# Readmission Rate of Diabetic Patients

By

Ramya Sree V

# Problem Statement

- **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 hospital to save money also improve the healthcare services..**
- **With this project, the aim is to find the best model for readmission prediction and the factors which most likely affect the readmission.**

# Project Flow

Data Collection

Data Wrangling(Missing values, Data cleaning, standardization etc.)

EDA & Visualizations

Modelling:
Logistic
Decision Trees
Random Forest
AdaBoost Classification

# Data Description

**Medical records in the dataset include 50 attributes that are the risk factors,in addition to a label indicating the readmission status of a patient indicates whether a patient was readmitted to the hospital within 30 days or not.**

**The dataset encounters satisfy the following conditions:**

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

# Null Value Handling

race 2273
weight 98569
payer_code 40256
medical_specialty 49949

diag_1 21
diag_2 358
diag_3 1423

- The features having more than 85% null values are been dropped.
- As diagnostic tests (diag_1, 2,3) and medical_specialty are either continuous or discrete i have replace null values with “unknown’’ as the null values are below 10%.

# Feature Engineering

**In Healthcare sector service utilization factor plays a key role to know how efficient the hospital is providing services.**

```
data['service_utilization'] = data['number_outpatient'] + data['number_emergency'] + data['number_inpatient']
```

**Understanding Diagnostic test Codes:**

**According to ICD-9-CM- Diagnostic codes are defined with specific terms within interval of numbers. Ref for diagnostic code: https://www.findacode.com/code-set.php?set=ICD9**

```
def map_now():
    listname = [('infections', 139),
                ('neoplasms', (239 - 139)),
                ('endocrine', (279 - 239)),
                ('blood', (289 - 279)),
                ('mental', (319 - 289)),
                ('nervous', (359 - 319)),
                ('sense', (389 - 359)),
                ('circulatory', (459-389)),
                ('respiratory', (519-459)),
                ('digestive', (579 - 519)),
                ('genitourinary', (629 - 579)),
                ('pregnancy', (679 - 629)),
                ('skin', (709 - 679)),
                ('musculoskeletal', (739 - 709)),
                ('congenital', (759 - 739)),
                ('perinatal', (779 - 759)),
                ('ill-defined', (799 - 779)),
                ('injury', (999 - 799))]
```

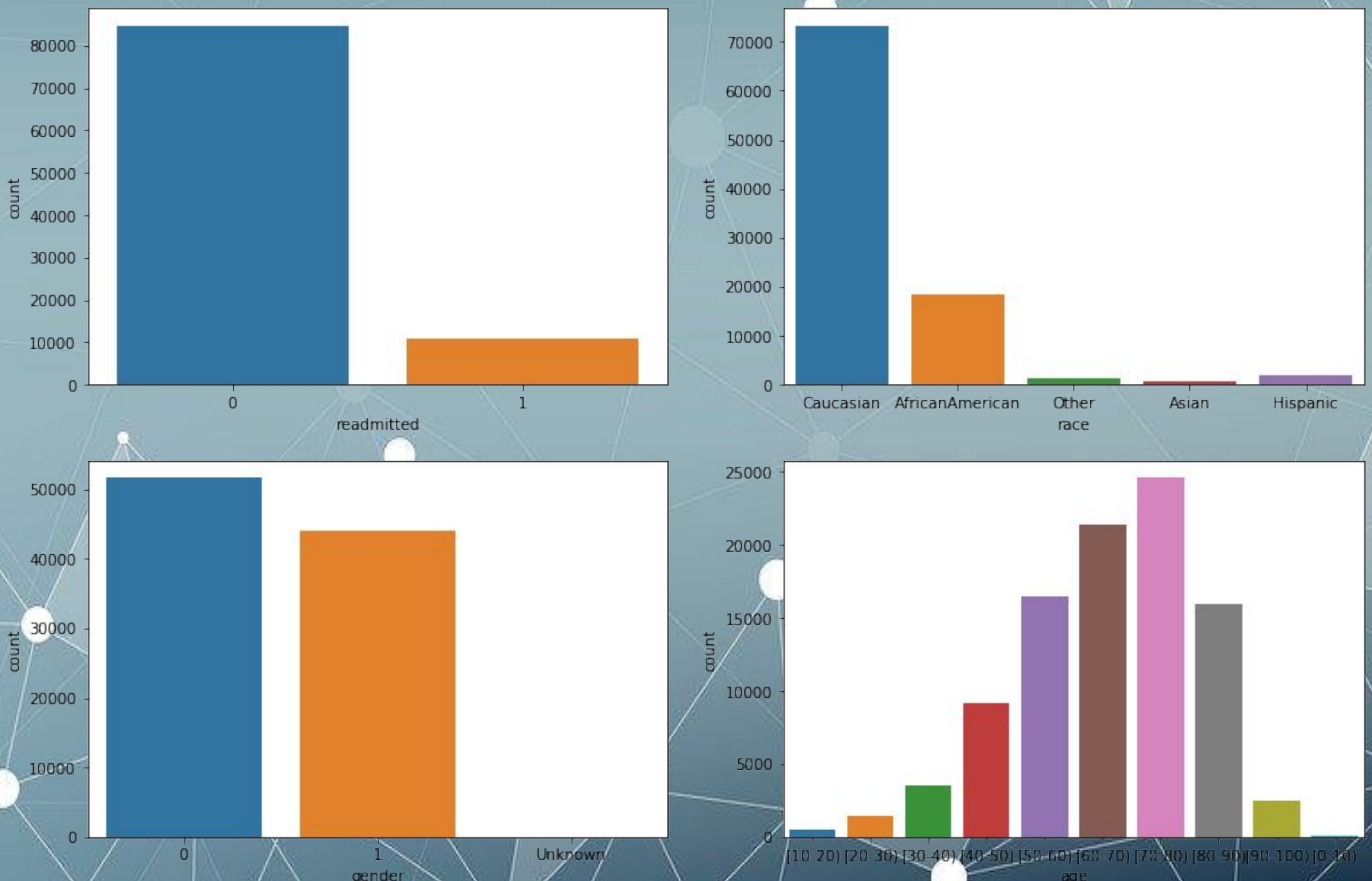
# Exploratory Data Analysis

- **Analyzing the numerical and categorical data separately.**

- **This picture discuss about numerical data present in the dataset**

# EDA…

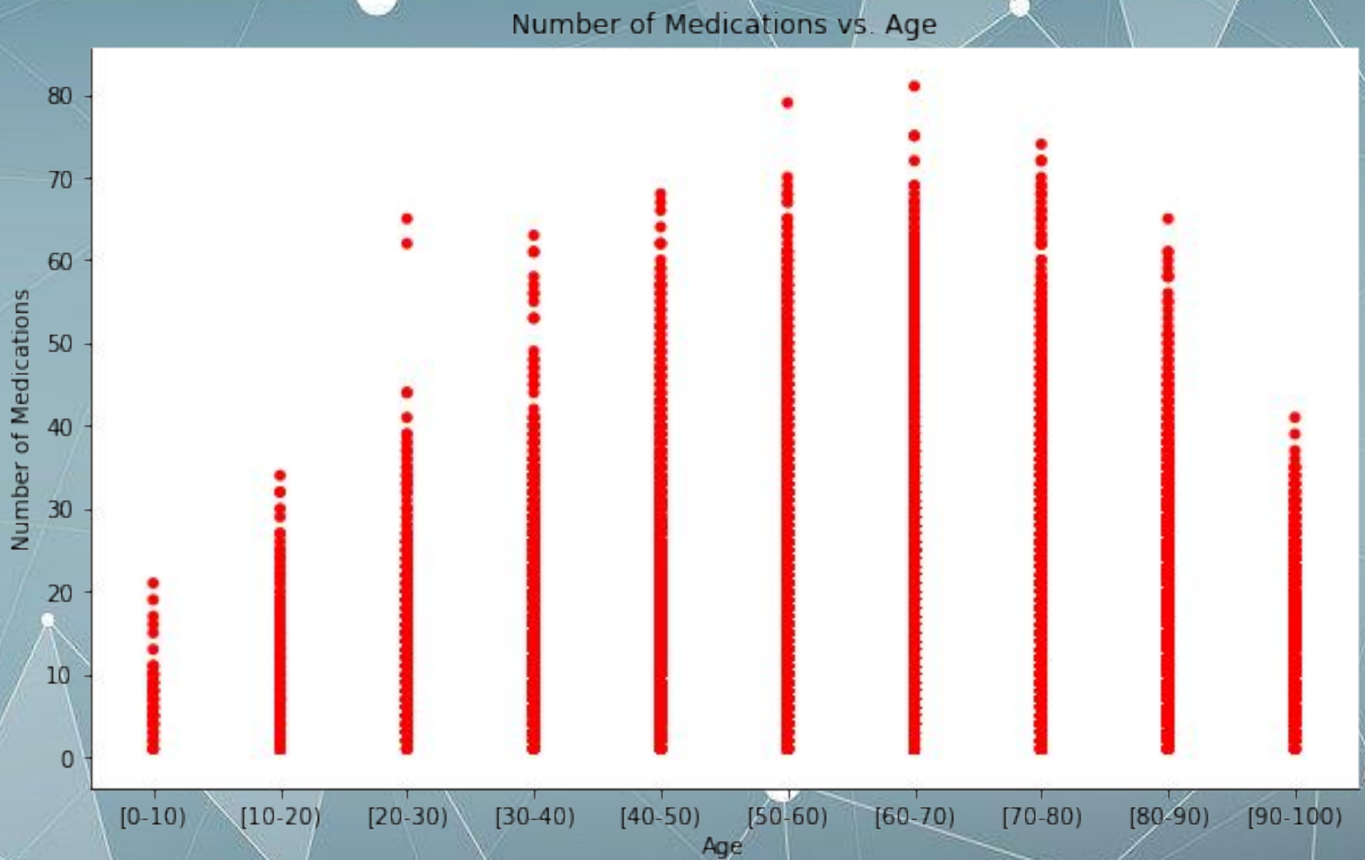
**This picture discuss about categorical data like Readmitted patients, Race, Gender and Age present in the dataset.**



# EDA

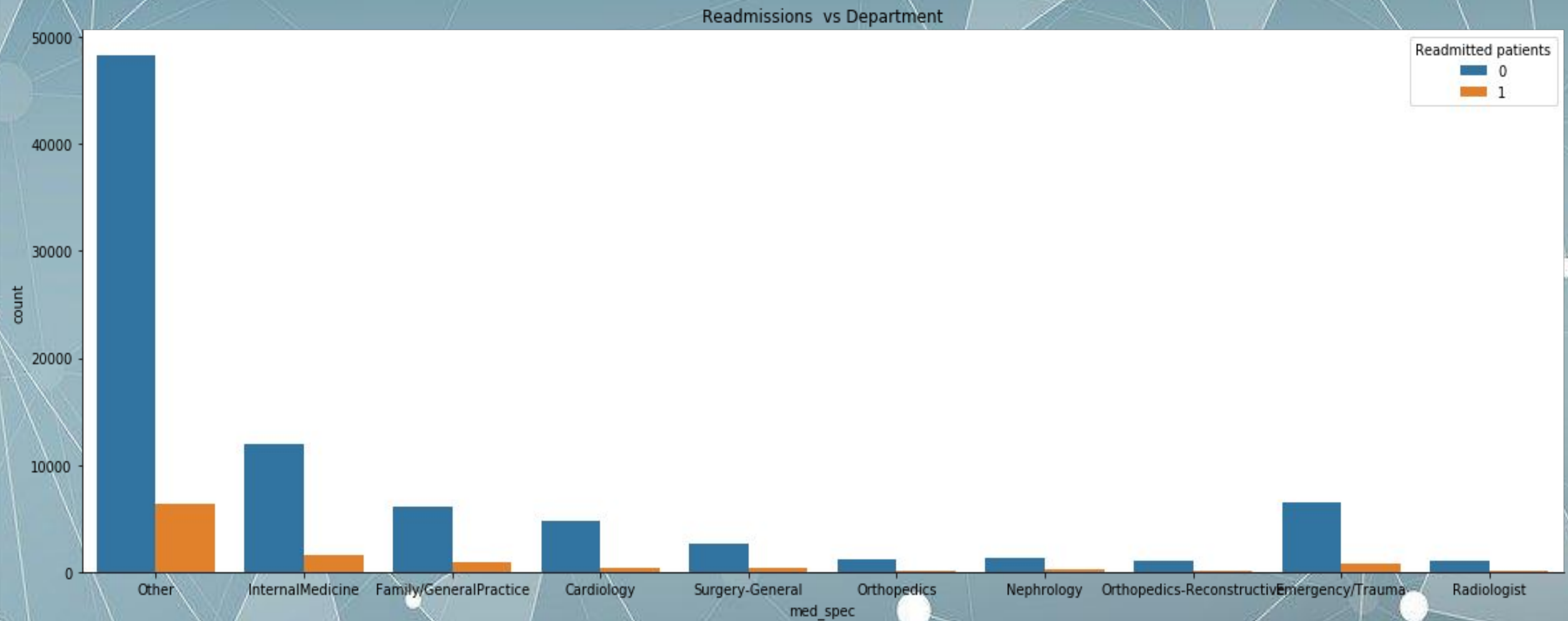
**This picture discuss about relationship between number of medications provided to the patients who admitted in different departments.**

# No.of medications vs age




When the doctor gives medicine to patient he considers the patients age according to age he prescribes the power of medicine & no.of medicines.

# No. of Readmission per Department




**The above chart gives the how many patients patients are readmitted per department**

# Data Journey

1. Importing All Libraries.
2. Loading dataset amd reading it.

**Data Profiling**

1. Null handling
2. Check uniques for every feature.
3. Drop null values
4. Drop highly correlated values

**Feature Engineering**

1. Encode features
2. Check for categ -orical features.
3. Check for Numerical features

**EDA**

1. Plot categorical & numerical data
2. Analyze medicine vs age.
3. Check service of the hospital

**Feature Analysis**

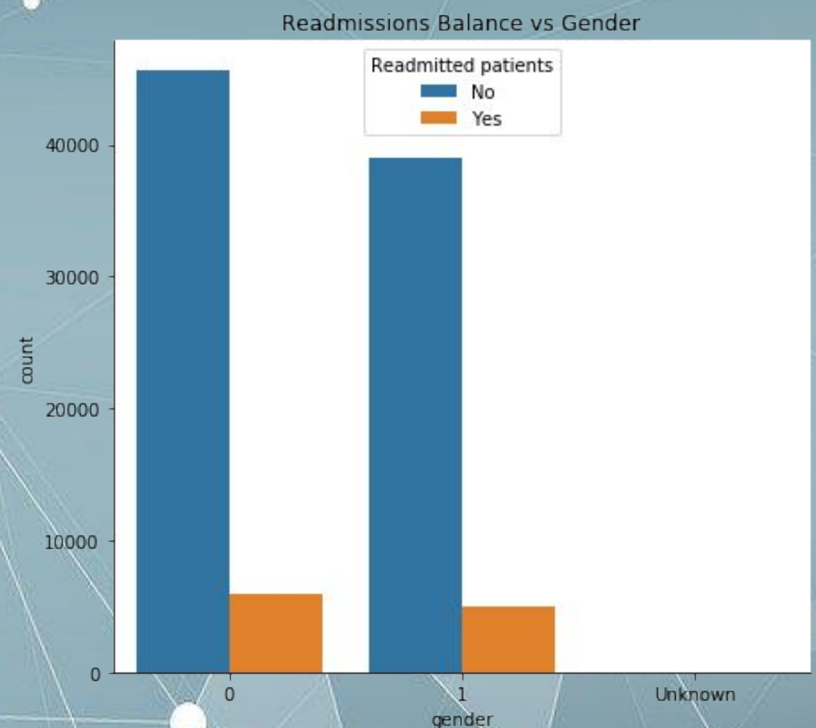
1. Confirm target.F.
2. Analyze corr between Target.F & other features
3. Remove skewne- ss & kurtosis.

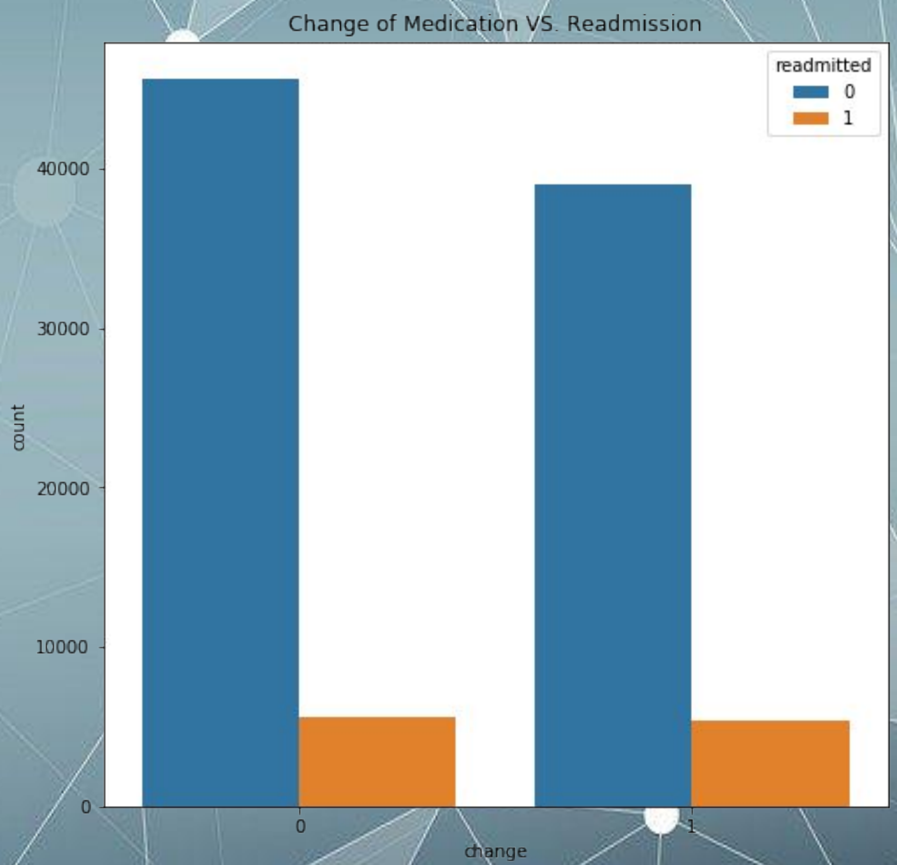
**Modelling**

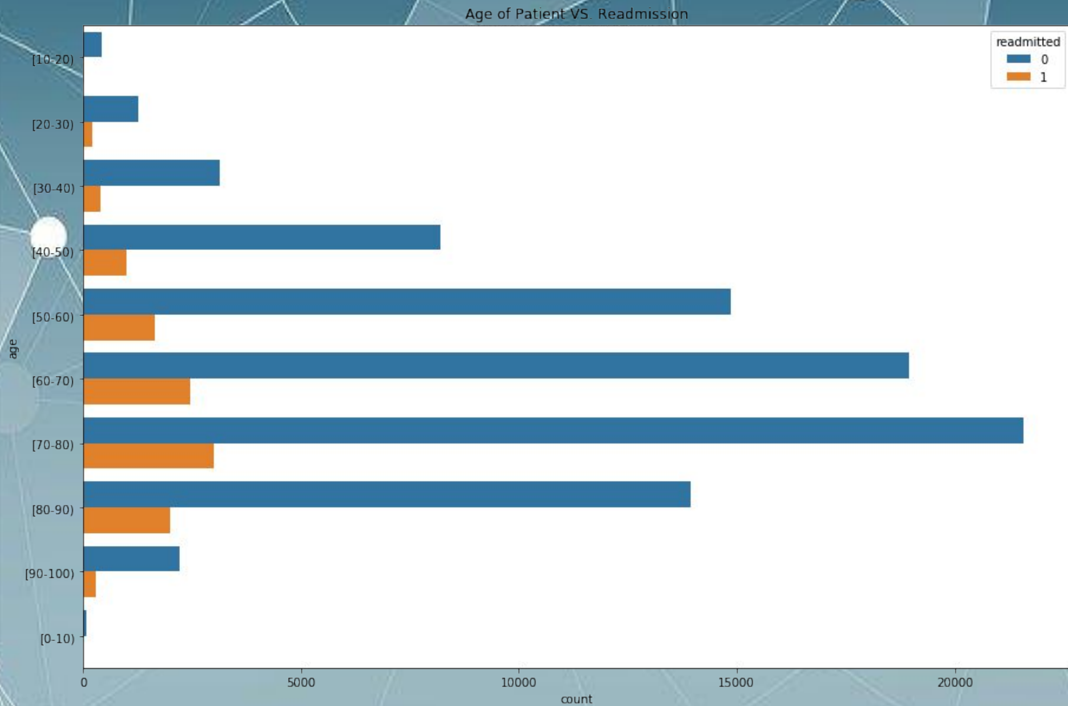
1. Modelling with different classification techniques.
2. Comparing models.

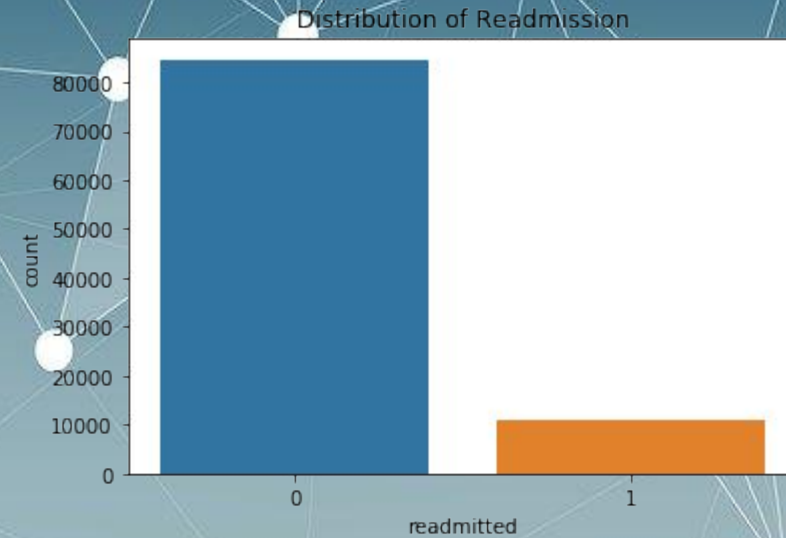
# Feature Analysis

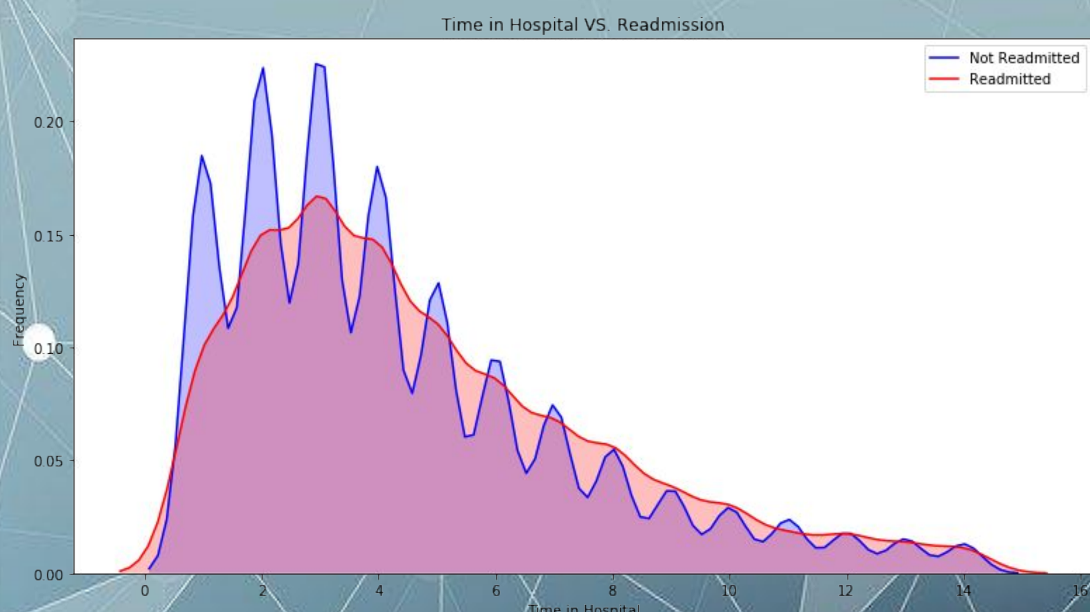
## Checking behaviour of each feature with  target variable

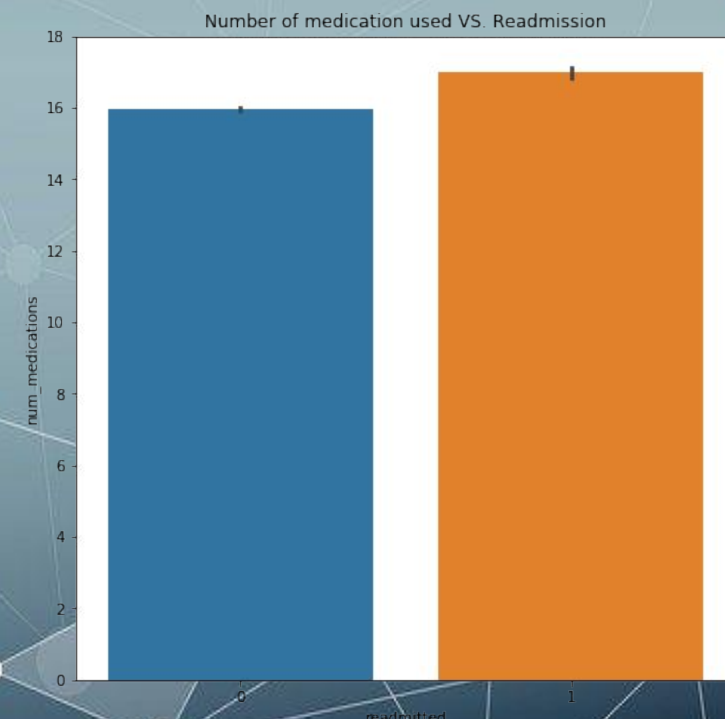



Age of Patient VS. Readmission
readmitted
0
1
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
[0-10)
age
0
5000
10000
15000
20000
count


Distribution of Readmission
80000
70000
60000
50000
40000
30000
20000
10000
0
count
0
1
readmitted


Time in Hospital VS. Readmission
Not Readmitted
Readmitted
0.20
0.15
0.10
0.05
0.00
Frequency
0
2
4
6
8
10
12
14
16
Time in Hospital


Number of medication used VS. Readmission
18
16
14
12
10
8
6
4
2
0
num_medications
0
1
readmitted

# Modelling

**The Models i have chosen are:**

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **Adaboost Classification**
- **Naive Bayes**
- **Gradient Boost Algorithm**

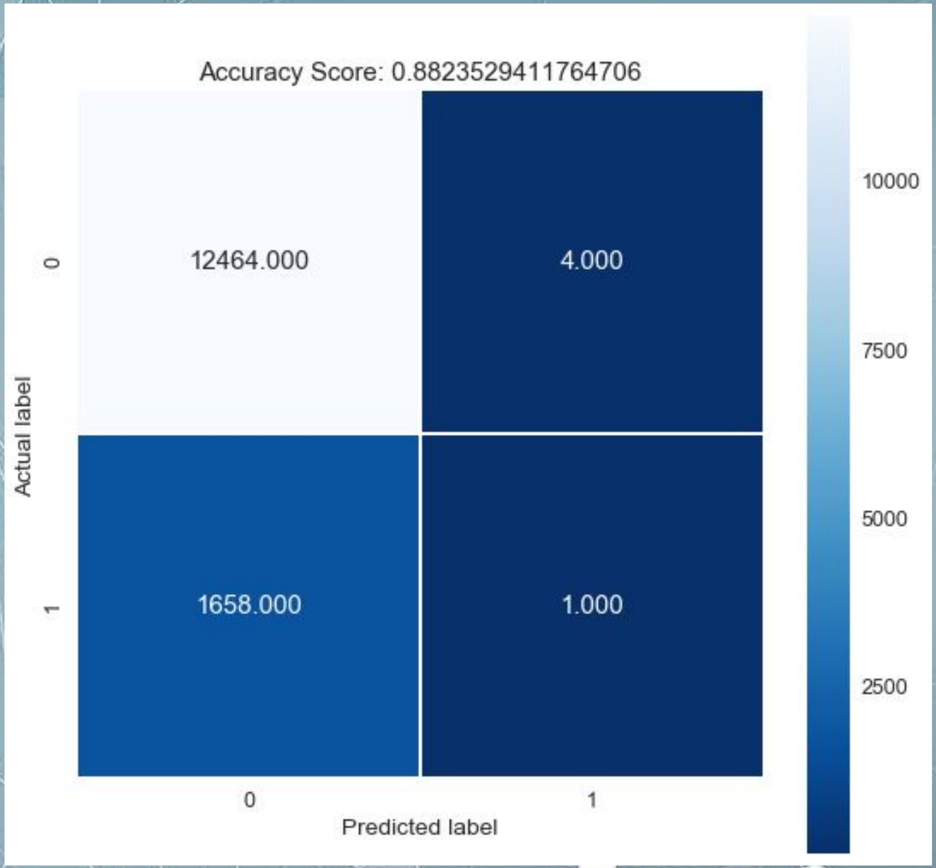
**As per problem statement (prediction of readmission of diabetic patients) i have selected  the above models**

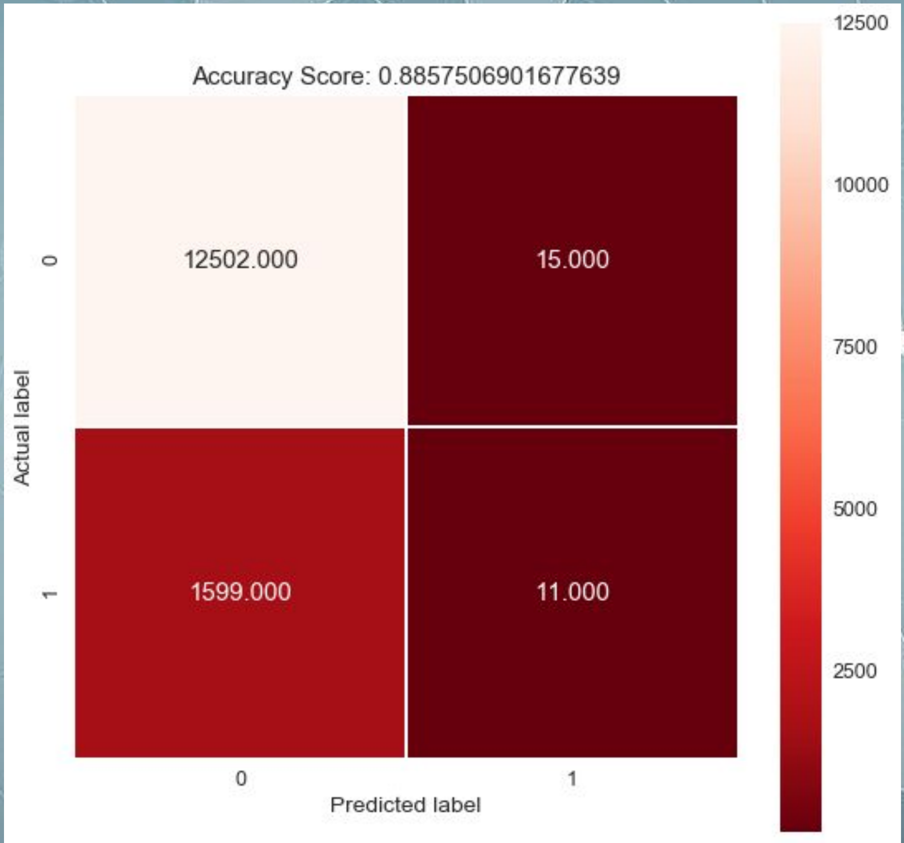
# Evaluation

**To evaluate the model performance the following parameters are been used:**

1. **Accuracy**
2. **Precision**
3. **Recall**
4. **Confusion Matrix**

# Evaluations of Models

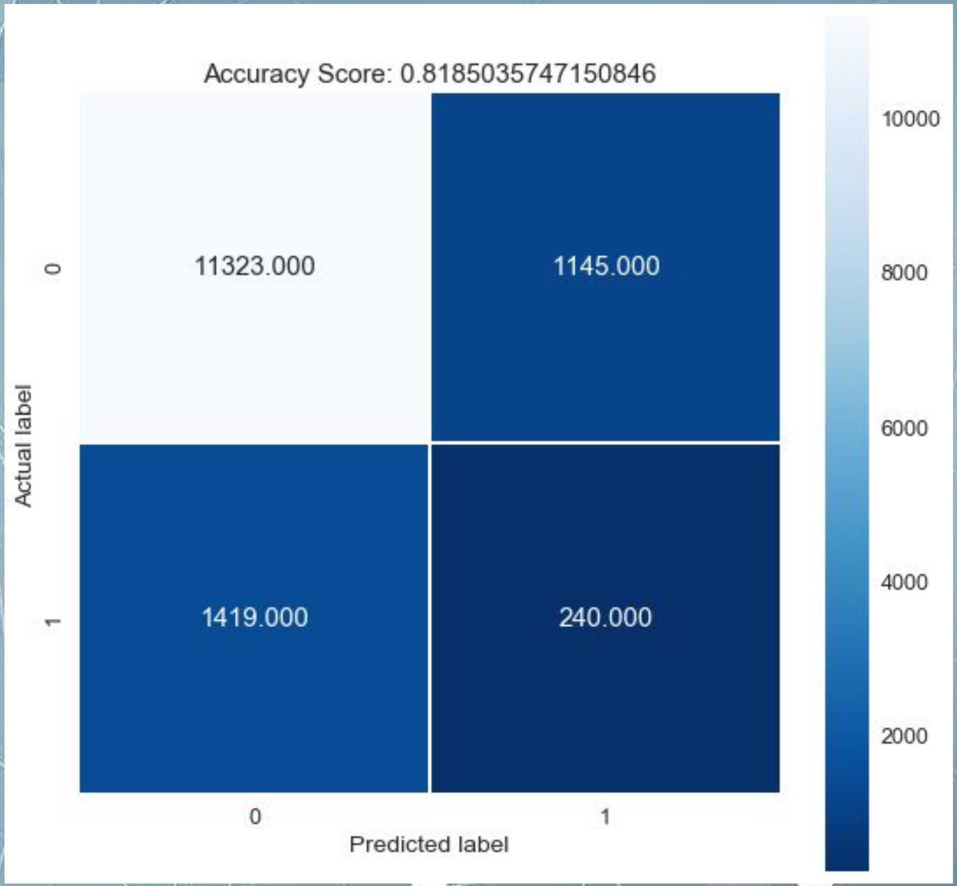


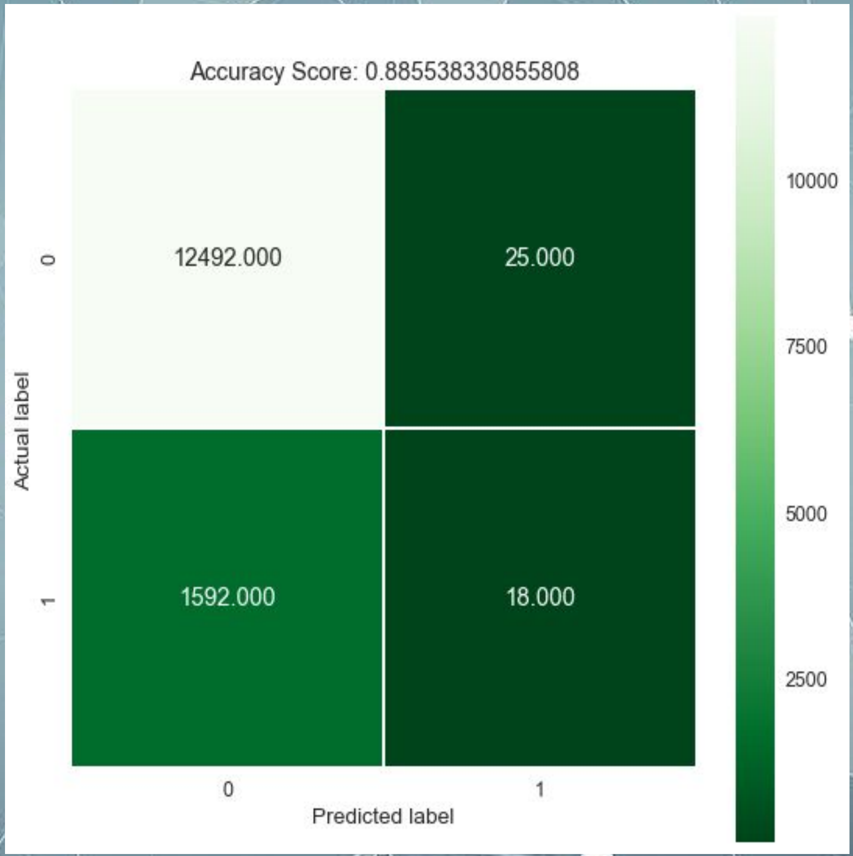

**Confusion Matrix of Logistic Regression**

**Confusion Matrix of Random Forest**

# Evaluation of Models

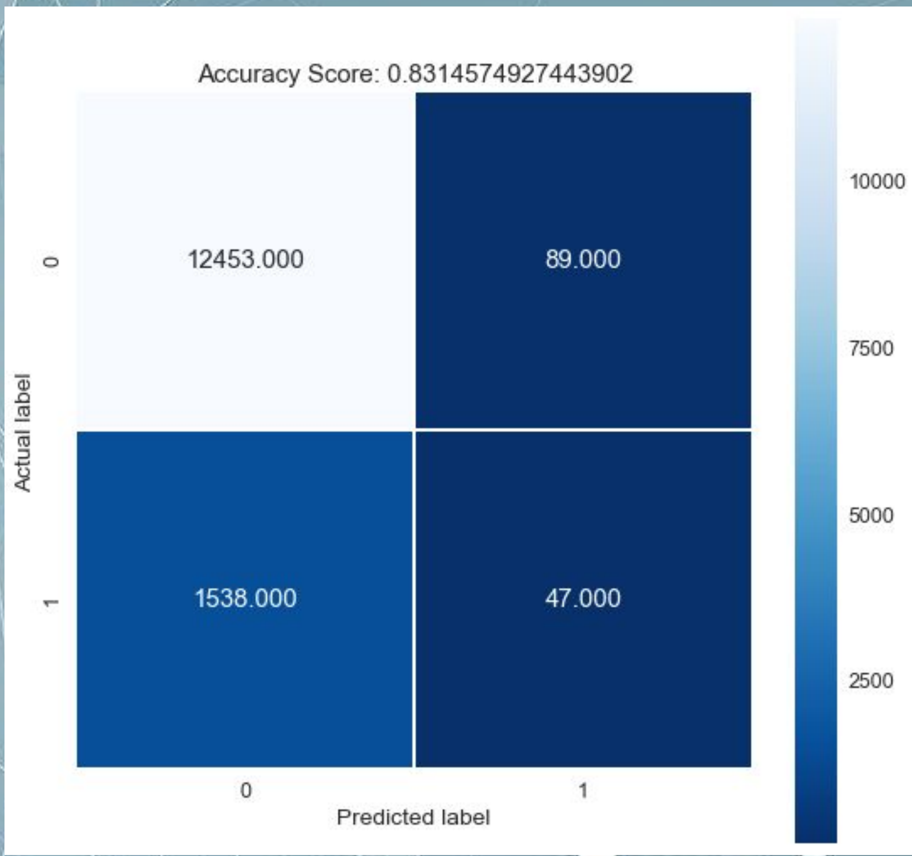



Confusion Matrix of Decision Tree

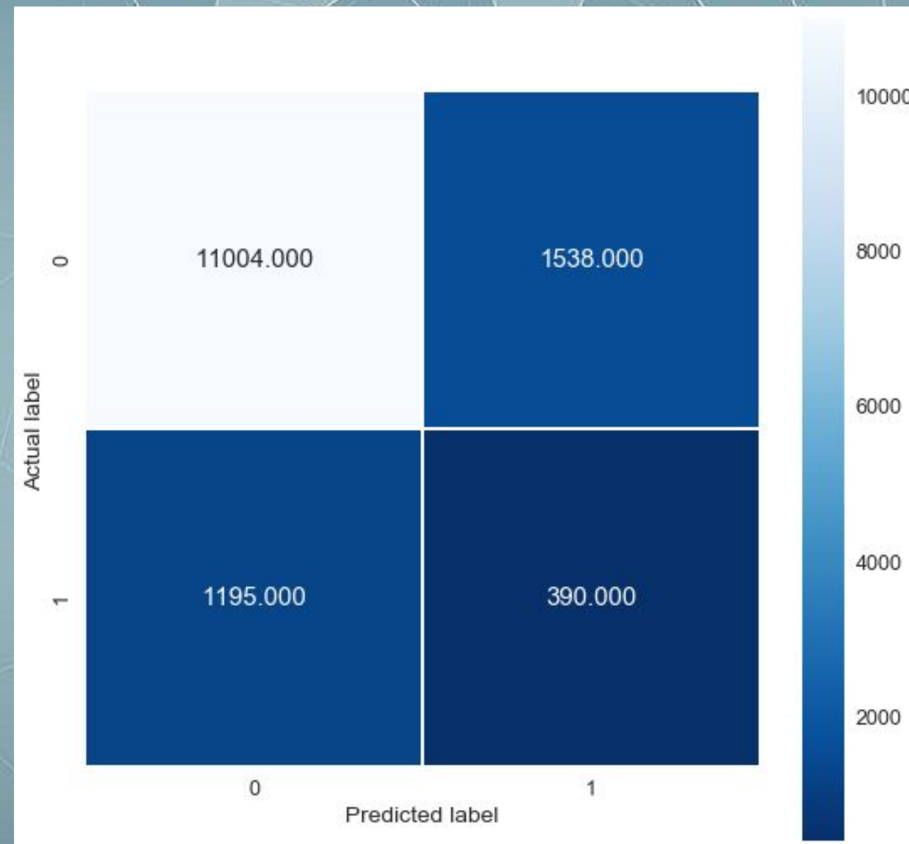



Confusion Matrix of AdaBoost Classification Model

# Evaluation of Models

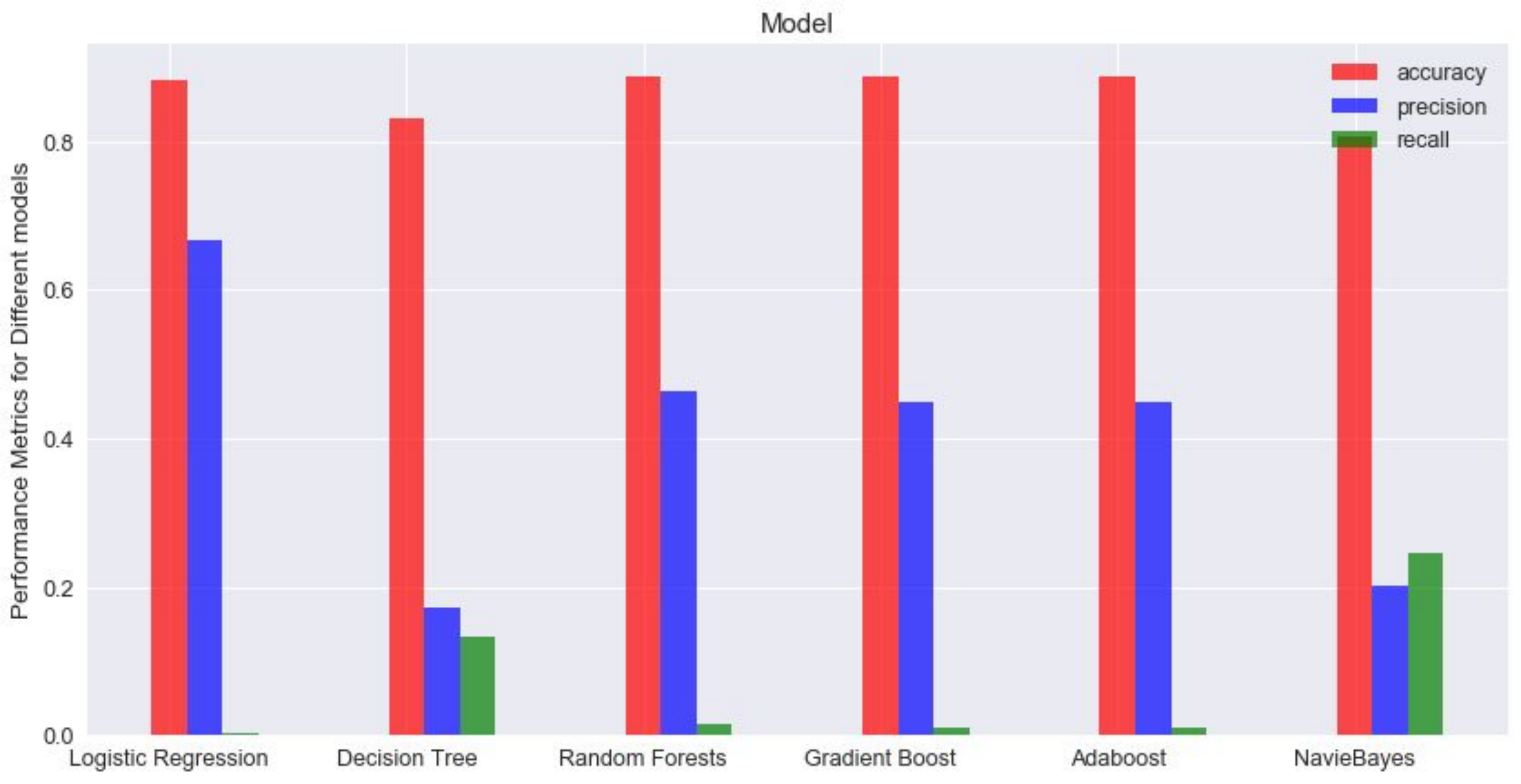



**Confusion Matrix of Gradient Boost Algorithm**




**Confusion Matrix of Naive Bayes Model**

# Comparison of Models



# Business Recommendations

- **After processing models we ended having Random Forest is the good model for classification.**
- **We ended with some of the important features which hospitals have to consider. Therefore hospital authorities know  which feature is showing more effect on readmission rate.**