.



Variables types

Numeric	13
Categorical	34
Boolean	1
Date	0
URL	0
Text (Unique)	0
Rejected	2
Unsupported	0



Dataset info

Number of variables	50
Number of observations	101766
Missing cells	0 (0.0%)
Duplicate rows	0 (0.0%)
Total size in memory	38.8 MiB
Average record size in memory	400.0 B

Data Cleaning

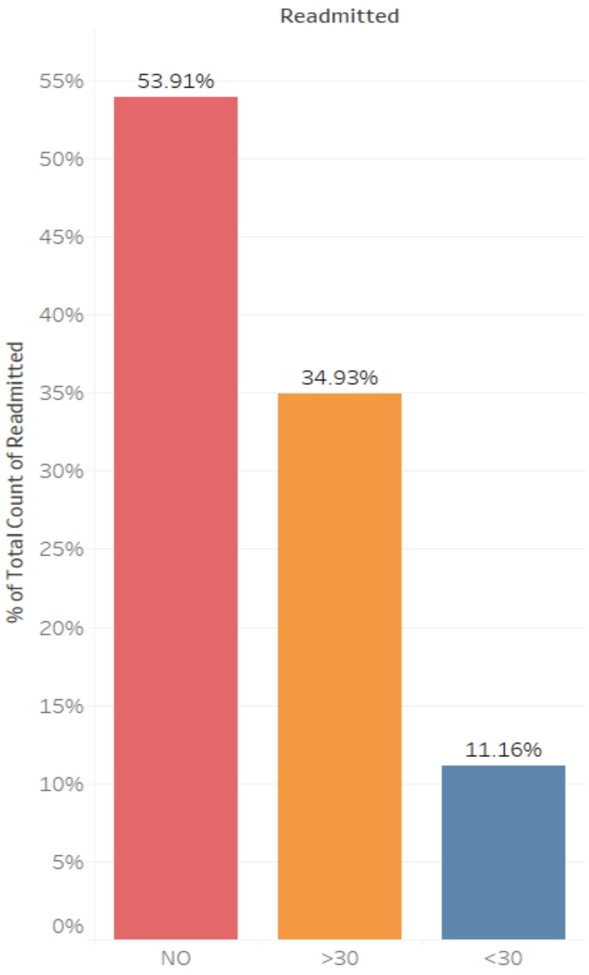
- In the dataset there are many question marks (?) present. In our analysis these values were firstly converted to Null values and were then treated accordingly.
- We have dropped ‘Weight’, ‘Payer code’ and ‘Medical_speciality’ columns as it had more than 40% of missing values.
- In the ‘Race’ column, we have replaced missing values with mode since we have less than 5% missing values in that particular column.
- Certain codes for discharge_disposition_nbr were assigned to patients who are dead or in hospice, those data points were also dropped from the analysis.
- We dropped the variables ‘Examide’, ‘glimepiride-pioglitazone’ and ‘Citoglipton’ because all the values in the variables were same, i.e., ‘No’. Both the rows didn’t add any information to our data.
- In the dataset, the features which contains numeric values are of type Discrete Quantitative and has a finite set of values. Discrete data can be both Quantitative and Qualitative. So, treating outliers in this dataset is not possible.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a general approach to exploring datasets by the means of simple graphic visualizations in order to gain a deeper understanding of the data as well as to gain insights of any hidden pattern in the data. It is a practice to understand the data first and try to gather as many insights from it. The scope of EDA is more pronounced in the given dataset because of the nature of the attributes we have in hand. Further, the practice of EDA not only accounts for the detection of anomalies and hidden patterns but also testing of hypotheses and assumptions surrounding the dataset. Thus, a comprehensive EDA will help in the retention of gainful insights and inferences about the data.

Univariate Analysis

- **Distribution of Readmitted**



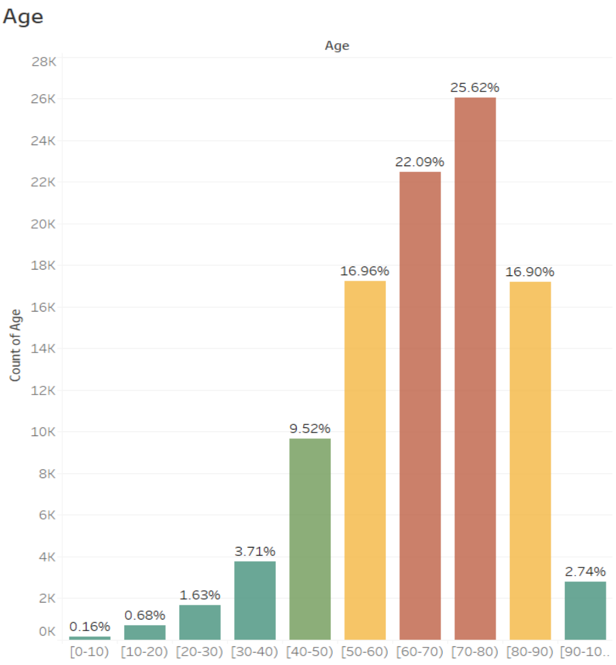
Inference:

- Less than 15% patients were readmitted in less than 30
- The number of counts for readmission before 30 days is very less
- The number of counts for readmission after 30 days is comparatively more than readmission before 30 days

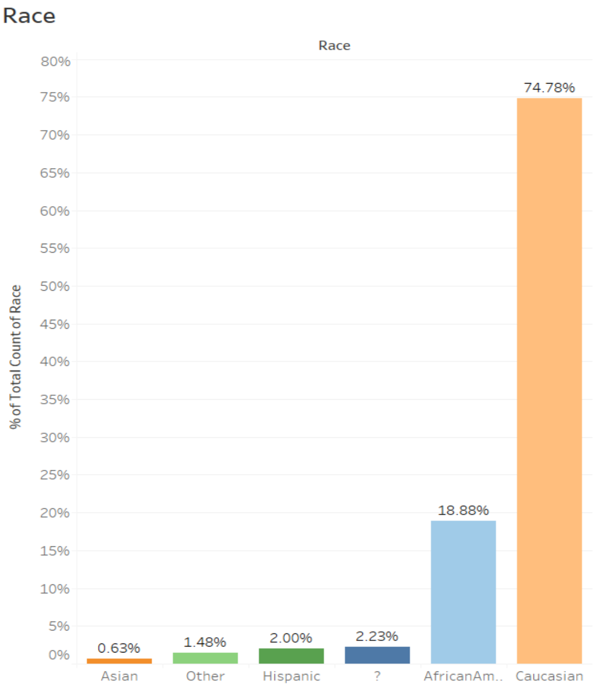
- **Distribution of Age**

Inference:

- Most of the patients having diabetes are in the group 50-80.
- The peak is at patients of 70-80 years of age.
- There are also some special cases below 30 years of age which contribute around 2.09% of the total patients.
- We see a fall in the number of patients who are above 80 years of age, which might seem counterintuitive.



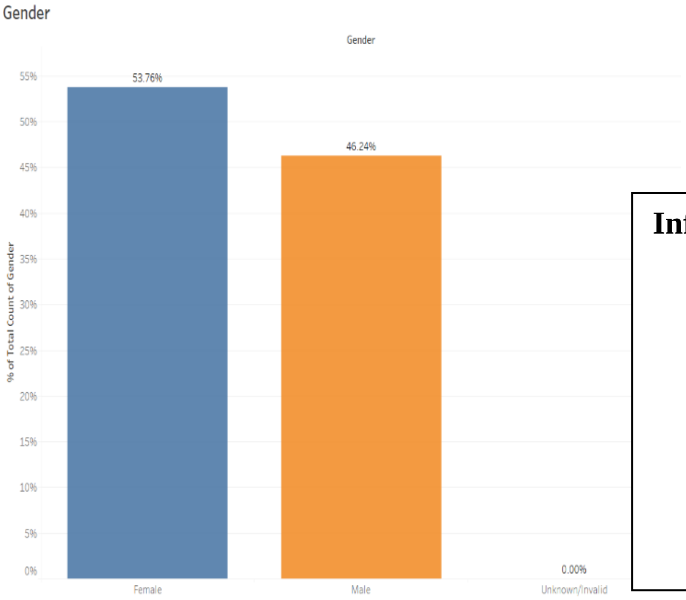
- **Distribution of Race**



Inference:

- We have records of 5 race types of patients, i.e., Caucasian, African American, Hispanic, Asian and Other.
- 75.2% of the total patients are Caucasian.
- The character "?" here represents the missing values which are 2.2% of the total values
- Asians have the least number of entries in our dataset.

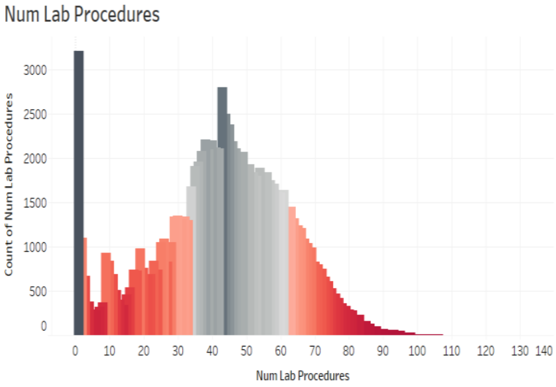
- **Distribution of Gender**

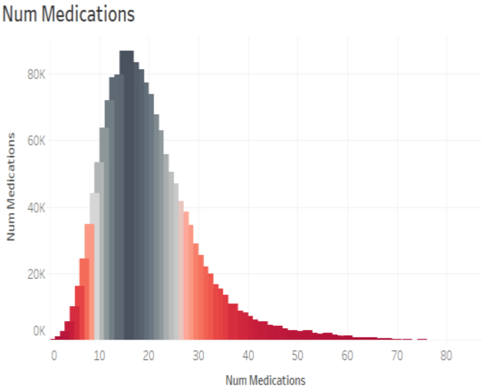


Inference:

- According to the data set, more females were diagnosed with diabetes.
- There are some Unknown/Invalid entries in the dataset which needs to be dealt with.

- **Distribution of num_lab_procedures and num_medication**

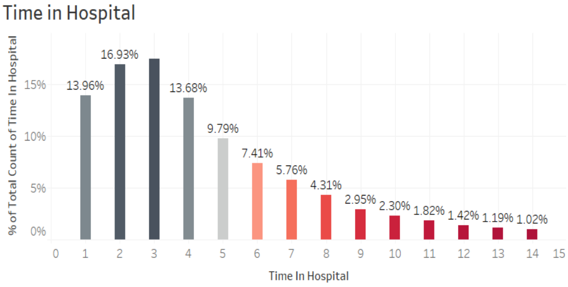


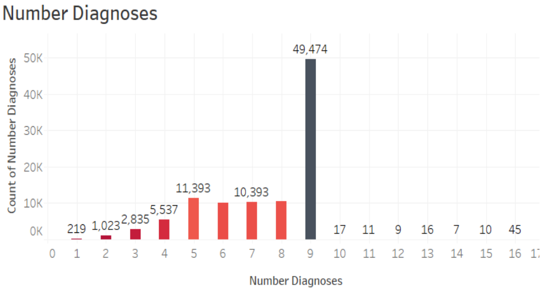


Inference:

- There is only one patient who performed the maximum number of lab test i.e. 132.
- Maximum number of patients have only one lab test. i.e Most of the patients performed only one test.
- The average number of lab test performed by most of the patients is 42.
- Most of the patients were given between 10-25 numbers of medicines.
- There are some cases were number of medication given to a patient is greater than 70 with max value being 81.

- **Distribution of time_in_hospital and num_of_diagnoses**



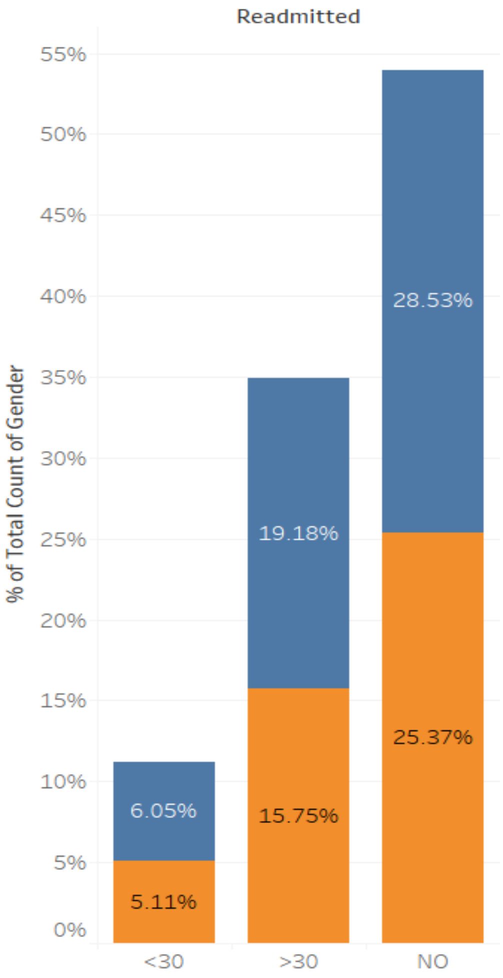


Inference:

- The minimum time spent in hospital by a patient is 1 day and the maximum being 14 days.
- About 50% of the patients have spent between 1-3 days in the medical facility.
- The Maximum value of the Number Diagnosis column has been capped to 9, which is also having the maximum number of entries.

Bivariate Analysis

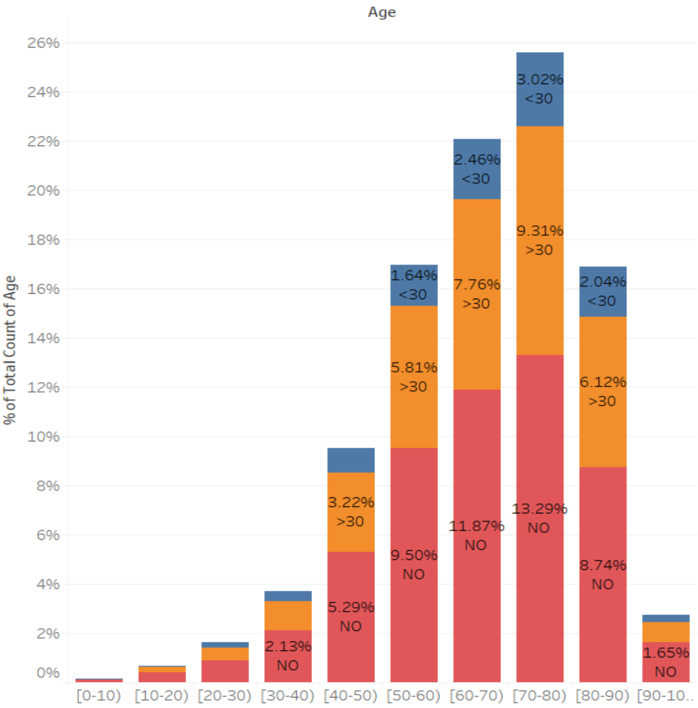
- **Gender vs Readmitted**



Inference:

- Female has readmitted more in <30 days.
- Majority is of 'no record of readmission' in male and female.

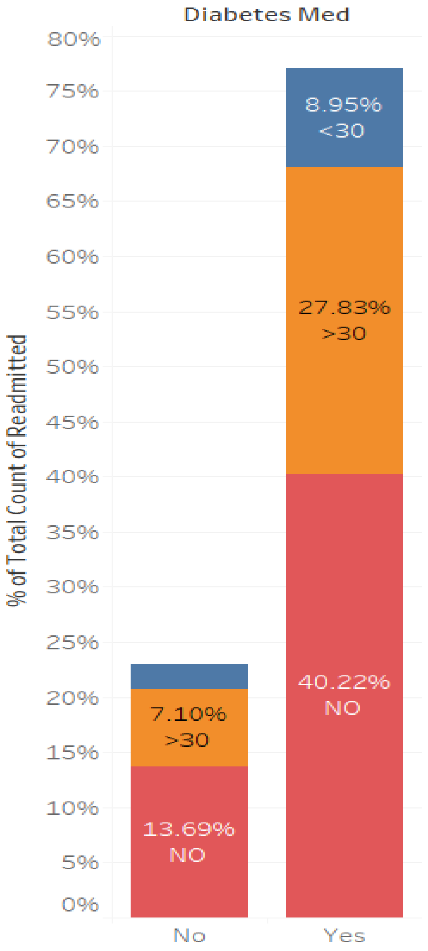
- **Age vs Readmitted**



Inference:

- Patients in the age group of 0-10 are not much affected by diabetes and hence we see minimum number of entries.
- Majority all the age groups having 'no readmission'.
- Patients in the age group of 70-80 are the once who are getting readmitted (<30) the most.

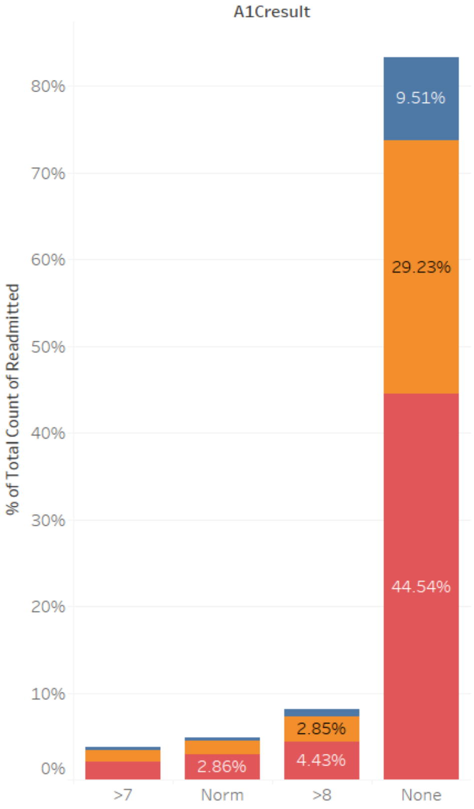
- **DiabMed vs Readmitted**



Inference:

- About 9% of the people with diabetic medication were again readmitted in less than 30 days.
- Most of the people who didn't get readmitted were already into Diabetes medication.

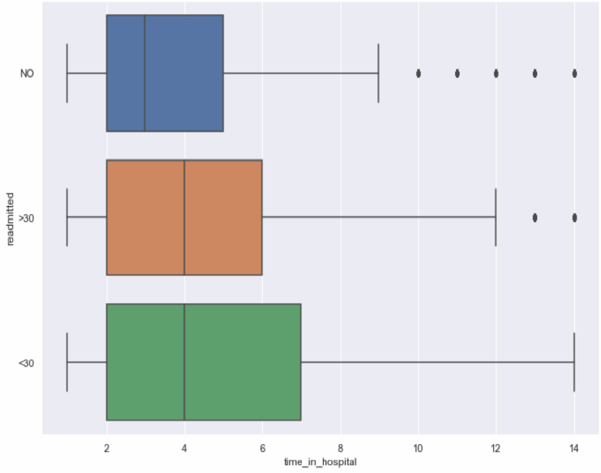
- **A1CResult vs Readmitted**

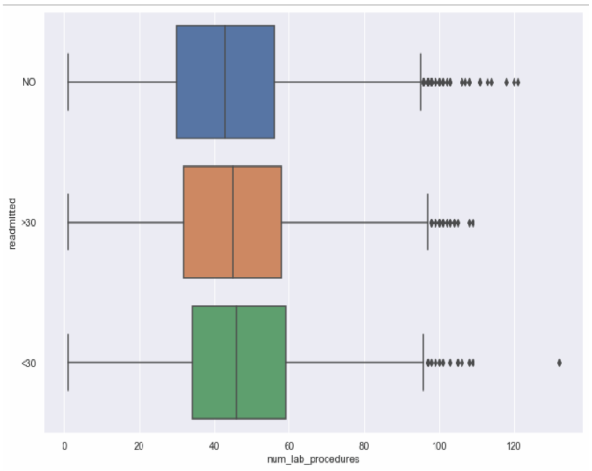


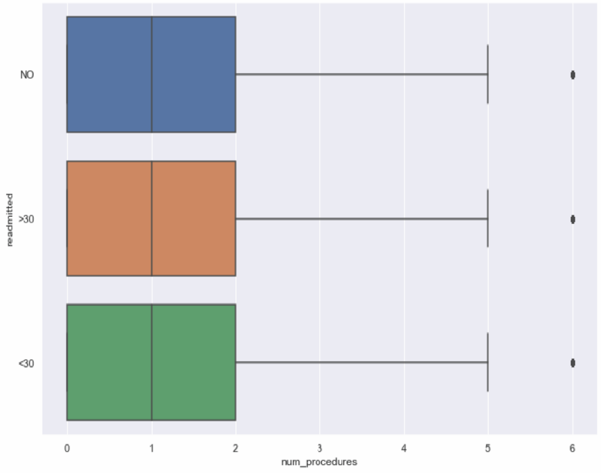
Inference:

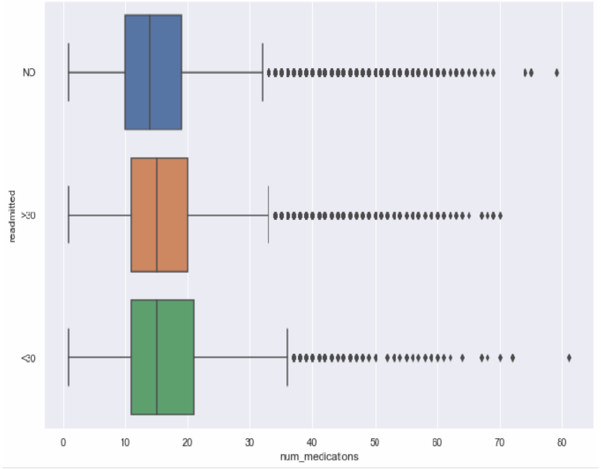
- Patients A1C test is not being measured that is 'none' having the more number of 'no readmission' cases, average number of >30 days of readmission.
- The 'none' value has number of patients readmitted in <30 days.

Outlier Checking









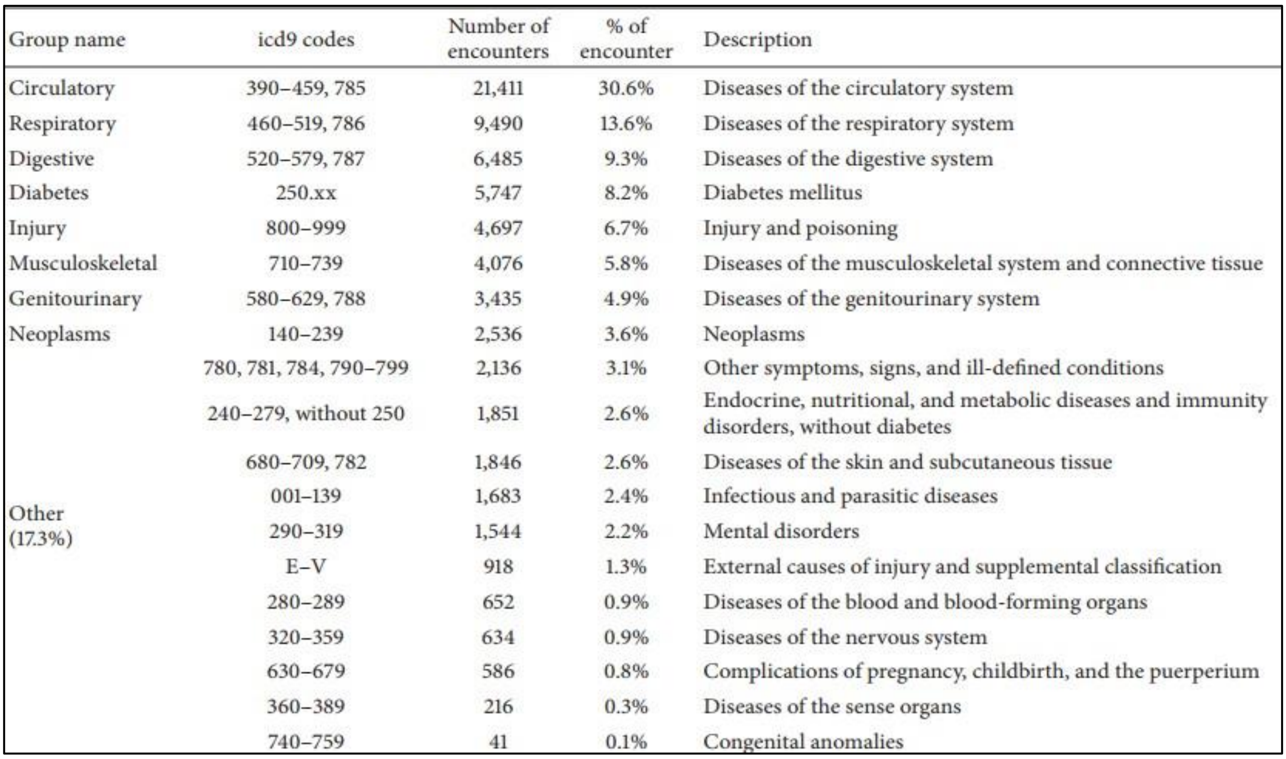
Observations

- For the Patients who didn’t get readmitted we see a few outliers in the ‘time_in_hospital’ column.
- ‘Num_lab_procedures’ column has uniform distribution across all categories with respect to the readmitted column.

FEATURE ENGINEERING

Given the ambiguity that we face because of some of the features in this dataset, the importance of Feature Engineering gets even more magnified for this analysis. Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. In this project the following steps and measures were taken to simplify some of the critical features in this dataset without any loosing up on any relevant information.

- The entries in the diag_1, diag_2, diag_3 are about the primary, secondary and additional secondary diagnosis which were performed during encounter. These variables had more than 800 levels and each of these values were coded in accordance with the nature of disease they relate to. In our analysis we have remapped these values to their respective diseases and thereby have reduced the number of levels.

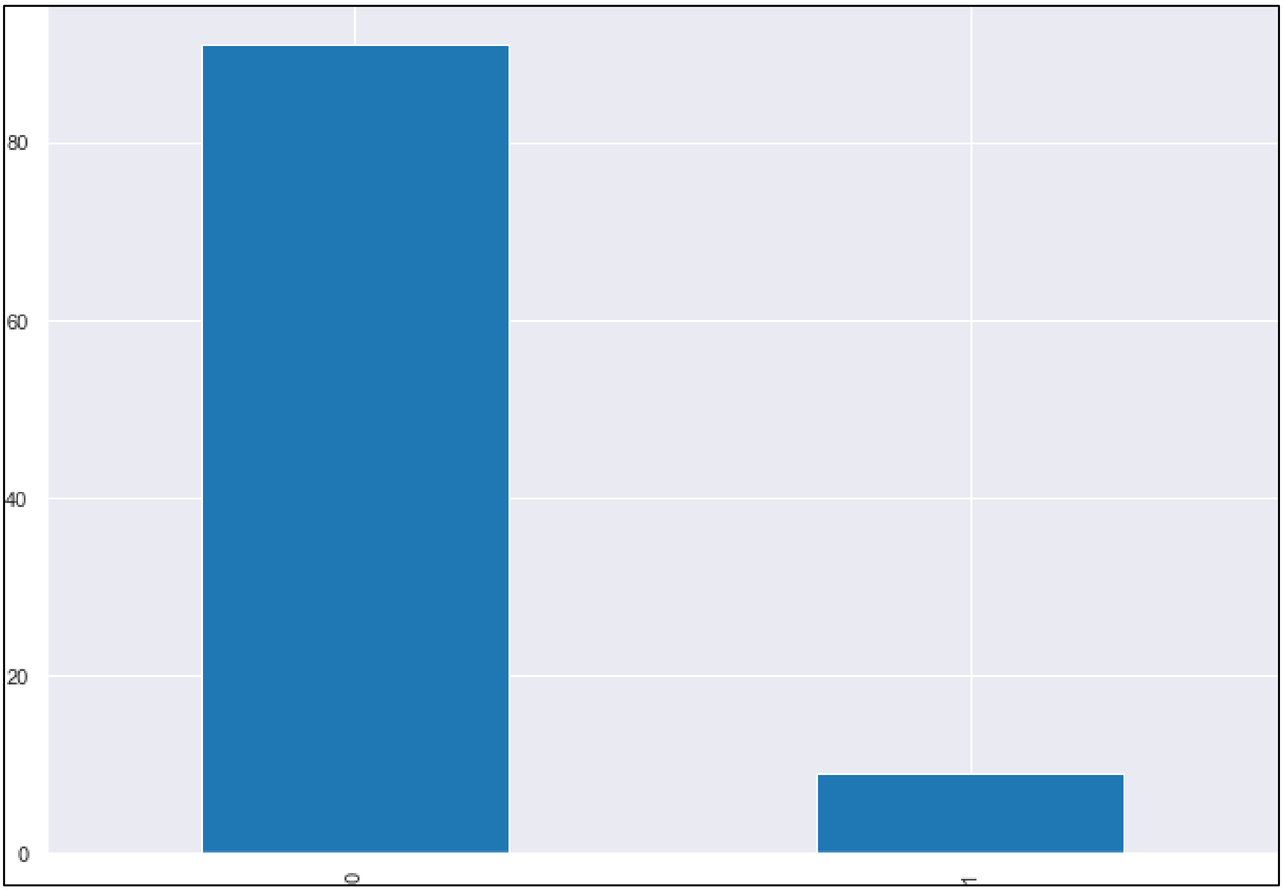


Group name	icd9 codes	Number of encounters	% of encounter	Description
Circulatory	390–459, 785	21,411	30.6%	Diseases of the circulatory system
Respiratory	460–519, 786	9,490	13.6%	Diseases of the respiratory system
Digestive	520–579, 787	6,485	9.3%	Diseases of the digestive system
Diabetes	250.xx	5,747	8.2%	Diabetes mellitus
Injury	800–999	4,697	6.7%	Injury and poisoning
Musculoskeletal	710–739	4,076	5.8%	Diseases of the musculoskeletal system and connective tissue
Genitourinary	580–629, 788	3,435	4.9%	Diseases of the genitourinary system
Neoplasms	140–239	2,536	3.6%	Neoplasms
Other (17.3%)	780, 781, 784, 790–799	2,136	3.1%	Other symptoms, signs, and ill-defined conditions
	240–279, without 250	1,851	2.6%	Endocrine, nutritional, and metabolic diseases and immunity disorders, without diabetes
	680–709, 782	1,846	2.6%	Diseases of the skin and subcutaneous tissue
	001–139	1,683	2.4%	Infectious and parasitic diseases
	290–319	1,544	2.2%	Mental disorders
	E–V	918	1.3%	External causes of injury and supplemental classification
	280–289	652	0.9%	Diseases of the blood and blood-forming organs
	320–359	634	0.9%	Diseases of the nervous system
	630–679	586	0.8%	Complications of pregnancy, childbirth, and the puerperium
	360–389	216	0.3%	Diseases of the sense organs
	740–759	41	0.1%	Congenital anomalies

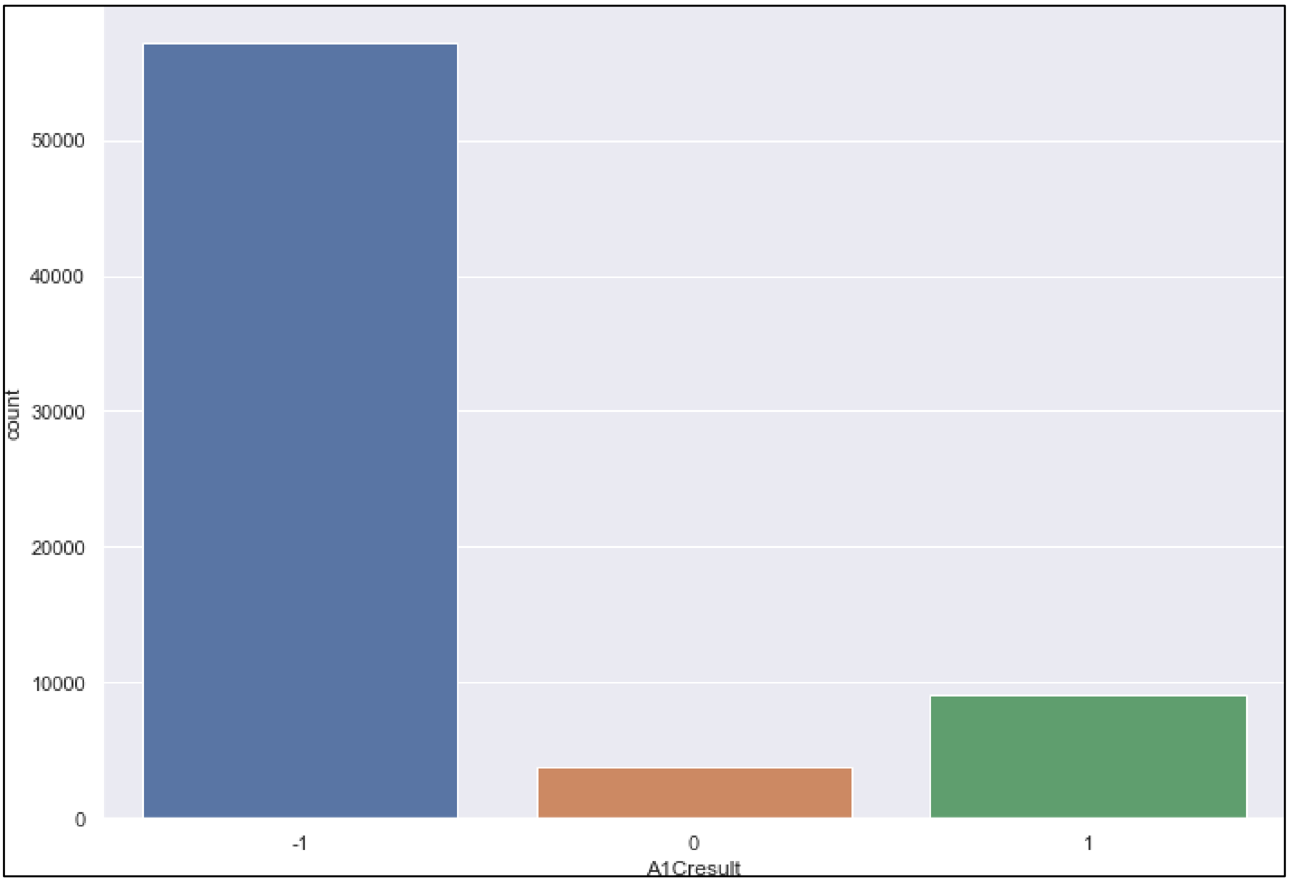
- The ‘AGE’ feature had values depicted in range of tens, we categorized them into the max- value of the age ranges they were assigned. Example: If a patient was assigned (40-50], we assigned 50 as their age.

- The variables ‘total_outpatients’, ‘number_inpatients’, ‘number_emergency’ all were summed up to create a a new variable “total_visits”.

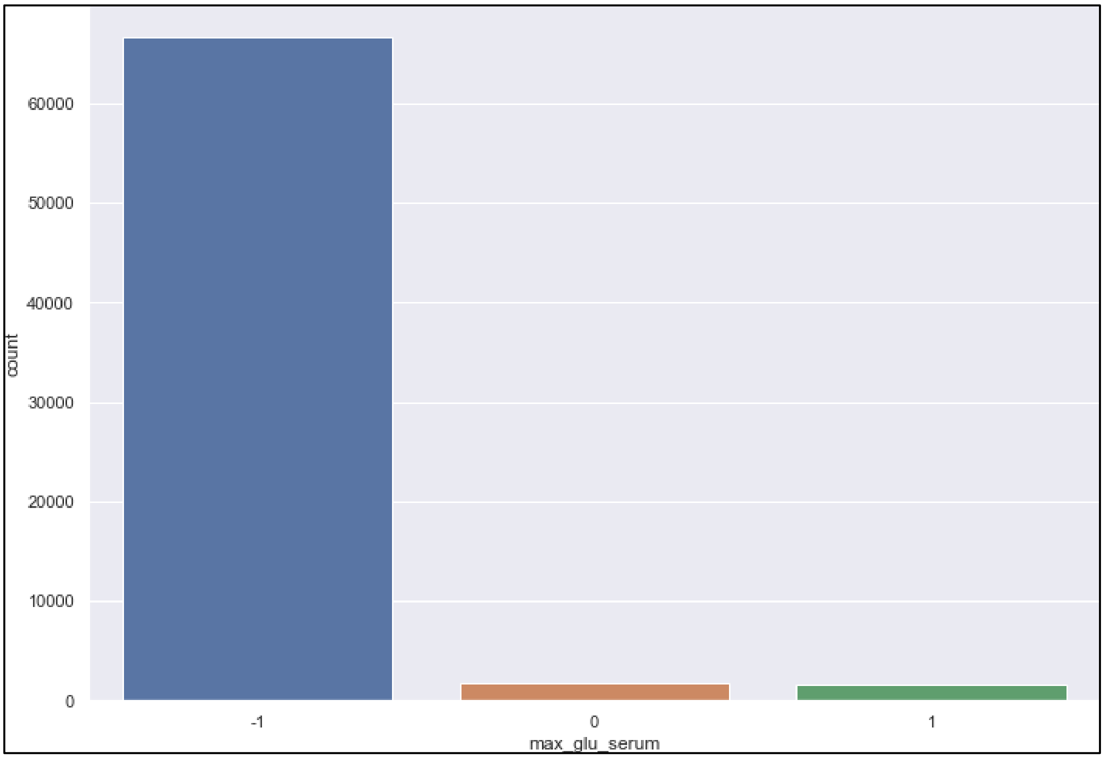
- The readmitted column was value were then reassigned with “>30” and “No” categorized as 0 while the “<30” has been categorized as 1.



- The feature ‘A1Cresult’ is the variable containing levels of Glycated Hemoglobin. Glycated hemoglobin is a form of hemoglobin that is chemically linked to a sugar. The usual sugar is glucose. The formation of the sugar-Hb linkage indicates the presence of excessive sugar in the bloodstream, often indicative of diabetes. So, we categorized ‘A1Cresult’ above 7 to the value ‘1’, A1Cresult ‘Normal’ i.e. below 5.7% to ‘0’ and the ‘A1Cresult’ that has not been calculated to ‘-1’.



- The feature ‘max_glu_serum’ is the glucose level of the patients in mg/dl. If this value is greater than 200 then the patient tends to be diabetic. Accordingly, we categorized the values greater than 200 to ‘1’, ‘Normal’ to ‘0’ and ‘None’ to ‘-1’.



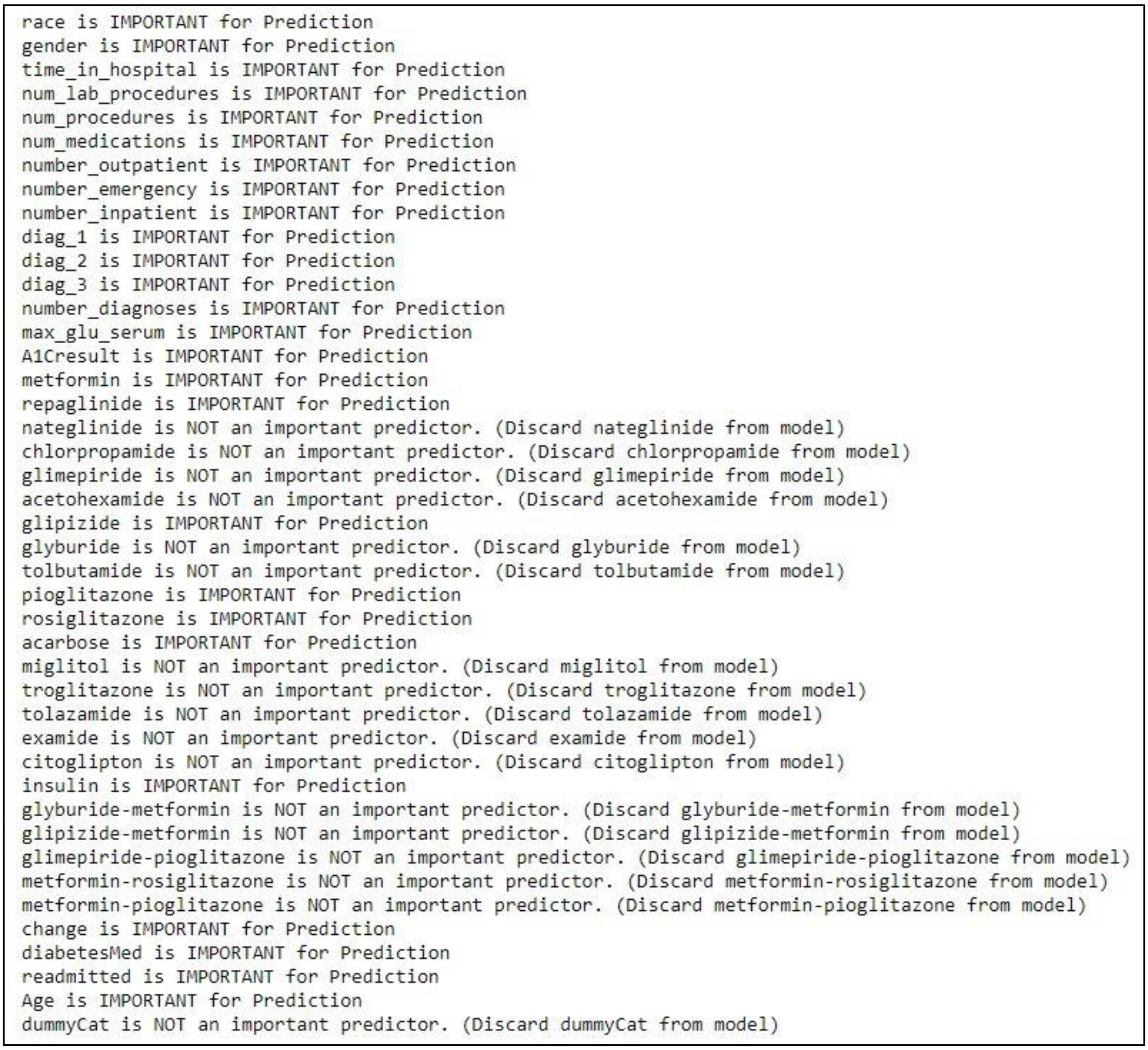
- In the dataset, the features which contains numeric values are of type Discrete Quantitative and has a finite set of values. Discrete data can be both Quantitative and Qualitative. So, treating outliers in this dataset is not possible.
- By performing Chi-2Contengency Test with target variable (Readmitted) we got to know that all other medicines except for insulin are irrelevant. We have to determine which treatment is working well solo insulin or conjunction of other drugs with insulin, so wherever the value of insulin alone is 1 we have considered it as solo insulin (insulin) and wherever there is combination of insulin and other drugs we have considered it as insulin + others (I + O). There are some patients who were solely treated with medicines other than insulin and some patients were not given any medications which were separately accounted for.

STATISTICAL ANALYSIS

Statistical testing was done to look at the evidence for a particular hypothesis being true. This helped us to accomplish decisions. Hypothesis will be performed to statistically validate the impact of respective independent variables on the outcome. Since all the variables are discrete, we will perform Chi-Square test to test feature significance.

- **Hypothesis**

  Ho: *The feature is not significant predictor of Target.*
  Ha: *The feature is significant predictor, i.e., it has high association with Target.*



```
race is IMPORTANT for Prediction
gender is IMPORTANT for Prediction
time_in_hospital is IMPORTANT for Prediction
num_lab_procedures is IMPORTANT for Prediction
num_procedures is IMPORTANT for Prediction
num_medications is IMPORTANT for Prediction
number_outpatient is IMPORTANT for Prediction
number_emergency is IMPORTANT for Prediction
number_inpatient is IMPORTANT for Prediction
diag_1 is IMPORTANT for Prediction
diag_2 is IMPORTANT for Prediction
diag_3 is IMPORTANT for Prediction
number_diagnoses is IMPORTANT for Prediction
max_glu_serum is IMPORTANT for Prediction
A1Cresult is IMPORTANT for Prediction
metformin is IMPORTANT for Prediction
repaglinide is IMPORTANT for Prediction
nateglinide is NOT an important predictor. (Discard nateglinide from model)
chlorpropamide is NOT an important predictor. (Discard chlorpropamide from model)
glimepiride is NOT an important predictor. (Discard glimepiride from model)
acetohexamide is NOT an important predictor. (Discard acetohexamide from model)
glipizide is IMPORTANT for Prediction
glyburide is NOT an important predictor. (Discard glyburide from model)
tolbutamide is NOT an important predictor. (Discard tolbutamide from model)
pioglitazone is IMPORTANT for Prediction
rosiglitazone is IMPORTANT for Prediction
acarbose is IMPORTANT for Prediction
miglitol is NOT an important predictor. (Discard miglitol from model)
troglitazone is NOT an important predictor. (Discard troglitazone from model)
tolazamide is NOT an important predictor. (Discard tolazamide from model)
examide is NOT an important predictor. (Discard examide from model)
citoglipton is NOT an important predictor. (Discard citoglipton from model)
insulin is IMPORTANT for Prediction
glyburide-metformin is NOT an important predictor. (Discard glyburide-metformin from model)
glipizide-metformin is NOT an important predictor. (Discard glipizide-metformin from model)
glimepiride-pioglitazone is NOT an important predictor. (Discard glimepiride-pioglitazone from model)
metformin-rosiglitazone is NOT an important predictor. (Discard metformin-rosiglitazone from model)
metformin-pioglitazone is NOT an important predictor. (Discard metformin-pioglitazone from model)
change is IMPORTANT for Prediction
diabetesMed is IMPORTANT for Prediction
readmitted is IMPORTANT for Prediction
Age is IMPORTANT for Prediction
dummyCat is NOT an important predictor. (Discard dummyCat from model)
```

As we can see that most of the features here, which is shown as not important, are medicines that are used to treat diseases so we need to combine the medicines to make the medicines relevant.

MODEL BUILDING

DEFINE X AND Y VARIABLE

Now, before applying any model, we have prepared the data and segregate the features and the label of the dataset. Variable X contains all the independent variables that are necessary to make prediction. Variable y has readmitted as target variable.

- **Creating Dummy for X Variables**

  The variable x contains all the independent variables which are necessary to make the prediction. Here there are many categorical columns which are in text format as object data type. As we cannot build the model using the categorical text data, we have created dummy variables which creates numerical column.

- **Independent Variables used for model building**
  - 'time_in_hospital',
  - 'num_lab_procedures',
  - 'num_procedures',
  - 'num_medications',
  - 'number_diagnoses',
  - 'max_glu_serum',
  - 'A1Cresult',
  - 'Age',
  - 'total_visits',
  - 'race_Asian',
  - 'race_Caucasian',
  - 'race_Hispanic',
  - 'race_Other',
  - 'gender_Other',
  - 'diag_1_Diabetes',
  - 'diag_1_Digestive',
  - 'diag_1_Genitourinary',
  - 'diag_1_Injury',
  - 'diag_1_Muscoloskeletal',
  - 'diag_1_Neoplasms',
  - 'diag_1_Others',
  - 'diag_1_Respiratory',

  - 'diag_2_Diabetes',
  - 'diag_2_Digestive',
  - 'diag_2_Genitourinary',
  - 'diag_2_Injury',
  - 'diag_2_Muscoloskeletal',
  - 'diag_2_Neoplasms',
  - 'diag_2_Others',
  - 'diag_2_Respiratory',
  - 'diag_3_Diabetes',
  - 'diag_3_Digestive',
  - 'diag_3_Genitourinary',
  - 'diag_3_Injury',
  - 'diag_3_Muscoloskeletal',
  - 'diag_3_Neoplasms',
  - 'diag_3_Others',
  - 'diag_3_Respiratory',
  - 'change_No',
  - 'diabetesMed_Yes',
  - 'treatment_insulin',
  - 'treatment_no drug',
  - 'treatment_other'

- **Using Cross Val Predict**

  We are using Cross Val Predict in the model with cv =10 instead of train test split so that no pattern is untouched in data and we get a better f1 score and roc_auc score.

ALGORITHMS USED

Since our problem is a classification problem, we will be using the following algorithms in modelling:

- Logistic Regression
- KNN
- Tree Based Models

- Decision Tree Classifier
- Random Forest Classifier

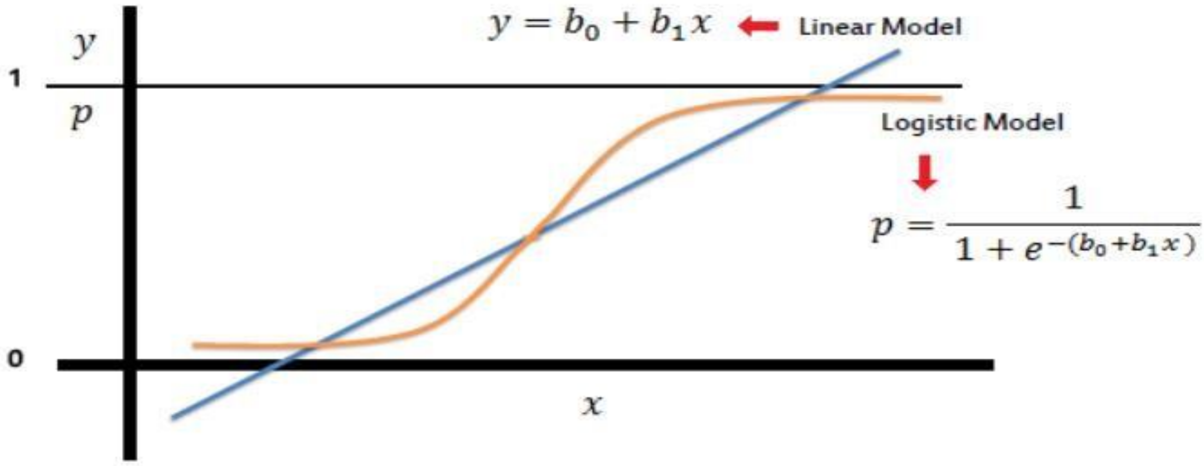
- **Logistic Regression**

  Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

  A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)

  Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

  On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



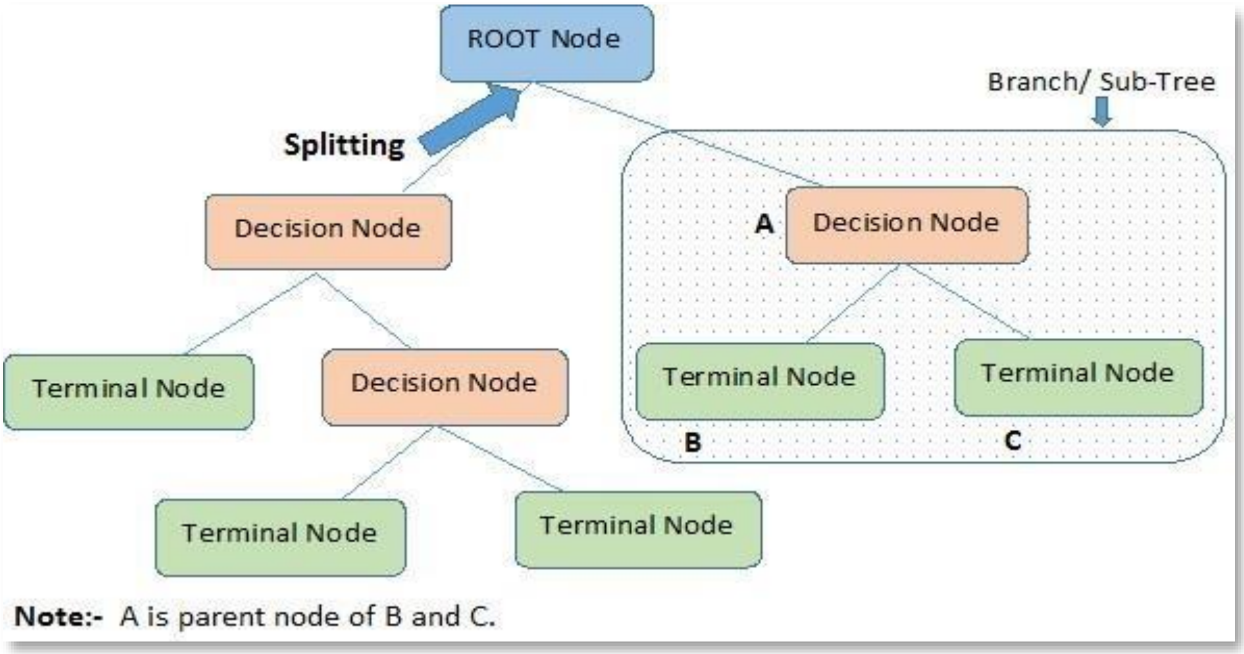
- **KNN (K-Nearest Neighbors)**

  KNN is a simple yet powerful classification algorithm. It requires no training for making predictions, which is typically one of the most difficult parts of a machine learning algorithm. The KNN algorithm has been widely used to find document similarity and pattern recognition.

  The intuition behind the KNN algorithm is one of the simplest of all the supervised machine learning algorithms. It simply calculates the distance of a new data point to all other training data points. The distance can be of any type e.g. Euclidean or Manhattan etc. It then selects the K-nearest data points, where K can be any integer. Finally, it assigns the data point to the class to which the majority of the K data points belong.

- **Decision Tree Classifier(CART)**

  A Decision tree (CART) is a schematic, tree-shaped diagram used to determine a course of action or show a statistical probability. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.



Note:- A is parent node of B and C.

- **Random Forest Classifier**

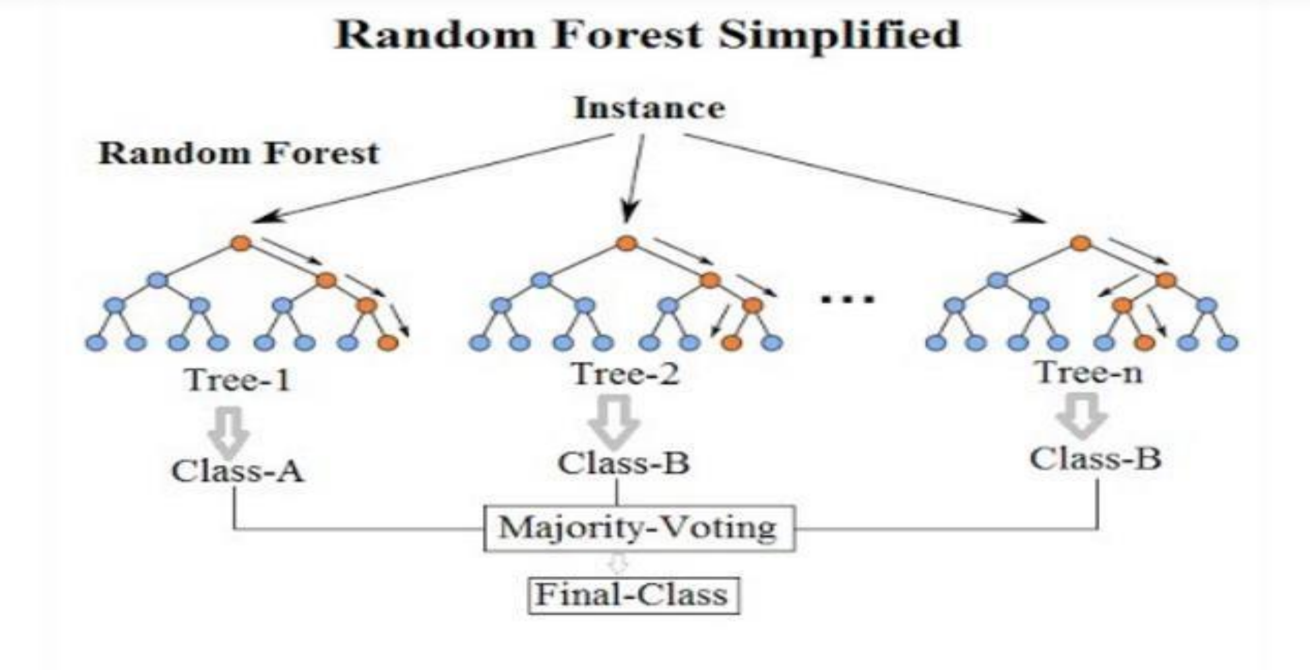
  Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

  To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

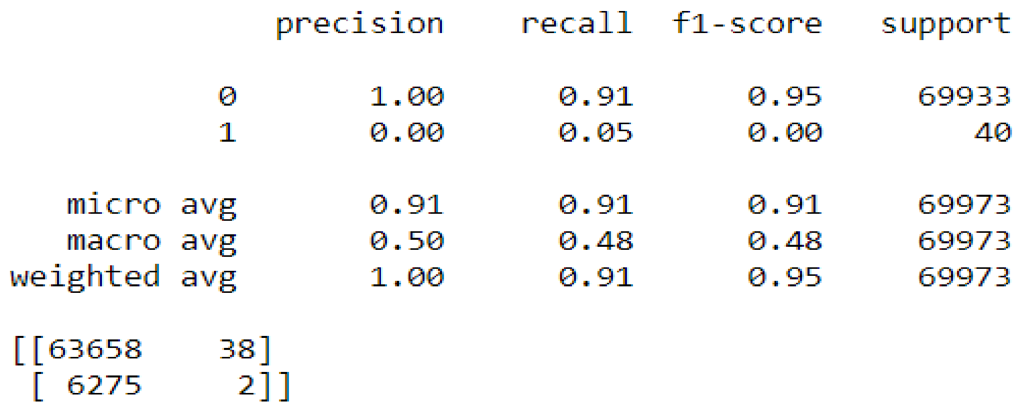
  One big advantage of random forest is, that it can be used for both classification and regression problems.

  Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Fortunately, we don't have to combine a decision tree with a bagging classifier and can just easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest regressor.

  Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

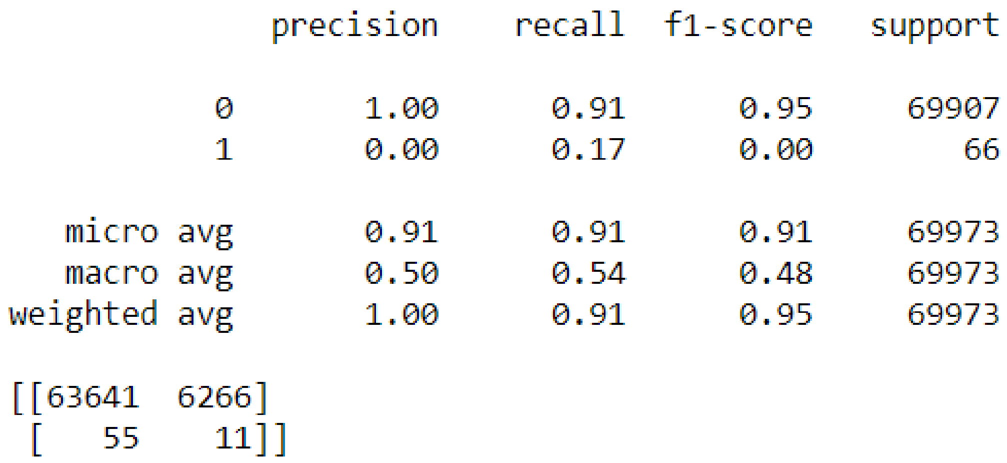


- **MODELING AND RESULTS**
  - **Classification Report for Logistic Regression (Base Model)**



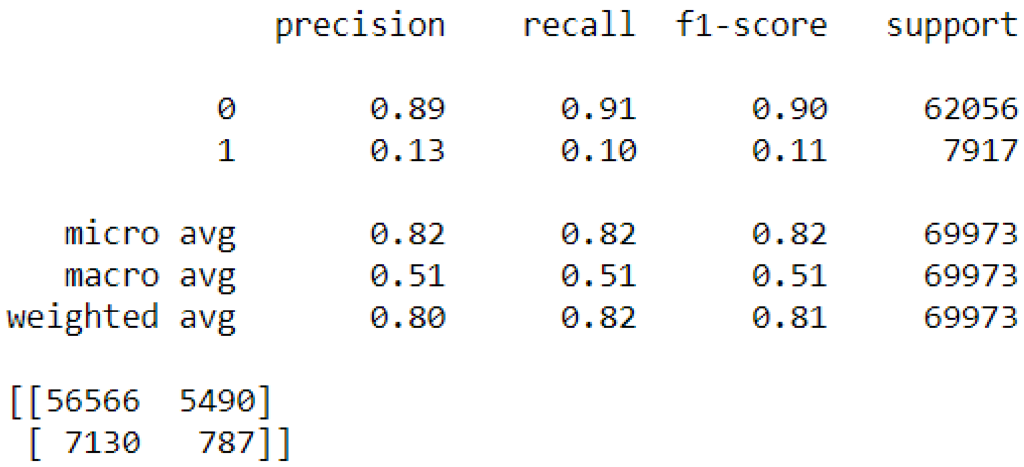
```
              precision    recall  f1-score   support

           0       1.00      0.91      0.95     69933
           1       0.00      0.05      0.00        40

   micro avg       0.91      0.91      0.91     69973
   macro avg       0.50      0.48      0.48     69973
weighted avg       1.00      0.91      0.95     69973

[[63658    38]
 [ 6275     2]]
```

  - **Classification Report for KNN (Base Model)**



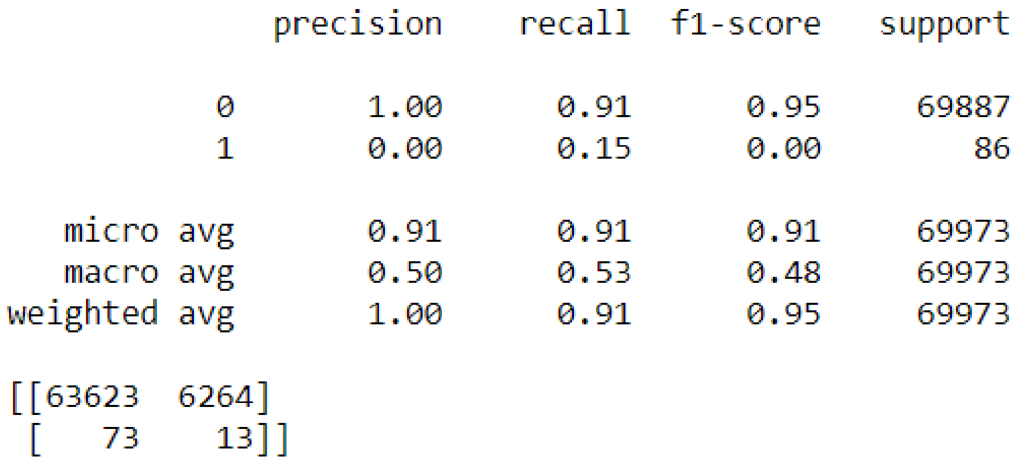
```
              precision    recall  f1-score   support

           0       1.00      0.91      0.95     69907
           1       0.00      0.17      0.00        66

   micro avg       0.91      0.91      0.91     69973
   macro avg       0.50      0.54      0.48     69973
weighted avg       1.00      0.91      0.95     69973

[[63641  6266]
 [   55    11]]
```

  - **Classification Report for Decision Tree Classifier (Base Model)**



```
              precision    recall  f1-score   support

           0       0.89      0.91      0.90     62056
           1       0.13      0.10      0.11      7917

   micro avg       0.82      0.82      0.82     69973
   macro avg       0.51      0.51      0.51     69973
weighted avg       0.80      0.82      0.81     69973

[[56566  5490]
 [ 7130   787]]
```

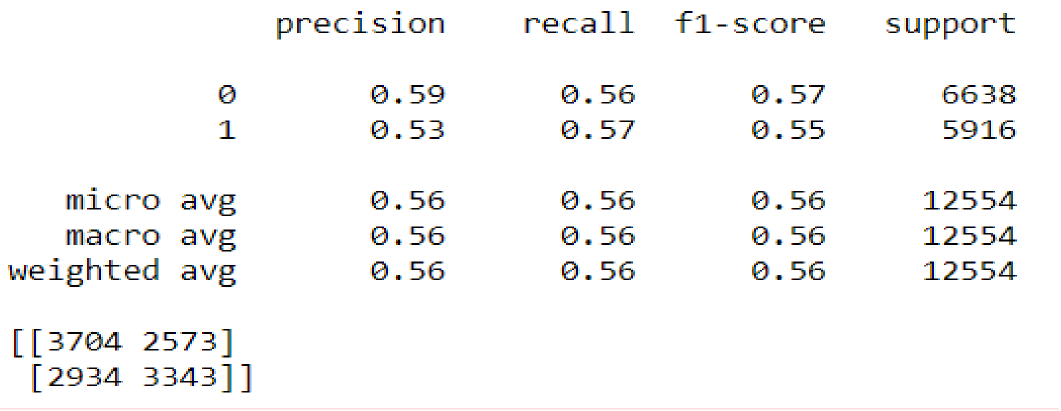
- **Classification Report for Random Forest Classifier (Base Model)**



```
              precision    recall  f1-score   support

           0       1.00      0.91      0.95     69887
           1       0.00      0.15      0.00        86

   micro avg       0.91      0.91      0.91     69973
   macro avg       0.50      0.53      0.48     69973
weighted avg       1.00      0.91      0.95     69973

[[63623  6264]
 [   73    13]]
```

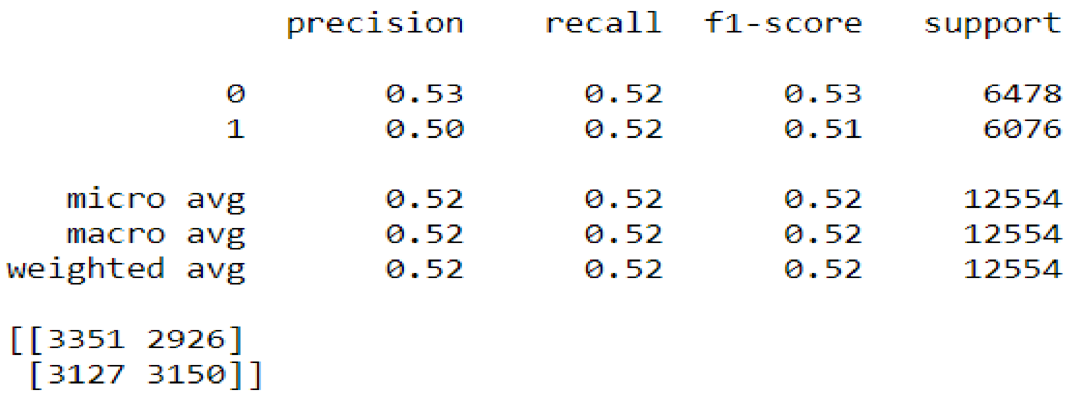
All General models are not performing well and totally misclassifying group1 because the data is highly imbalanced. We can improve the model by using resampling techniques.

- **Classification Report of Logistic Regression (With Under sampling)**



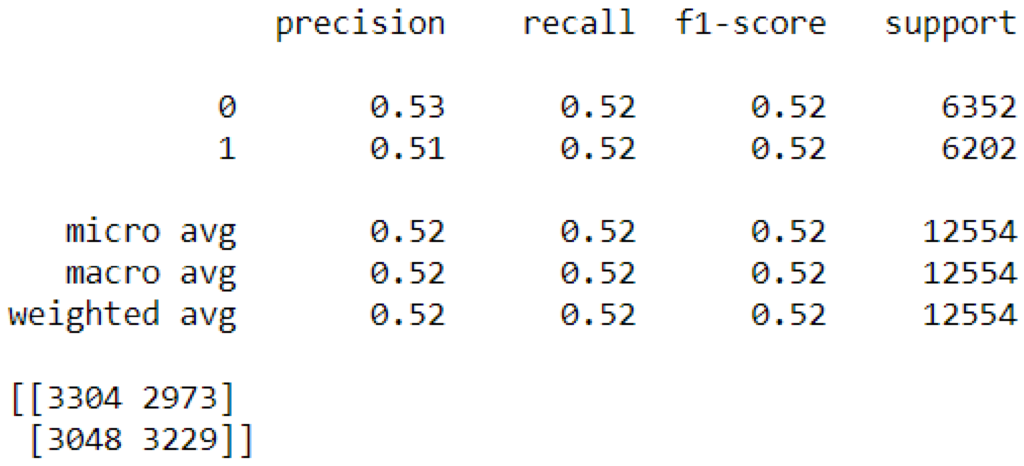
```
              precision    recall  f1-score   support

           0       0.59      0.56      0.57      6638
           1       0.53      0.57      0.55      5916

   micro avg       0.56      0.56      0.56     12554
   macro avg       0.56      0.56      0.56     12554
weighted avg       0.56      0.56      0.56     12554

[[3704 2573]
 [2934 3343]]
```

- **Classification Report of KNN (With Under Sampling)**



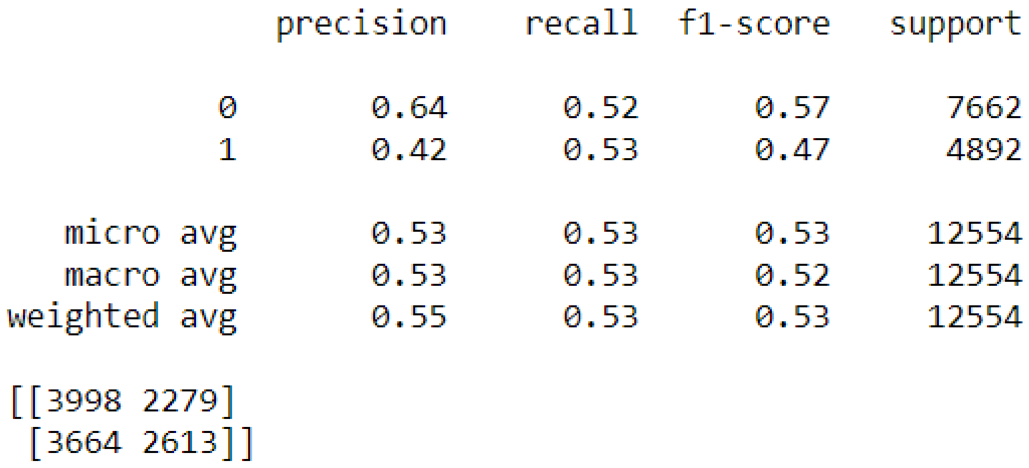
```
              precision    recall  f1-score   support

           0       0.53      0.52      0.53      6478
           1       0.50      0.52      0.51      6076

   micro avg       0.52      0.52      0.52     12554
   macro avg       0.52      0.52      0.52     12554
weighted avg       0.52      0.52      0.52     12554

[[3351 2926]
 [3127 3150]]
```

- **Classification Report for Decision Tree Classifier (With Under Sampling)**



```
              precision    recall  f1-score   support

           0       0.53      0.52      0.52      6352
           1       0.51      0.52      0.52      6202

   micro avg       0.52      0.52      0.52     12554
   macro avg       0.52      0.52      0.52     12554
weighted avg       0.52      0.52      0.52     12554

[[3304 2973]
 [3048 3229]]
```

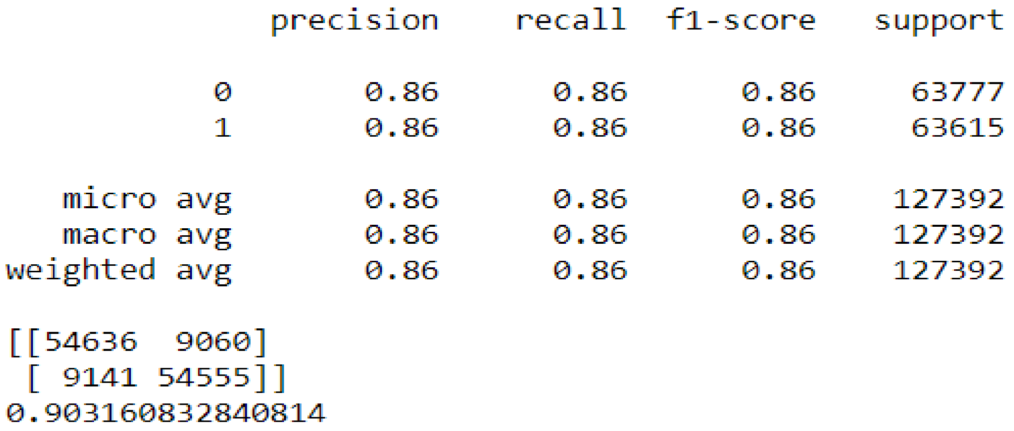
- **Classification Report for Random Forest Classifier (With Under sampling)**



```
              precision    recall  f1-score   support

           0       0.64      0.52      0.57      7662
           1       0.42      0.53      0.47      4892

   micro avg       0.53      0.53      0.53     12554
   macro avg       0.53      0.53      0.52     12554
weighted avg       0.55      0.53      0.53     12554

[[3998 2279]
 [3664 2613]]
```

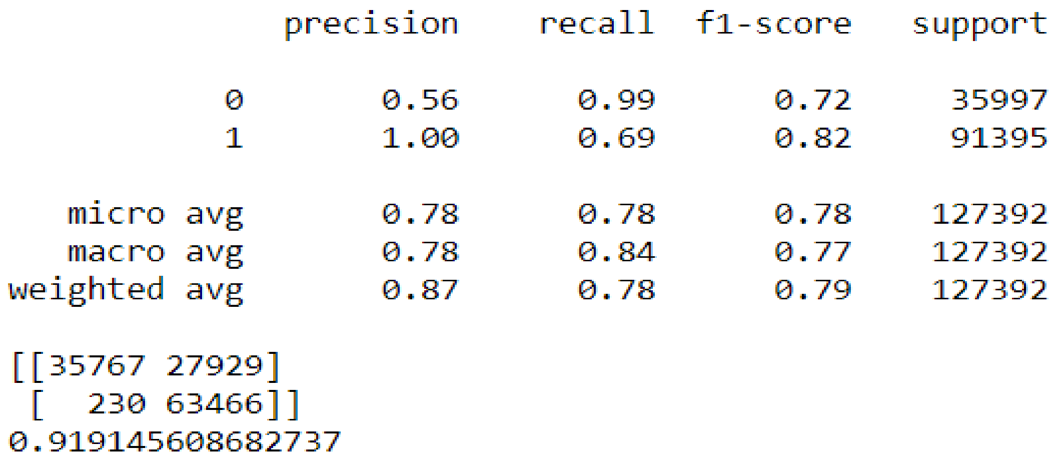
Although with under sampling the model is classifying the classes to some extent but still the result are not satisfactory so we can use Smote to improve our model.

- **Classification Report for Logistic Regression (With SMOTE)**



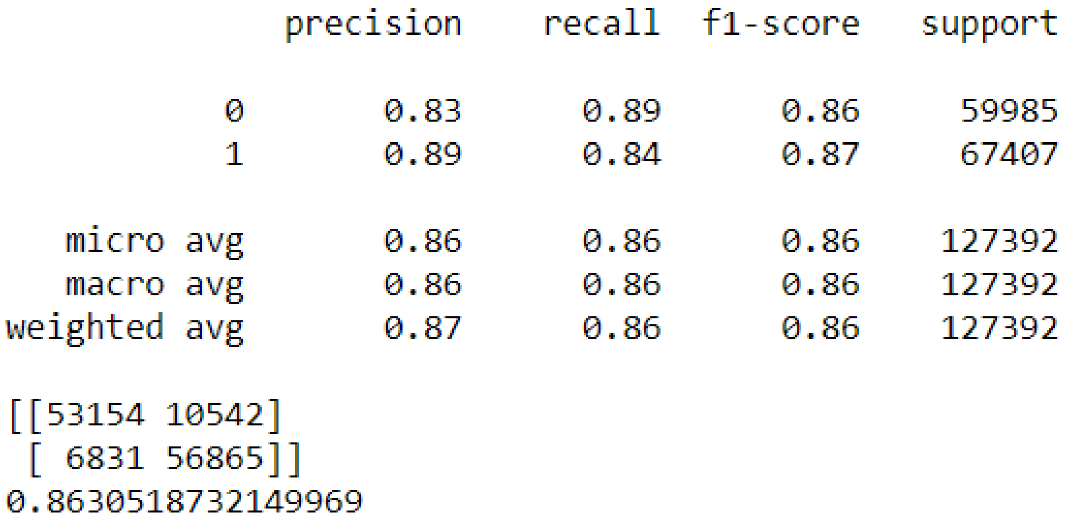
```
              precision    recall  f1-score   support

           0       0.86      0.86      0.86     63777
           1       0.86      0.86      0.86     63615

   micro avg       0.86      0.86      0.86    127392
   macro avg       0.86      0.86      0.86    127392
weighted avg       0.86      0.86      0.86    127392

[[54636  9060]
 [ 9141 54555]]
0.903160832840814
```

- **Classification Report for KNN (With SMOTE)**



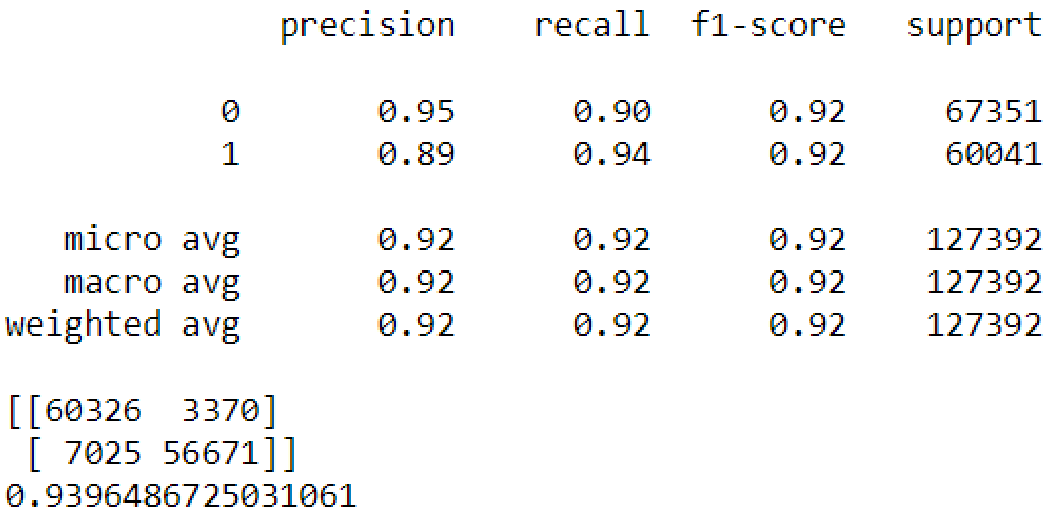
```
              precision    recall  f1-score   support

           0       0.56      0.99      0.72     35997
           1       1.00      0.69      0.82     91395

   micro avg       0.78      0.78      0.78    127392
   macro avg       0.78      0.84      0.77    127392
weighted avg       0.87      0.78      0.79    127392

[[35767 27929]
 [  230 63466]]
0.919145608682737
```

- **Classification Report for Decision Tree Classifier (With SMOTE)**



```
              precision    recall  f1-score   support

           0       0.83      0.89      0.86     59985
           1       0.89      0.84      0.87     67407

   micro avg       0.86      0.86      0.86    127392
   macro avg       0.86      0.86      0.86    127392
weighted avg       0.87      0.86      0.86    127392

[[53154 10542]
 [ 6831 56865]]
0.8630518732149969
```

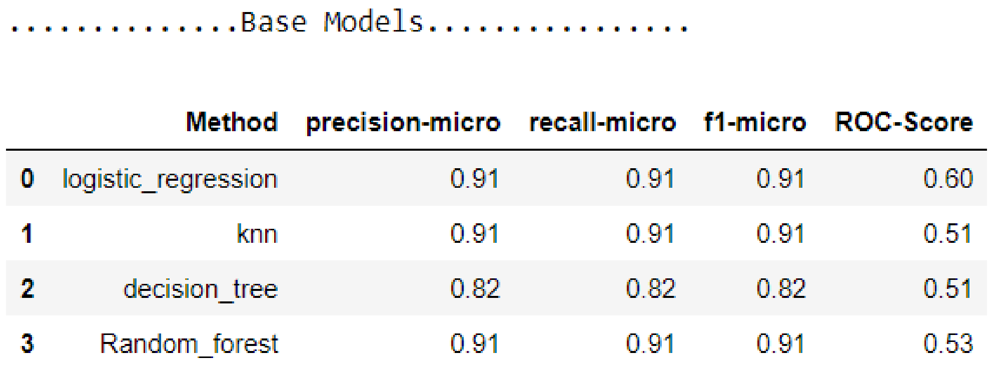
- **Classification Report for Random Forest Classifier (With SMOTE)**



```
              precision    recall  f1-score   support

           0       0.95      0.90      0.92     67351
           1       0.89      0.94      0.92     60041

   micro avg       0.92      0.92      0.92    127392
   macro avg       0.92      0.92      0.92    127392
weighted avg       0.92      0.92      0.92    127392

[[60326  3370]
 [ 7025 56671]]
0.9396486725031061
```

With SMOTE, Random Forest is giving best result with a F1 micro of 0.92 and Roc of 0.94 the model is performing best and classifying all the classes correctly.

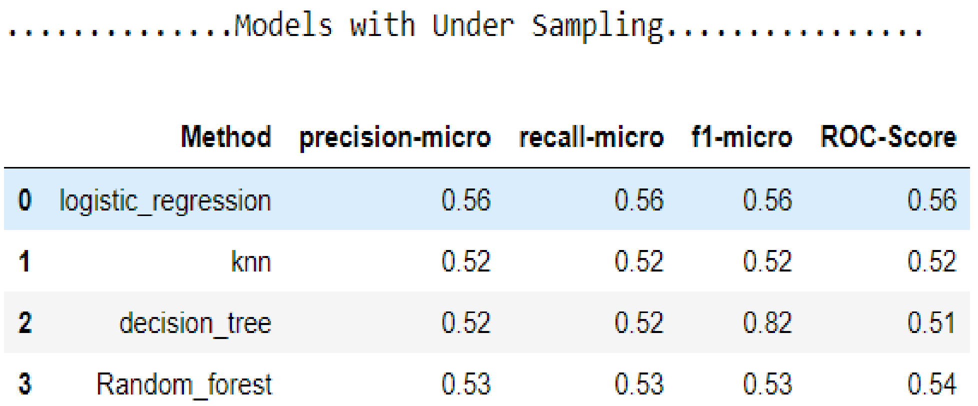
- **MODEL COMPARISON**
  - **Base Models**



```
..............Base Models................
```

	Method	precision-micro	recall-micro	f1-micro	ROC-Score
0	logistic_regression	0.91	0.91	0.91	0.60
1	knn	0.91	0.91	0.91	0.51
2	decision_tree	0.82	0.82	0.82	0.51
3	Random_forest	0.91	0.91	0.91	0.53

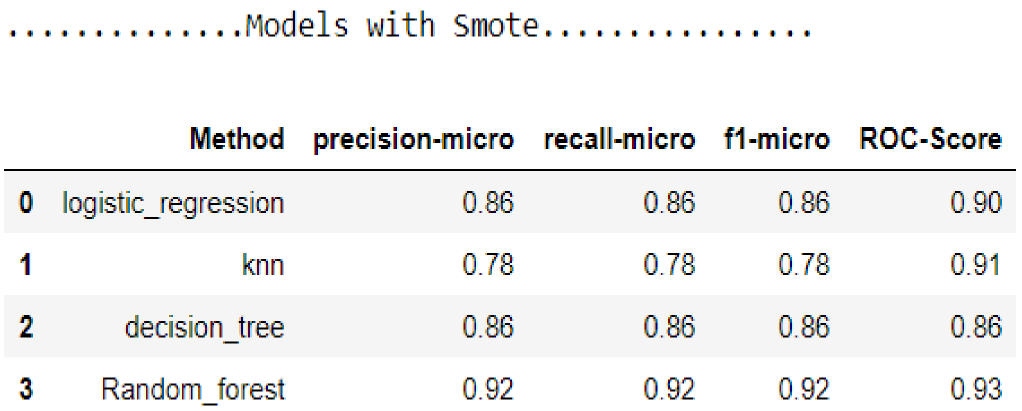
  - **Models with Under Sampling**



```
.............Models with Under Sampling...............
```

	Method	precision-micro	recall-micro	f1-micro	ROC-Score
0	logistic_regression	0.56	0.56	0.56	0.56
1	knn	0.52	0.52	0.52	0.52
2	decision_tree	0.52	0.52	0.82	0.51
3	Random_forest	0.53	0.53	0.53	0.54

  - **Models with SMOTE**



```
..............Models with Smote................
```

	Method	precision-micro	recall-micro	f1-micro	ROC-Score
0	logistic_regression	0.86	0.86	0.86	0.90
1	knn	0.78	0.78	0.78	0.91
2	decision_tree	0.86	0.86	0.86	0.86
3	Random_forest	0.92	0.92	0.92	0.93

With SMOTE the model is giving best result (roc_auc_score = 0.93) with random forest classifier. Because the data is highly imbalanced so base models without any resampling techniques are not giving good result. Even with undersampling the model is not showing satisfactory result because the data size decreases drastically to around 6000 from 63000(approx.).

One of the key reasons why SMOTE is giving better results than Over-sampling is that Over-sampling simply replicates the observations from the minority class, though this method does lead to no loss in information unlike Under-sampling, The disadvantage of using this method is that, The process of replication of observations in original data set, ends up adding multiple observations of several types, thus leading to over fitting. Although, the training accuracy of such data set will be high, but the accuracy on unseen data will be worse. Here's where the advantages of SMOTE come into play, Even though SMOTE is also an oversampling technique, it creates synthetic observations instead of reusing existing observations and hence the chances of over fitting get lowered down.

CONCLUSION

The treatment of a patient in a hospital plays a pivotal role in reducing the chances of readmission while simultaneously ensuring deliverance of quality health care at the first place, further abidance of medical prescriptions and instructions can also play a key role in reducing the risk of readmission. Conducting tests of Blood Glucose Levels, A1C test results at a wider level at regular intervals will also help in supervising the fluctuations in the records and maintenance of such records will lead avoidance of future emergencies and hence will help in avoiding readmission. The importance of Blood Glucose Test and A1C Test result has also been highlighted by model. An improvement in the dataset can be made by provision of features such as BMI(Body Mass Index) of the patient, Hereditary records of the patients as well as Blood Pressure values to enhance the accuracy and the explainability factor of the model further.

REFERENCES

[1] - Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records, *Hindawi Publishing Corporation BioMed Research International Volume 2014*, Article ID 781670, 11 pages http://dx.doi.org/10.1155/2014/781670

[2] - Predicting 30-Day Hospital Readmission for Diabetes Patients using Multilayer Perceptron, *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 2, 2019

[3] - G. E. Umpierrez, S. D. Isaacs, N. Bazargan, X. You, L.M.Thaler, and A. E. Kitabchi, “Hyperglycemia: an independent marker of in-hospital mortality in patients with undiagnosed diabetes,” *Journal of Clinical Endocrinology and Metabolism, vol. 87*, no. 3, pp. 978–982, 2002.