Akshatha Kamath 11409022
INFO 5082 Fall 2021

PREDICTION OF HOSPITAL READMISSION RATE OF DIABETIC PATIENTS

INTRODUCTION AND BACKGROUND OF THE PROJECT

Diabetes is a disease that is prevalent very prominently in today's world. It is a condition that occurs when an individual's blood sugar level is higher than normal. Insulin, which is a hormone secreted in the pancreas, helps in regulating blood glucose levels. Diabetes has no cure, which is why it is important to take care of being diagnosed with diabetes. Millions of people are admitted to hospitals in the early stages of diagnosis and when the effects become worse. Getting admitted once should guide the patient towards a steady recovery of health. If the patient is required to get readmitted into a hospital, it can question the hospital's capability to help people recover from diseases. This research is carried out with a motive to determine the trend in which diabetes affects, which group of people is affected, and what are the reasons if any for readmission and assess the right type of diabetes medications. The progress so far is the collection of the data set, data cleaning, identifying the different attributes of the dataset, shortlisting Machine Learning algorithms that could be implemented in the further stages of the research, and the literature review of the related works.

In the world of data science, if there is any domain that involves a variety of data that is concentrated towards a cause or a result, it is the field of health care. There is no exact set of features that can be considered to determine a certain condition or disease. Therefore, this topic was selected, to make use of the features to carry out an analysis of data, make use of models provided by technologies like Machine Learning to draw inferences and suggest a better approach and treatment of diabetes and control the readmission rate. Through this research, we will be able to apply the methods of data analysis and use them to make decisions and suggest solutions towards an existing problem. Being able to apply data science into a real-world scenario and bring results will complete my initial pursuit towards learning data science. Medical data is very important, yet sensitive data which needs thorough studying and analysis to obtain the right results. Not only does this focus on an individual’s health, but also the health care industry. Mainly because the analysis is being carried out on a specified set of data. This is a small step towards analysis of medical data, and there are many scientists with experience who have carried out detail research. I wanted to study more applications of machine learning and data analysis and looking forward to applying advance technologies like deep learning and neural networks. Below there are a few related

works that I have reviewed that have detailed research on diabetes related data. The ideal goal of this research is to be able to implement it with live, real-time data, which is currently out of scope for academic-level research. To conclude, this research is to understand the gender and age group affected, predict its frequency, and predict the probability of patient readmission rate with the help of data science and visualization concepts.

STATEMENT OF THE PROJECT PROBLEM

This research has been carried out to predict the rate of readmission to hospitals of diabetic patients. Hospital readmission is one of the major concerns in The United States of America due to health care being very expensive. The readmission rate of the patients is an indicator for hospitals to improve or work on their quality of services. A higher readmission rate will lead to penalties for hospitals and might affect the cost of care. This research, using the dataset with relevant features, was taken up to figure out which patients have a higher chance of readmission and what factors affect the readmission rate with the help of data analysis, which starts right from the collection of data to using machine learning models and visualizing the results.

RELATED WORK

1. **Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records**

Purpose of Study

The purpose of this study was to study and analyze the way the measurement of hemoglobin (HbA1c) can make an impact on the rate at which patients are admitted or readmitted to the hospital. For this study, a database of 70.000 patient records was made use of. Since the research that is being carried out is related to the readmission rate and its factors, this paper was selected to study the reasons for readmission.

Related Work

Readmission rate and its impact on the national economy and an individual's life is crucial. To publish this paper, the authors made references to 18 papers. On publishing this paper, it has been cited 269 times.

Research Design

To carry out this research, a few assessments of diabetes care in hospitalized patients were considered to serve as a baseline. The analysis of the database was carried out to take a closer look at the pattern of diabetes care of patients admitted to a hospital and address the steps to be taken in the future to improve the condition of the patient. The use of hemoglobin as a marker for diabetes care was examined. The authors hypothesized that the reduction of the rate of readmission of patients to hospitals can be associated with the measurement of hemoglobin in the patient. The authors defined the readmission attribute as "readmitted" if the patient is admitted within 30 days of discharge, and "otherwise" which includes readmission post 30 days or no readmission. The patient records were examined to determine the first inpatient visit after discharge. The authors examined the frequency of hemoglobin test ordering and the response to the result which was mapped as the change in medication. The authors considered four groups of encounters: no test performed, test performed at a normal range, result with more than 8% and no medication change, and result with more than 8% and medication change. The authors analyzed a single encounter per patient to keep it independent. The significance level was determined at $p<0.01$. They used multivariate logistic regression to fit HbA1c measurements against readmission. Each step had tests of significance of the higher degree of freedom variables and the deviance table was analyzed. This was followed by a sensitivity analysis. A logistic model was fitted without the HbA1c and called the core model. Then HbA1c is added. Then pairwise interactions are added without HbA1c, and the significant ones were kept. Finally, the pairwise interactions are added with HbA1c.

Conclusion

After carrying out the analysis, the authors posted the statistical results of the analysis showing the percentages of the number of times HbA1c was measured throughout the treatment and the measures taken after the measurement, considering the various factors that were taken from the dataset. The results were plotted and represented graphically. The authors concluded that the measurement HbA1c for diabetic patients turned out to be a very useful factor to predict the rate of readmission. It provided insights to develop strategies to reduce readmission rates, reducing the costs incurred. The reason for readmission differed significantly between patients that had their samples collected during their diagnosis of diabetes and patients that had their samples collected for other causes, which had a higher rate. Readmission of diabetic patients was associated with the decision to test for hemoglobin.

2. Factors Affecting Costs and Utilization of Type 2 Diabetes Healthcare: A Cross-Sectional Survey among 15 Hospitals in Urban China.

Purpose of Study

The authors of this paper have considered several hospitals across China and have analyzed the possible factors that can affect the costs and the utilization of resources for the treatment of diabetes. Since hospital readmission rate is the main factor that can directly affect the nation's economy and the health care systems, this paper was included as a related work. In this paper, the authors have considered two main factors which are considered as continuous variables, Direct Medical Costs and Out of Pocket costs. Based on a few factors, the authors have brought about research as to how the nation's economy and people are affected.

Related Work

The discussions and analyses that have been carried out by the authors involve directly the national economy and the costs that are involved in the treatment of diabetes and the effect of the treatment on the economy of the health care systems. This paper has been cited 21 times and accessed 7126 times. 34 similar works have been referred by the authors to publish this research paper.

Research Design

Type 2 Diabetes is something that can affect people of all ages, and when this does happen, it becomes a burden on the nation's economy and the health care systems. This paper concentrates on the investigation of the determinants of the medical costs and the proportion of the cost that must be paid by the patient. The direct healthcare costs of diabetes range from 2.5% to 15% of the nation's annual health care budgets based on the treatments available. The expenses incurred by a diabetic person are two times greater than an average human being in China. The levels of hemoglobin and insulin treatments were associated with the medical costs. Different methods have been incorporated to investigate the determinants. The study was set across four major cities, Shanghai, Beijing, Guangzhou, and Chengdu, spanning across all ends of the country. The subjects who were eligible for this study were adult outpatients that received treatment for Type 2 Diabetes. Patients were interviewed in the hospitals with a survey, that had questions related to the patient's demographics, the characteristics of diabetes, and the complications caused along with the history of its treatment. The outcomes of the survey were that the total expenditures of treating Type 2 Diabetes. The out-of-pocket proportion was a secondary outcome.

The survey had questions related to an individual's marital status, education level, employment, monthly income, and if there was any history in the treatment of diabetes.

Direct medical costs (DMC) and out-of-pocket (OOP) were measured as continuous variables. Generalized estimating equations (GEE) models were used to determine the factors that are associated with these costs with a Gamma distribution and a log-link function. A GEE model with a Poisson distribution and log-linear function was performed. There were a total of 1530 subjects identified as Type 2 Diabetes patients. With the increase in the complications that were caused because of Diabetes, the DMC costs went on increasing. 41.7% of them were male and 58.7% were in tier 3 hospitals. Among them, 91% were taking medications and 33% were receiving insulin therapy. There was an observation that male patients paid a higher DMC as compared to female patients. Employed outpatients paid 12% lesser annual DMC than the unemployed outpatients. As the complications of type 2 diabetes increased, there was a rise in the annual DMC by 33%. China has a health insurance system that is providing good coverage for the people that are employed but not quite the case for those employed in rural areas. This brought about inequalities in the health care provisions.

Conclusion

Type 2 Diabetes-related DMC varied concerning the characteristics of the disease, whereas the out of pockets proportion was mainly determined with an individual's socioeconomic status and his level of health care affordability. Factors like employment and place of living also played a major role. The study concluded that overall, DMC can be reduced if the appropriate, effective treatments are provided in the diagnosis of diabetes. Preventing the occurrence of complications can reduce health care costs, and this is only possible with the help of early detections and treatment. Health consequences can be delayed, reducing the costs. To make this possible, norms of healthcare must be reformed. This also includes the reformation of the nation's medical insurance system and efficient distribution of patients in hospitals so that everyone can be a recipient of effective and quality treatment.

3. Factors Affecting Health-Care Costs and Hospitalizations Among Diabetic Patients in Thai Public Hospitals

Purpose of Study

In this paper, the authors have studied the attributes that are affecting the hospitals with the healthcare cost and the admissions of diabetic patients in Thai Public hospitals. Diabetes is a very common but serious chronic disease that stays forever and the complications due is this is life-long. All over the world, every age group includes people that have diabetes. This disease is affecting patients due to healthcare costs. In Thailand, an estimation of cost for diabetes was conducted in 7 Thai government hospitals in 4 regions of Thailand and Bangkok, the cost per patient for diabetes was approximately 6017 baht, the annual average total cost was 13,751 baht. The purpose of this study was to see the total healthcare costs and the admission of patients.

Related Work

The authors thought for this study was by having some information about these attributes might help improve patient management and maybe reduce the costs in the future. This paper has been cited 7 times and 18 similar works have been referred by the authors to publish this research paper.

Research Design

The data source was a study that was performed to collect data of diabetic patients from October 1, 2002, till September 30, 2003. Demographic characteristics, medical history of illness, health-care utilization, and medical costs were all included in the data. Medical costs were any charges incurred because of a patient's underpayment (i.e., capitation, fee-for-service, or out-of-pocket). The social security office gives hospitals a predetermined amount of money each year to cover the health-care benefits of UC patients, and patients additionally pay 30 baht per visit ($US 1 = 35baht). Out of pocket was for patients who pay the cost all by themselves. Patient Selection was another method where diabetic patients must have at least one claim with diabetes mellitus as the primary, secondary, or tertiary diagnostic code according to the International Statistics Classification Diagnostics and Health Problems tenth version. Statistical Analysis used total healthcare expenses and hospitalizations as the dependent variables. Demographic characteristics, health-care usage, complications, comorbidities, and payment modalities were all independent variables. Multivariate statistical analysis was also performed.

Conclusion

Diabetes patients who used insulin had considerably greater healthcare expenses and a higher likelihood of hospitalization than those who did not. Comorbidities such as hypertension, cancer, nephropathy, retinopathy, and so on were similarly linked to higher healthcare expenditures. Attributes that are influencing healthcare costs and patient admission might help healthcare professionals to improve patient management and minimize healthcare expenses in the future.

4. Diabetes and Gender

Purpose of Study

The purpose of this study is to discuss and analyze the relationship between the occurrence of diabetes and gender. It is a very common and general assumption that it is negligible or zero sex bias when it comes to classifying people as Type 1 Diabetic or Type 2 Diabetic. Sex bias is a feature of autoimmune diseases. In this research, older literature is explored to figure out the hypothesis that males are the victims of the rising incidence of the disease. Diabetes can be hereditary as well. It has been shown that men with type 1 diabetes are more likely to transmit diabetes to their offspring as compared to women affected with diabetes. The authors have observed experiments that might suggest ways of influencing the early course of both times of diseases.

Related Work

This paper brings about research on gender bias and how diabetes can affect people of all genders. This paper can help with narrowing down the age groups that are more likely affected by diabetes. This paper was cited 332 times and has been accessed 8021 times. More than 100 points of references have been used to publish this research.

Research Design

The authors have considered many factors to determine the relationship between diabetes and gender. They considered that the phenotype of autoimmune diabetes changes with age. There was a careful review of sex differences in children. For children under the age of 15, there was a minor bias inclining towards the male in Europe and there was an inclination towards females in Africa and Asia. It was observed that the populations that had an incidence higher than 23 for every 100,000 had a male dominant bias and those with a rate of 4.5 for every 100,000 had a female dominant bias. With increasing age, the

phenotype becomes hard to separate from the type 2 diabetes syndrome. They however determined the existence of diabetes based on the rate and amount of insulin intake. A closer analysis of this paper shows that many cases were not treated with insulin, and this also included women with gestational diabetes. The authors questioned whether male predominance was a new phenomenon in young diabetic adults. The best data came from Norway, where there was an evident dominance of male diabetes in childhood and early adult life. Before introducing insulin there were more deaths in male diabetic patients as compared to female patients up to the age of 50. At this point, there is evidence that western societies have a male bias in diabetes over a wide age range. It is observed in mouse models that the female immune system is rigorously responsive as compared to the male immune system. There could be some conclusions drawn from that perspective towards humans. The authors carried out studies that could be related to hormonal changes, sex, puberty, and pregnancy when it comes to diabetes. The authors could not clearly state whether hormonal changes are responsible for the susception of diabetes of both types.

Conclusion

The authors conclude that the effects of both male and female genders are present in both Type 1 and Type 2 diabetes forms. The observational experiments considering different factors like age, puberty, hormonal changes, and the existence of other autoimmune diseases like Rheumatoid arthritis, which could lead to the existence of diabetes in an individual. The authors concluded that there is a challenge in type 1 and type 2 diabetes to identify and unlock the corrective forces before the individual being diagnosed with the disease.

5. **Impact of Diabetes on Hospital Admission and Length of Stay Among a General Population Aged 45 Year or More: A Record Linkage Study.**

Purpose of Study

In this paper, the authors have searched for the risk involved in hospitalization and the added risk of admission of patients into hospitals due to diabetes. Obesity and age are being considered as a few of the main factors that can cause diabetes in an individual, concluding that diabetes can be considered an epidemic. The key to control diabetes is early detection and practices to improve diabetes care by using health services. The study shows that people with diabetes have a higher readmission rate and longer stay in hospitals than those who don't have diabetes. It is necessary to access the risk of hospitalizing diabetic

patients and how it could impact the general population. In this paper, the authors have described the relationship between hospitalization and a variety of demographic variables, socioeconomic status, lifestyle, and wellbeing of people who have and don't have diabetes. Also, to investigate the increased risk of hospitalization that may be associated with diabetes.

Related Work

The discussion is carried out by the authors to show that diabetes' growing cases and its great influence on the usage of healthcare services, particularly hospitals, is a source of worry for healthcare planners. This paper has been cited 18 times and 40 similar works have been referred by the authors to publish this research paper.

Research Design

Data was collected from The Sax Institute's 45 and Up Study is a cohort study of over 250,000 people. Participants were recruited by filling out a baseline questionnaire and agreeing to long-term follow-up. Most participants were classified as diabetic based on their responses to the question 'Has a doctor ever told you that you have diabetes?' Sedentary physical activity was defined as no physical activity recorded, whereas adequate physical activity included at least 150 minutes of walking in at least 5 sessions per week. The main outcome from this research was any admission of a patient aged 45 or above within 12 months. It was first segregated into being hospitalized at least once or no hospitalization, then the age of the person was seen. The average of hospital admissions, the number of days, and the period of stay of patients with and without diabetes were estimated for 12 months. Methods like descriptive, univariate, and multivariate models were used to perform the analysis. Zero Inflated Poisson was created to account for the large number of people who were hospitalized throughout the research period. The correlations between all causes, non-elective admissions, and ACSC changed over time. In the multivariate model, missing data was also included as "missing values". There was no difference in the intensity of the relationship between a participant's characteristic and hospitalization between people with and without diabetes. Separate models, however, were developed for people with and without diabetes. SAS version 9.3 was used for all analyses.

Conclusion

Hospital admission rates for all-cause admission were 631.3 and 454.8 per 1,000 people for 12 months, respectively, and the mean duration of stay was 8.2 and 7.1 days this is for people with and

without diabetes. Age, gender, smoking, BMI, and so on are all linked to an increased risk of hospitalization. Diabetic male patients who had low income, current smokers, hypertension, and so on were more likely to get admitted. This research is one of the few studies that look at the influence of diabetes on the admission of people in a large non-clinical group.

The reduction of risk, which is directly associated with the characteristics, is mainly responsible for the relation between diabetes and factors like age and obesity. There is an increased risk among individuals that have diabetes, taking gender and income rate into account. This is considering the impact on their health status and their access to health services.

OBJECTIVES OF THE STUDY

1. To determine the spectrum of the age group that is diabetic.
2. To discover the effect of diabetes across races and gender.
3. To discover the correlation between the age group and the number of medications prescribed for diabetic patients.
4. To identify the attributes and factors that play a major part in the readmission of diabetic patients
5. To implement machine learning models to find the best suited algorithm for the problem.

DATA COLLECTION

The source of the dataset is from the UCI website, it includes over 50 attributes showing the patient and hospital outcomes. There are more than 100,000 records in this dataset. The dataset contains data of about 10 years at 130 hospitals in The United States of America. The data is present in the form of a CSV file format. Since the volume of the data is large with over 100,000 records and over 50 attributes it should be sufficient to carry out data analysis even after cleaning the data. There is a wide option of features to choose from to use in the Machine Learning models, making it possible to try the implementation of a model with a different set of features. It contains key features like the type of medication, age group, race, gender, admission type, number of lab procedures, number of medications, number of times the patient has been in the hospital which can be used among the models, and so on.

Dataset source:

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

Variables 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	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.

Img src: https://www.hindawi.com/journals/bmri/2014/781670/tab1/

RESEARCH DESIGN AND METHODOLOGY

1. Data collection.
2. Substitution the missing data values.
3. Data cleaning and preparation.
4. Standardizations.
5. Performing Exploratory Data Analysis.
6. Data visualization.
7. Data Modeling.
8. Prediction using Machine Learning models
9. Visualization of results
10. Comparison of results of Machine Learning models.
11. Determining the most accurate model for the dataset.

The process of the research, after the initial proposal, will start with a collection of data. Every data set is prone to having null values, insignificant attributes, and missing data. To bring uniformity in the data and make it suitable to carry out data analysis, the missing values are either filled or dropped completely. The same applies to the attributes that might not be significant towards the data analysis and prediction. In the dataset, there is a column called *readmitted,* which has 3 types of values. To make it categorical, so that it can be used with ML models, another column called *readmissions* was created and it has the values 1 and 0, for yes and no respectively. Along with the main CSV file, there is another csv file which has the ID mapping. In the main dataset, there were ID allotted to people that signified that they had expired. Entries with these ID mappings were removed from the dataset. After cleaning and preparing the data, the data was brought into a uniform format, making it convenient to use for further steps of research.

Once the data is ready exploratory data analysis was performed on the data. It involves finding the correlation between the features, understanding the features that are present in the data set, and visualize the relation of the variables with the help of graphs and plots. For the dataset, fields like these were taken *'num_procedures', 'num_medications', 'number_emergency', 'time_in_hospital'* and *'num_lab_procedures'* and a scatter matrix was plotted. Next, histograms were plotted for the numerical variables present in the dataset. After this, variables like *‘readmitted’, ’race’, ‘gender’* and *‘age’* were

visualized and plotted against their counts. For the last part of the EDA, the relation between medication and age groups, gender and readmissions, age and readmissions, and their balances were plotted.

After EDA, the data is used with Machine Learning models like Logistic Regression, Decision Tree, Random Forest Classifier, AdaBoosted Classification Model, Hyperparameters Tuning for AdaBoosted, KNN Classifier and Support Vector Machines for Classification. Before using the data for Machine Learning Models, the data had to be prepped. The columns of type number and the columns of type object were selected, and the missing data were filled with 0 and "unknown" respectively. The diagnosis of patients in the dataset are given ICD9 codes (International Classification of Diseases). The columns *'diag_1', 'diag_2'* and *'diag_3'* have the diagnosis codes. These conditions and codes have been mapped with the help of functions *'condition_mapping'* and *'code_mapping'*. On completing this, the data must be normalized. For this *'StandardScaler'* has been used. Now, we can store the data into variables X and Y. Y is going to have the column readmissions and X is going to have the remaining columns of the dataset. X and Y will be split into training and testing data. The split being considered here is 80% and 20%. For the machine learning models, the training and testing data were used to run the models. After running the models, we obtained the accuracy, confusion matrix, and classification report. Each model has a visualization of the confusion matrix. After running all models, an ROC curve was plotted for each algorithm and plotted against each other, to determine the best algorithm among all. Below is a list of algorithms used, and the accuracy scores obtained. The ROC curve plot is also provided in the section below for all algorithms.

MACHINE LEARNING MODELS USED

- ### Logistic Regression

  Logistic regression is a method that is used for statistical analysis. It is used to make a prediction that is based on prior observations of the dataset. It is an essential tool when it comes to machine learning. It predicts a dependent variable by assessing and analyzing the relationship between other independent variables.

- **Decision Tree**

  Decision tree is a supervised machine learning model which functions on the splitting of data continuously, keeping a certain parameter into consideration. A decision tree constitutes of two main parts, namely nodes and leaves. Leaves are usually don't have child nodes and they are usually considered as the outcome of the algorithm. There is classification as well as regression trees.

- **Random Forest**

  Random Forest classifier is an ensemble machine learning method. It is used for both regression and classification applications. Random forest works by creating multiple decision trees. The output is considered as the majority class that is selected by most trees. Random forests outperform decision trees.

- **AdaBoost**

  AdaBoost stands for Adaptive Boosting. It is a boosting technique. Like random forest, this is also an ensemble machine learning method. In this, weights are reassigned to every instance. Instances that are classified correctly have lesser weights assigned to it, and incorrect classifications of instances are allotted higher weights.

- **Hyperparameter tuned AdaBoost**

  The hyperparameters of AdaBoost were tuned and was run on the data. GridSearch was implemented by creating a parameter grid of different number of estimators and learning rates.

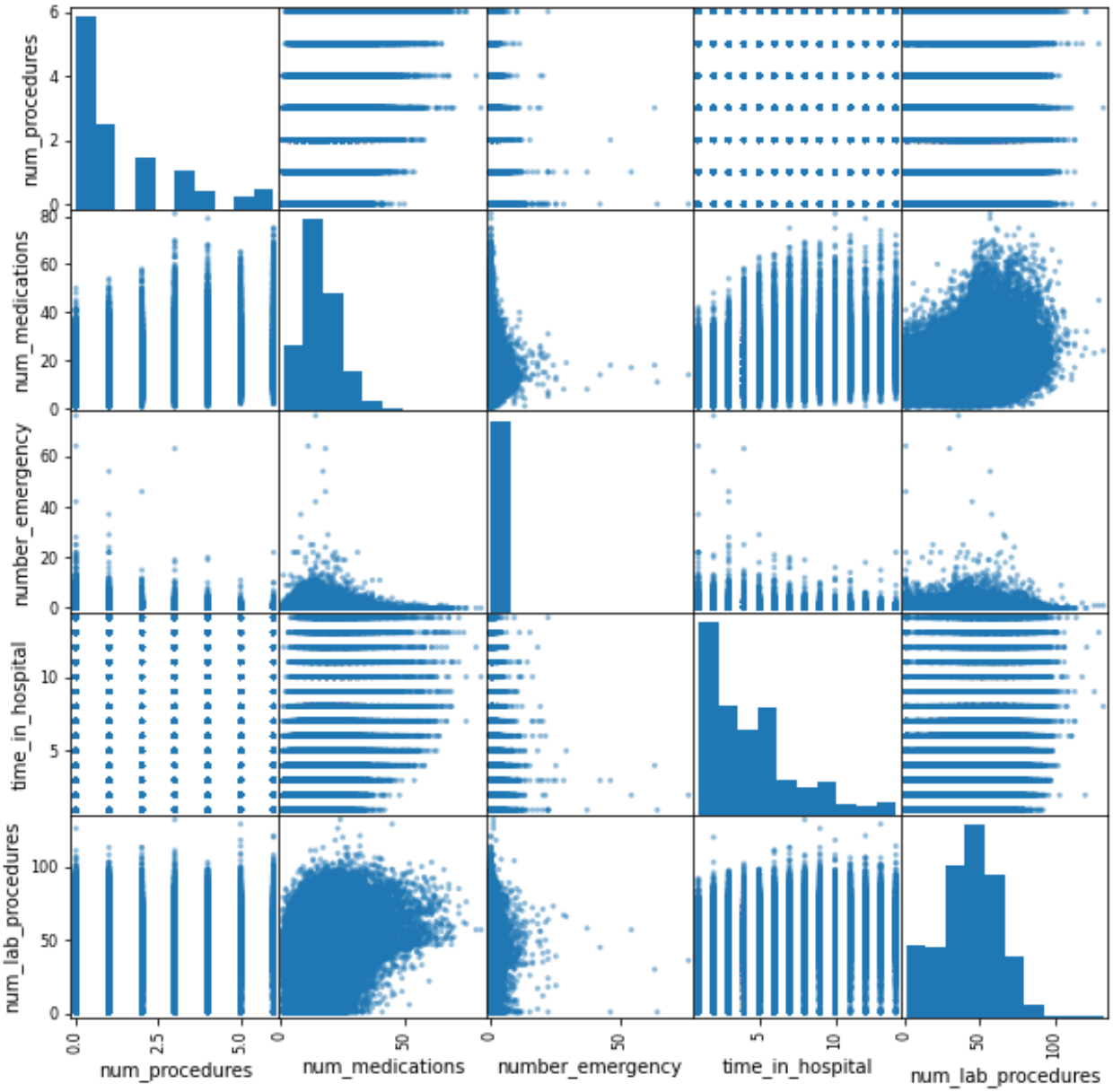
- **KNN Classifier**

  KNN stand for K Nearest Neighbors algorithm. This algorithm is a supervised machine learning algorithm that is also used for regression and classification tasks. The number of nearest neighbors is K. KNN calculates the distance from every data point in the data set and it segregates the data points with the shortest distances.

- **Gaussian Naïve Bayes**

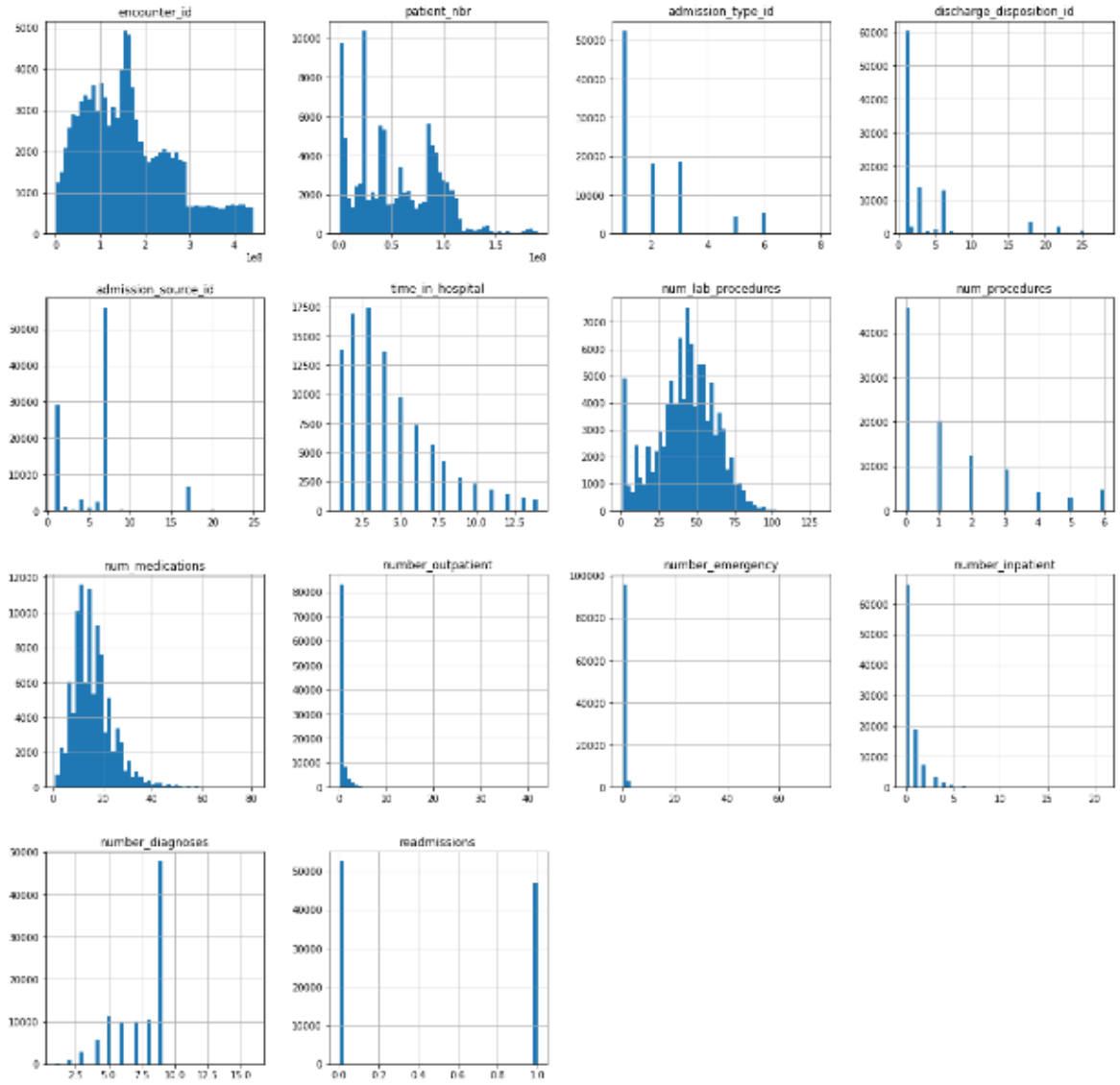
  Gaussian Naïve Bayes is a rendition of the Naïve Bayes algorithm that has the Gaussian distribution as its blueprint. It is a suitable algorithm when it comes to using continuous data. These data, when associated with a class, are distributed according to a Gaussian distribution.

DATA ANALYSIS, DATA VISUALIZATIONS AND RESULTS



The above graph shows the scatter matrix plotted for the variables *'num_procedures', 'num_medications', 'number_emergency', 'time_in_hospital'* and *'num_lab_procedures'*.

```
In [16]: #Plotting of numerical variables
         %matplotlib inline
         df.hist(bins=50, figsize=(20,20))
         plt.show()
```

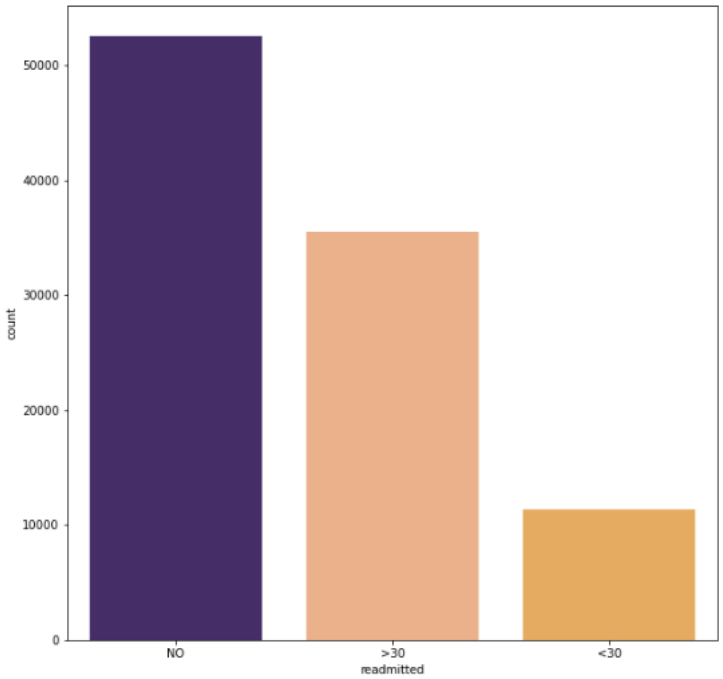


The above image shows the histogram plotted for the numerical variables like *'num_procedures', 'time_in_hospital', 'admission_source_id'* and more in the dataset.

The below graph shows the number of patients *'readmitted'* to the hospital within 30 days, after 30 days and patients who didn't get readmitted at all.

```
In [17]: #Readmissions
         fig, ax = plt.subplots(figsize=(10,10), ncols=1, nrows=1)
         sns.countplot(x="readmitted",data=df, palette=['#432371',"#FAAE7B","#FAAE4B"])
```

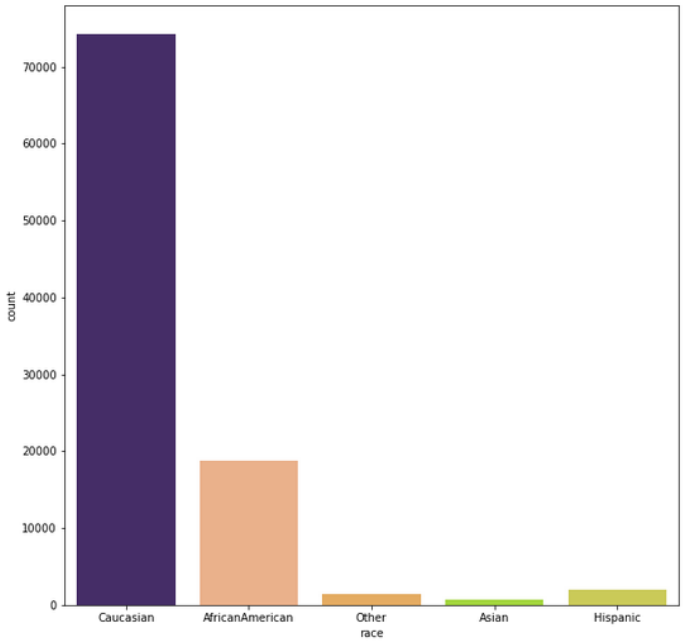
```
Out[17]: <AxesSubplot:xlabel='readmitted', ylabel='count'>
```



The below graph shows the race with the number of people highly affected with diabetes.

```
In [18]: #Races Affected
         fig, ax = plt.subplots(figsize=(10,10), ncols=1, nrows=1)
         sns.countplot(x="race",data=df, palette=['#432371',"#FAAE7B","#FAAE4B","#ABEF23","#DDDD45"])
```

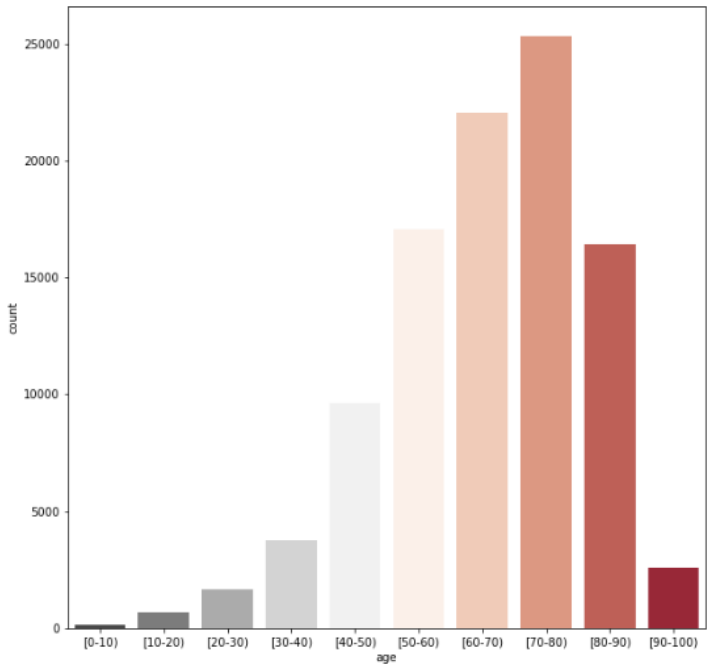
```
Out[18]: <AxesSubplot:xlabel='race', ylabel='count'>
```



The below graph shows all the age group which is affected with diabetes

```
In [19]: #Ages Affected
         fig, ax = plt.subplots(figsize=(10,10), ncols=1, nrows=1)
         sns.countplot(x="age",data=df, palette="RdGy_r")
```

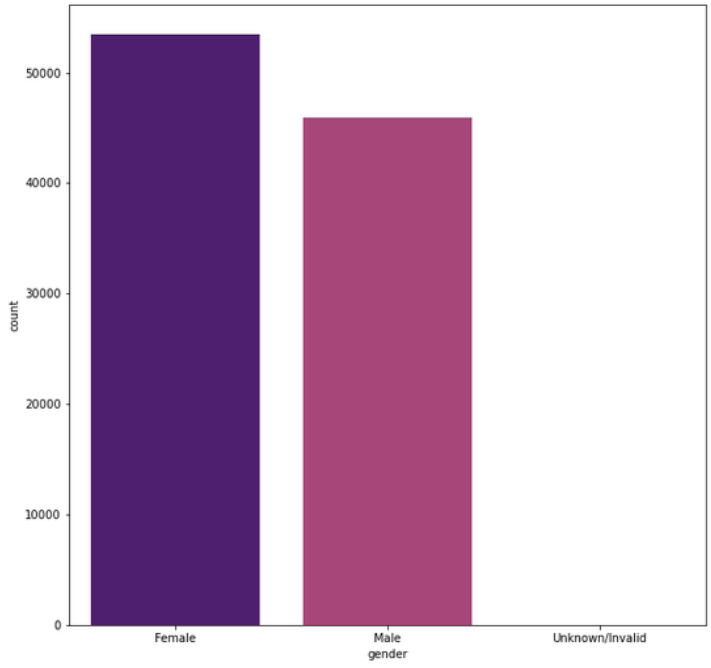
```
Out[19]: <AxesSubplot:xlabel='age', ylabel='count'>
```



The below graph shows the gender affected with diabetes

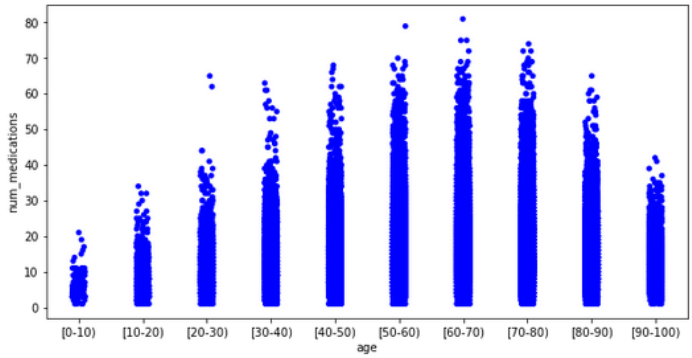
```
In [20]: #Gender
         fig, ax = plt.subplots(figsize=(10,10), ncols=1, nrows=1)
         sns.countplot(x="gender",data=df, palette='magma')
```

```
Out[20]: <AxesSubplot:xlabel='gender', ylabel='count'>
```



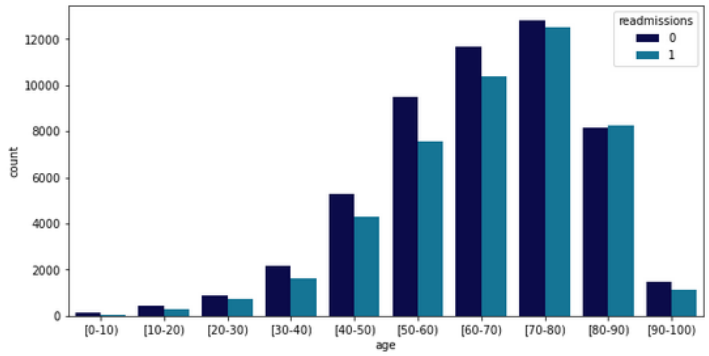
The below graph shows the relations between number of medications and the age group of people.

```
In [21]: #Medications vs Age
age_sort = df.sort_values(by = 'age')
med_plot = sns.stripplot(x = "age", y = "num_medications", data = age_sort, color = 'blue')
med_plot.figure.set_size_inches(10, 5)
plt.show()
```



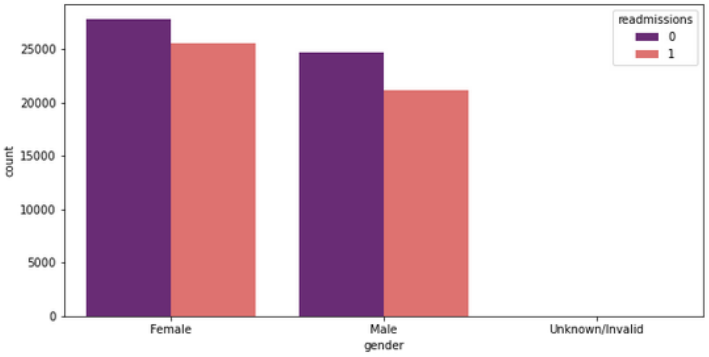
The below graph shows the relations between the readmission rate and the age group of people.

```
In [22]: #Readmission vs Age balance (0 -> No, 1 -> Yes)
ar = df.age.unique()
ar.sort()
ar_sort = list(ar)

plot = sns.countplot( x = 'age', hue = 'readmissions', data = df, order = ar_sort, palette = 'ocean')
plot.figure.set_size_inches(10, 5)
plt.show()
```

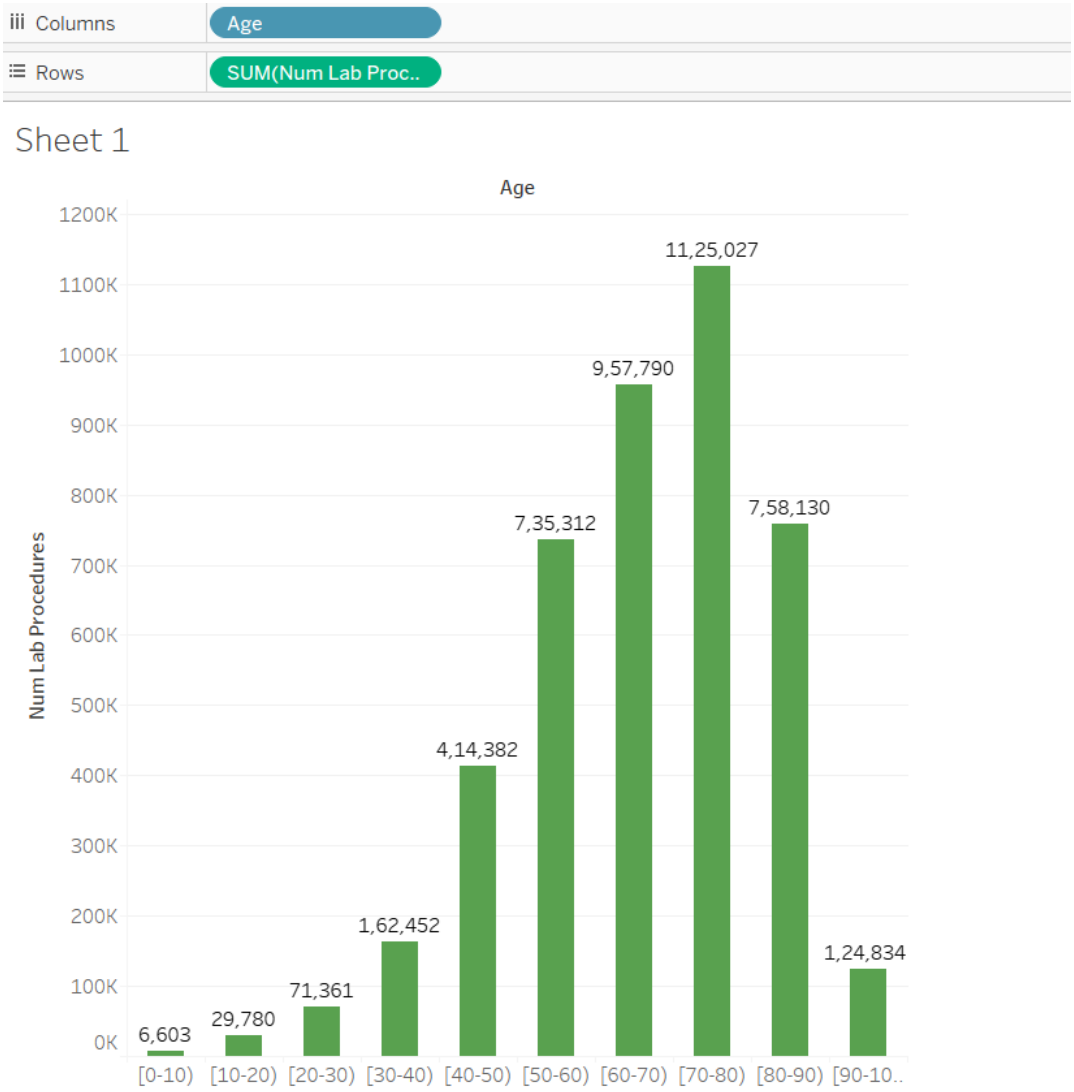


The below graph shows the relations between readmission rate and the gender.

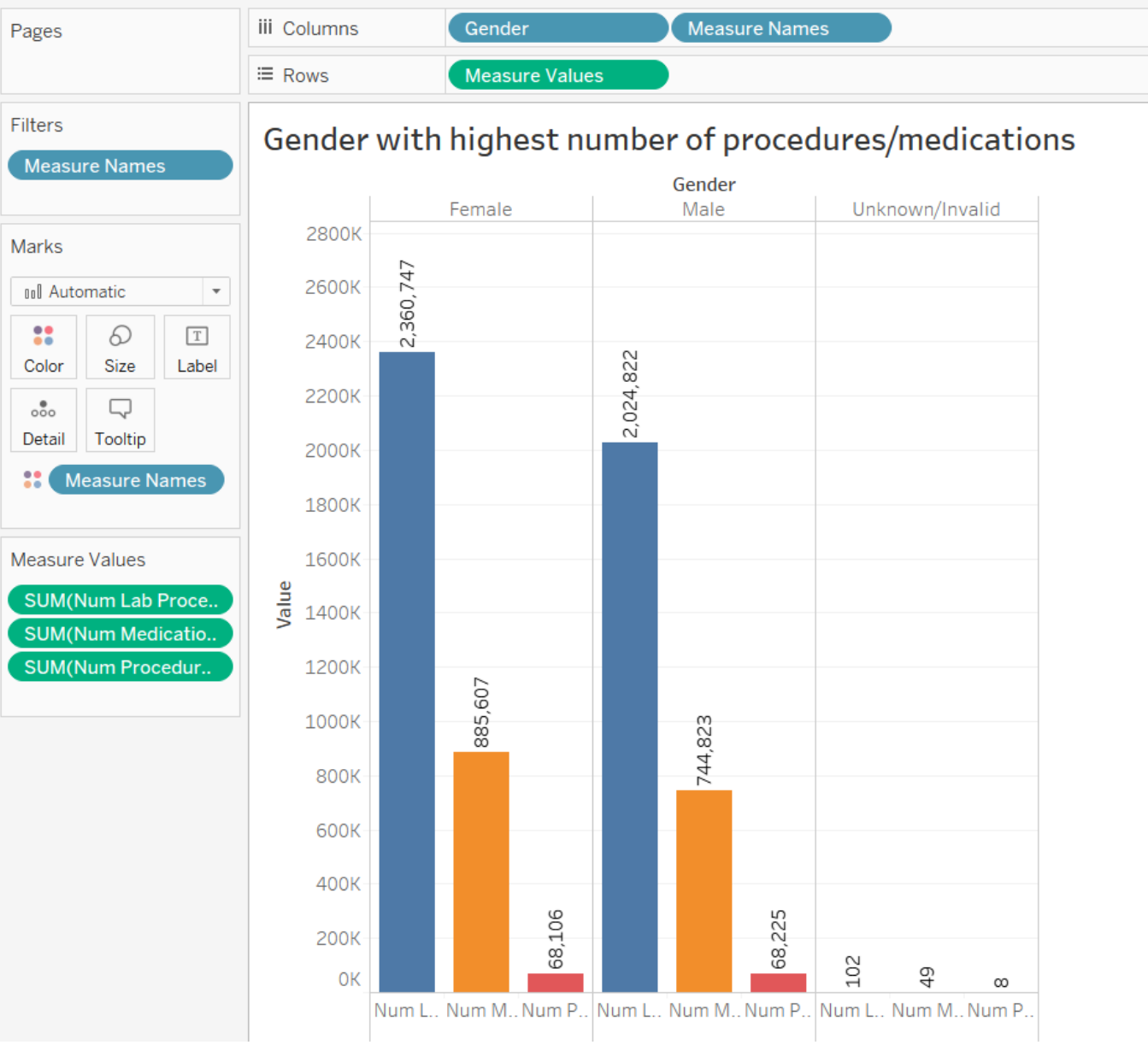
```
In [23]: #Readmission vs Gender (0 -> No, 1 -> Yes)

gr = sns.countplot( x = 'gender', hue = 'readmissions', data = df, palette = 'magma')
gr.figure.set_size_inches(10, 5)
plt.show()
```



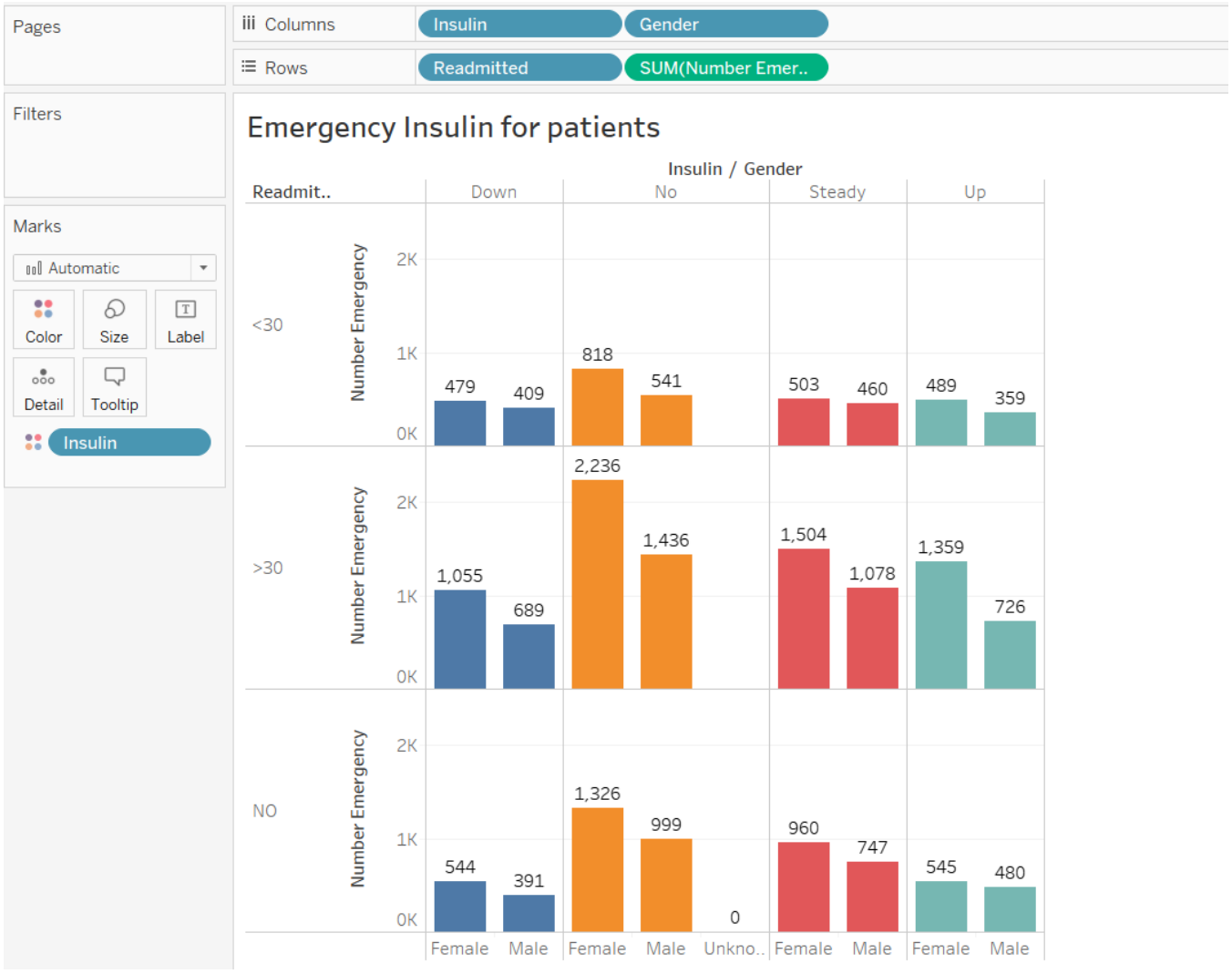
From the below visualization was done in tableau to understand the dataset better. Below we can see that people who fall in the age group of 70-80 years have the highest number of lab procedures.



The below visualization shows which gender has the highest number of lab procedures, num of medications and number of procedures. We can from the visualization that Females have the highest lab procedures, procedures and medications.



The below visualization shows the number of emergency cases that will arrive at a hospital who need insulin or not during the emergency case with readmission details and gender.



The data is split into X_train,Y_train, X_test and Y_test. These training and testing data are used in the Machine Learning models that are implemented.

In [32]:
```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report

X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size = 0.2, random_state = 7, stratify = Y)
print("X_train : {}, X_test : {}".format(X_train.shape, X_test.shape))
```

```
X_train : (79474, 172), X_test : (19869, 172)
```

Logistic Regression

Accuracy

In [33]:
```
log = LogisticRegression(penalty='12', C=0.0005)
log.fit(X_train, Y_train)
log1 = log.predict(X_test)
print("Accuracy: ", log.score(X_test, Y_test))
```

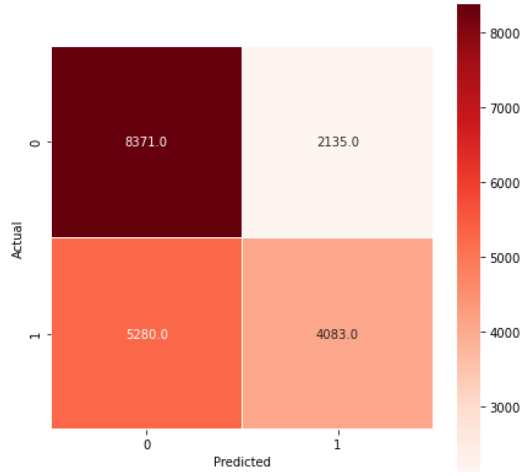
```
Accuracy:  0.6268055765262469
```

Confusion Matrix

In [34]:
```
print(confusion_matrix(Y_test, log1))
```

```
[[8371 2135]
 [5280 4083]]
```

In [35]:
```
plt.figure(figsize=(7,7))
sns.heatmap(confusion_matrix(Y_test, log1), annot = True, fmt=".1f", linewidths= 0.5, square = True, cmap = 'Reds');
plt.ylabel('Actual');
plt.xlabel('Predicted');
```



Classification report

In [36]:
```
print(classification_report(Y_test, log1))
```

```
              precision    recall  f1-score   support

           0       0.61      0.80      0.69     10506
           1       0.66      0.44      0.52      9363

    accuracy                           0.63     19869
   macro avg       0.63      0.62      0.61     19869
weighted avg       0.63      0.63      0.61     19869
```

Decision Tree

Accuracy

In [37]:
```
from sklearn.tree import DecisionTreeClassifier
```

In [38]:
```
dec = DecisionTreeClassifier(criterion = "gini", splitter = "best", random_state = 7, max_depth = 5, min_samples_leaf = 5)
dec.fit(X_train, Y_train)
prediction = dec.predict(X_test)
prediction
```

Out[38]:
```
array([1, 1, 0, ..., 1, 0, 1])
```

In [39]:
```
print("Accuracy: ", accuracy_score(Y_test, prediction))
```

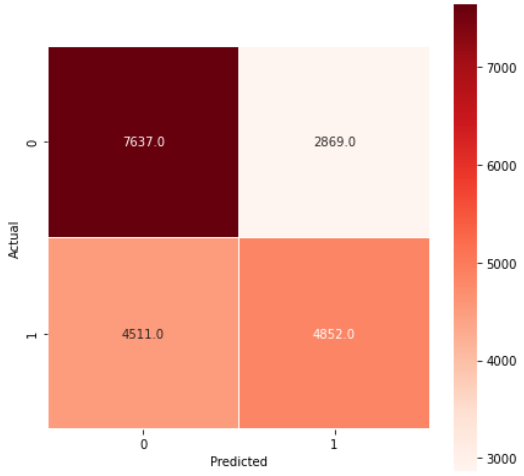
```
Accuracy:  0.6285671146006342
```

Confusion Matrix:

In [40]:
```
print(confusion_matrix(Y_test, prediction))
```

```
[[7637 2869]
 [4511 4852]]
```

In [41]:
```
plt.figure(figsize=(7,7))
sns.heatmap(confusion_matrix(Y_test, prediction), annot = True, fmt=".1f", linewidths= 0.5, square = True, cmap = 'Reds');
plt.ylabel('Actual');
plt.xlabel('Predicted');
```



Classification Report

In [42]:
```
print(classification_report(Y_test, prediction))
```

```
              precision    recall  f1-score   support

           0       0.63      0.73      0.67     10506
           1       0.63      0.52      0.57      9363

    accuracy                           0.63     19869
   macro avg       0.63      0.62      0.62     19869
weighted avg       0.63      0.63      0.62     19869
```

Random Forest Classifier

Accuracy

In [43]:
```
from sklearn.ensemble import RandomForestClassifier

random = RandomForestClassifier(criterion = "gini", random_state = 7, n_estimators = 200)
random.fit(X_train, Y_train)
random1 = random.predict(X_test)
```

In [44]:
```
print("Accuracy: ", random.score(X_test, Y_test))
```

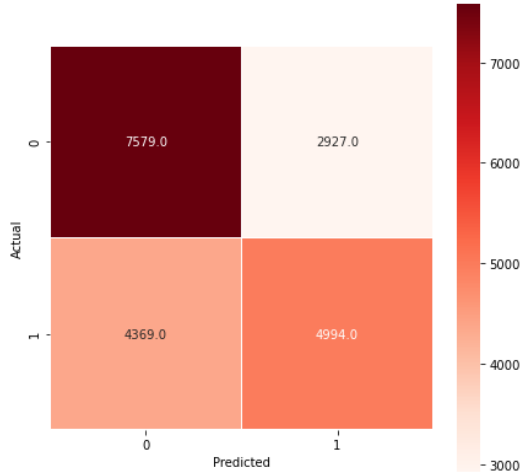
```
Accuracy:  0.6327948059791635
```

Confusion Matrix

In [45]:
```
print(confusion_matrix(Y_test, random1))
```

```
[[7579 2927]
 [4369 4994]]
```

In [46]:
```
plt.figure(figsize=(7,7))
sns.heatmap(confusion_matrix(Y_test, random1), annot = True, fmt=".1f", linewidths= 0.5, square = True, cmap = 'Reds');
plt.ylabel('Actual');
plt.xlabel('Predicted');
```



Classification Report

In [47]:
```
print(classification_report(Y_test, random1))
```

```
              precision    recall  f1-score   support

           0       0.63      0.72      0.68     10506
           1       0.63      0.53      0.58      9363

    accuracy                           0.63     19869
   macro avg       0.63      0.63      0.63     19869
weighted avg       0.63      0.63      0.63     19869
```

AdaBoosted Classification Model

Accuracy

In [48]:
```
from sklearn.ensemble import AdaBoostClassifier
ada = AdaBoostClassifier(n_estimators = 30, learning_rate = 0.3, random_state = 120)
ada.fit(X_train, Y_train)
ada1 = ada.predict(X_test)
```

In [49]:
```
print("Accuracy: ", ada.score(X_test, Y_test))
```

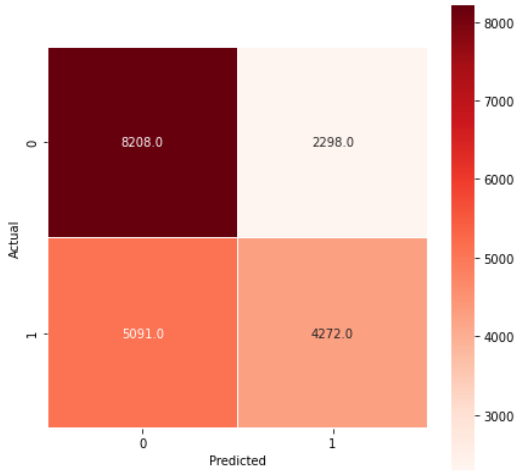
```
Accuracy:  0.6281141476672203
```

Confusion Matrix

In [50]:
```
print(confusion_matrix(Y_test, ada1))
```

```
[[8208 2298]
 [5091 4272]]
```

In [51]:
```
plt.figure(figsize=(7,7))
sns.heatmap(confusion_matrix(Y_test, ada1), annot = True, fmt=".1f", linewidths= 0.5, square = True, cmap = 'Reds');
plt.ylabel('Actual');
plt.xlabel('Predicted');
```



Classification Report

In [52]:
```
print(classification_report(Y_test, ada1))
```

```
              precision    recall  f1-score   support

           0       0.62      0.78      0.69     10506
           1       0.65      0.46      0.54      9363

    accuracy                           0.63     19869
   macro avg       0.63      0.62      0.61     19869
weighted avg       0.63      0.63      0.62     19869
```

Hyperparameter Tuning AdaBoosted Classification Model

Accuracy

In [53]:
```
from sklearn.ensemble import AdaBoostClassifier
from sklearn.model_selection import GridSearchCV

hyper = AdaBoostClassifier(n_estimators = 30, learning_rate = 0.3, random_state = 120)
param_grid = {'n_estimators': [100,150,200], 'learning_rate': [0.1,0.4,1.0]}
grid = GridSearchCV(hyper, cv = 3, n_jobs = 3, param_grid = param_grid)
grid.fit(X_train, Y_train)
grid1 = grid.predict(X_test)
```

In [54]:
```
print("Accuracy: ", grid.score(X_test, Y_test))
```

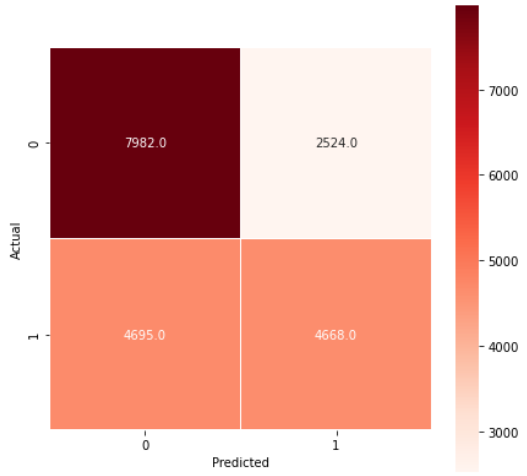
```
Accuracy:  0.6366701897428154
```

Confusion Matrix

In [55]:
```
print(confusion_matrix(Y_test, grid1))
```

```
[[7982 2524]
 [4695 4668]]
```

In [56]:
```
plt.figure(figsize=(7,7))
sns.heatmap(confusion_matrix(Y_test, grid1), annot = True, fmt=".1f", linewidths= 0.5, square = True, cmap = 'Reds');
plt.ylabel('Actual');
plt.xlabel('Predicted');
```



Classification Report

In [57]:
```
print(classification_report(Y_test, grid1))
```

```
              precision    recall  f1-score   support

           0       0.63      0.76      0.69     10506
           1       0.65      0.50      0.56      9363

    accuracy                           0.64     19869
   macro avg       0.64      0.63      0.63     19869
weighted avg       0.64      0.64      0.63     19869
```

KNN Classifier

Accuracy

In [58]:
```
from sklearn.neighbors import KNeighborsClassifier

KNN = KNeighborsClassifier(n_neighbors = 3)
KNN.fit(X_train, Y_train)
KNN1 = KNN.predict(X_test)
```

In [59]:
```
print("Accuracy: ", KNN.score(X_test, Y_test))
```

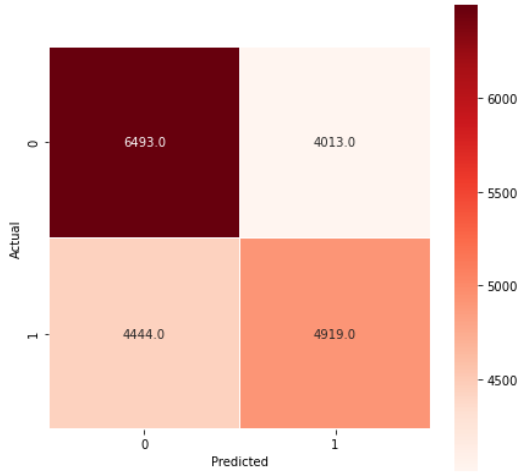
```
Accuracy:  0.5743620715687755
```

Confusion Matrix

In [60]:
```
print(confusion_matrix(Y_test, KNN1))
```

```
[[6493 4013]
 [4444 4919]]
```

In [61]:
```
plt.figure(figsize=(7,7))
sns.heatmap(confusion_matrix(Y_test, KNN1), annot = True, fmt=".1f", linewidths= 0.5, square = True, cmap = 'Reds');
plt.ylabel('Actual');
plt.xlabel('Predicted');
```



Classification Report

In [62]:
```
print(classification_report(Y_test, KNN1))
```

```
              precision    recall  f1-score   support

           0       0.59      0.62      0.61     10506
           1       0.55      0.53      0.54      9363

    accuracy                           0.57     19869
   macro avg       0.57      0.57      0.57     19869
weighted avg       0.57      0.57      0.57     19869
```

Gaussian Naïve Bayes

Accuracy

In [63]:
```
from sklearn.naive_bayes import GaussianNB

GNB = GaussianNB()
GNB.fit(X_train, Y_train)
GNB1 = GNB.predict(X_test)
```

In [64]:
```
print("Accuracy: ", GNB.score(X_test, Y_test))
```

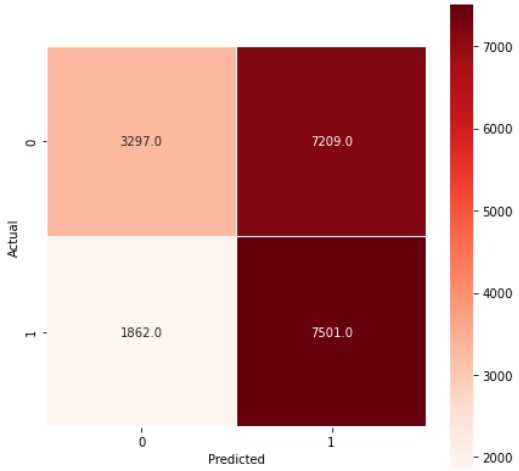
```
Accuracy:  0.5434596607780965
```

Confusion Matrix

In [65]:
```
print(confusion_matrix(Y_test, GNB1))
```

```
[[3297 7209]
 [1862 7501]]
```

In [66]:
```
plt.figure(figsize=(7,7))
sns.heatmap(confusion_matrix(Y_test, GNB1), annot = True, fmt=".1f", linewidths= 0.5, square = True, cmap = 'Reds');
plt.ylabel('Actual');
plt.xlabel('Predicted');
```



Classification Report

In [69]:
```
print(classification_report(Y_test, GNB1))
```

```
              precision    recall  f1-score   support

           0       0.64      0.31      0.42     10506
           1       0.51      0.80      0.62      9363

    accuracy                           0.54     19869
   macro avg       0.57      0.56      0.52     19869
weighted avg       0.58      0.54      0.52     19869
```

USAGE OF RESULTS

The below table shows the results of all the models implemented in this project.

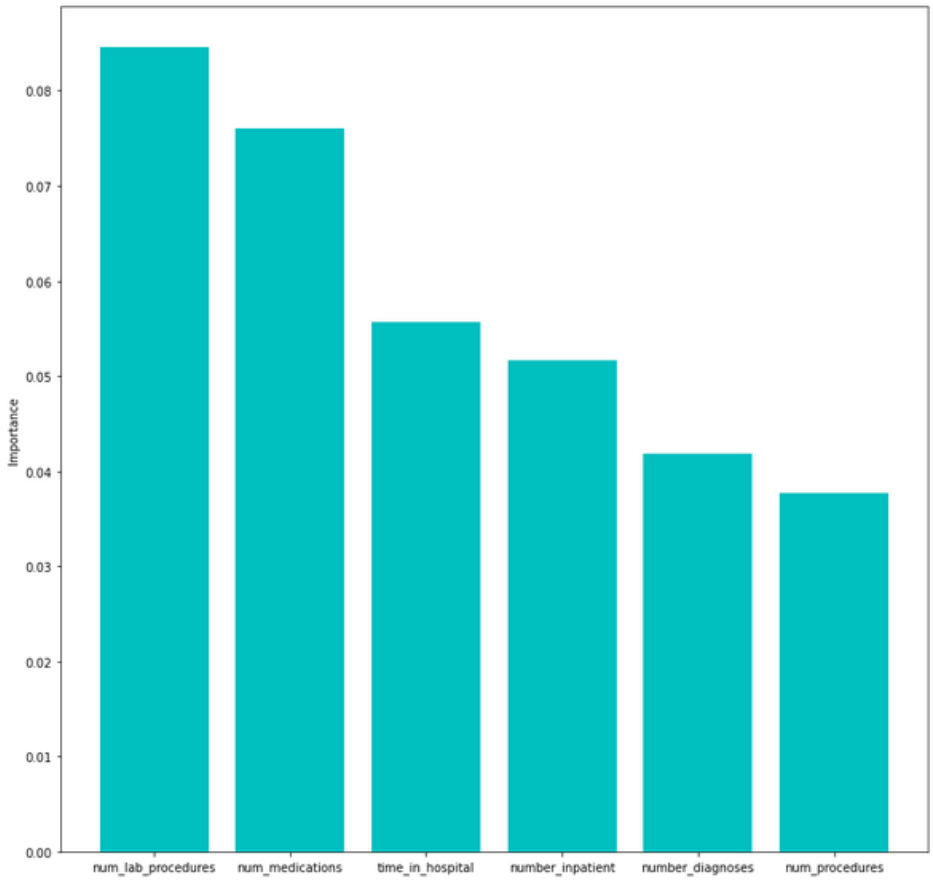
Model Name	Accuracy	Precision	Recall	F1 - Score	Support
Logistic Regression	62.68%	0 – 0.61 1 – 0.66	0 – 0.80 1 – 0.44	0 – 0.69 1 – 0.52	0 – 10506 1 – 9363
Decision Tree	62.86%	0 – 0.63 1 – 0.63	0 – 0.73 1 – 0.52	0 – 0.67 1 – 0.57	0 – 10506 1 – 9363
Random Forest Classifier	**63.28%**	**0 – 0.63** **1 – 0.63**	**0 – 0.72** **1 – 0.53**	**0 – 0.68** **1 – 0.58**	**0 – 10506** **1 – 9363**
AdaBoost Classifier	62.81%	0 – 0.62 1 – 0.65	0 – 0.78 1 – 0.46	0 – 0.69 1 – 0.54	0 – 10506 1 – 9363
Tuned AdaBoost Classifier	**63.67%**	**0 – 0.63** **1 – 0.65**	**0 – 0.76** **1 – 0.50**	**0 – 0.69** **1 – 0.56**	**0 – 10506** **1 – 9363**
KNN classifier	57.44%	0 – 0.59 1 – 0.55	0 – 0.62 1 – 0.53	0 – 0.61 1 – 0.54	0 – 10506 1 – 9363
Gaussian Naïve Bayes	54.35%	0 – 0.64 1 – 0.51	0 – 0.31 1 – 0.80	0 – 0.42 1 – 0.62	0 – 10506 1 – 9363

From the above table, we can see that the models Hyperparameter tuned AdaBoost, and Random Forest Classifier have performed the best among the other algorithms with an accuracy of 63.67% and 63.28% respectively. As the data is related to the medical field, there is only one degree of data similarity that we can expect. Higher degree of accuracies can be obtained with in-depth pre-processing. Supervised machine learning models did the job here, yet we obtained moderate values of accuracy due to the consistency of the data. These models can perform better on more training with larger amounts of data.

The above table has the precision, recall, f1 and accuracy scores of all models. Not just the accuracy, but even the precision and recall scores indicate that Random Forest and Hyperparameter tuned AdaBoost algorithms performed the best with the given data. In the graphs below, the main features that contribute towards readmissions are plotted.

Important Features

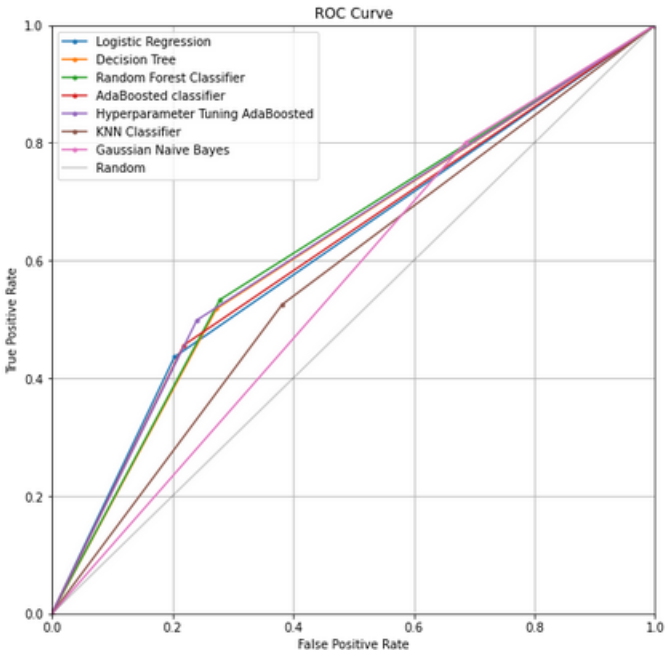
```
In [67]: #Features which affect the readmission rate
feature = X_train.columns
feature1 = random.feature_importances_
imp = pd.DataFrame([i for i in zip(feature, feature1)],
                   columns = ["Feature", "Importance"]).nlargest(6,"Importance")
imp.sort_values(by="Importance", inplace = False)
plt.figure(figsize=(13,13))
pd_len = range(len(imp))
plt.bar(pd_len, imp.Importance, color = 'c')
plt.xticks(pd_len, imp.Feature)
plt.ylabel('Importance')
plt.show()
```



In the above screenshot, the top 5 attributes that are responsible for the readmissions were found. They are num_lab_procedures, num_medications, time_in_hospital, number_inpatient, number_diagnoses, num_procedures. There are many more factors like diagnosis of specific diseases which also influence the increase in the readmission rate. American hospitals can be made aware of these factors, and they can be informed as to how these factors are increasing the rate of readmission. These statistics are constantly tracked, and hospitals with poor service and repeated admissions are majorly affected. Reducing readmissions can reduce costs for patients and at the same time build reputation for the hospitals.

ROC Curve

In [77]:
```
#ROC Curve comparing the models
plt.figure(figsize = (9,9))
plt.plot(fpr_logreg, tpr_logreg, marker = '.', lw = 1.5, label = 'Logistic Regression')
plt.plot(fpr_dec, tpr_dec, marker = '.', lw = 1.5, label = 'Decision Tree')
plt.plot(fpr_random, tpr_random, marker = '.', lw = 1.5, label = 'Random Forest Classifier')
plt.plot(fpr_ada, tpr_ada, marker = '.', lw = 1.5, label = 'AdaBoosted classifier')
plt.plot(fpr_hyper, tpr_hyper, marker = '.', lw = 1.5, label = 'Hyperparameter Tuning AdaBoosted')
plt.plot(fpr_KNN, tpr_KNN, marker = '.', lw = 1.5, label = 'KNN Classifier')
plt.plot(fpr_GNB, tpr_GNB, marker = '.', lw = 1.5, label = 'Gaussian Naive Bayes')
plt.plot([0,1], color='black', label='Random', lw = 1.5, alpha=0.2)
plt.title('ROC Curve')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.xlim([0,1])
plt.ylim([0,1])
plt.xticks(np.arange(0,1.1,0.2))
plt.yticks(np.arange(0,1.1,0.2))
plt.grid()
plt.legend()
plt.show()
```



To compare all the algorithms, the ROC curve was plotted for all the models implemented. In the graph, we can see that the plotting for Random Forest and Hyperparameter Tuned AdaBoost show the best performances as they are the closest to the top left corner of the graph. In this graph True positive rate of algorithms is plotted against False positive rate of algorithms.

SECURITY, PRIVACY, FAIRNESS AND ETHICS ISSUES

Along with the privilege of using medical data, it comes with a lot of privacy issues. Not every patient or hospital is comfortable with sharing the information that they collect on a day-to-day basis. Giving out medical information can sometimes lead to legal complications. There is a streamlined procedure that is involved when it comes to procuring sensitive information. It requires consent and non-disclosure agreements between the hospital/patient and the recipient. Sometimes, medical data is sold illegally for monetary benefits. All this must be taken care of when it comes to medical data. It should not fall in the wrong hands. Doing so can cause distress, grief, and complications to related parties.

LIST THE DATASETS TO BE USED

The dataset used for this project is

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

This link has two CSV files. One file contains the data, another contains the mapping. It is never sufficient when it comes to medical data because, training models with medical data needs a large amount of training data. Eventually, the goal could be making these predictions real time, as and when data is received. There are many datasets and parameters that can contribute to the rate of readmissions.

CONCLUSION

The topic of the research is to predict the hospital readmission rate of diabetic patients. With the dataset that is being used for the research, after cleaning the data, performing EDA and analysis using Machine Learning models. On carrying out the analysis of the data, two models performed better than the rest. The EDA analysis

showed the evident age group, races, and gender affected by diabetes; the medications prescribed for diabetic patients. The main attributes that were responsible for the readmission of diabetic patients were predicted. By carrying out this research, it was possible to determine the quality of the hospital and the cost incurred by the patients with the following results as health care in The United States is expensive. This research probably might show the capability or inadequacies of hospitals. It also can provide room

for improvement for hospitals by concentrating on the key areas. This can help patients control costs incurred by availing healthcare services, and hospitals can build their name by reducing the need to readmit patients at their facility. Application of data science to the field of medicine and health care is extremely helpful and important. It facilitates in taking appropriate decisions for the benefit of the patients as well as the health care givers.

BIBLIOGRAPHY

- Strack, B., DeShazo, J. P., Jennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014, April). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. https://doi.org/10.1155/2014/781670.
- BMC Health Services Research, Wang, W., Fu, C., Zhuo, H., Luo, J., & Xu, B. (2010). Factors affecting costs and utilization of type 2 diabetes healthcare: a cross-sectional survey among 15 hospitals in urban China. http://www.biomedcentral.com/1472-6963/10/244.
- Chaikledkaew, U., Pongchareonsuk, P., Chaiyakunapruk, N., & Ongphiphadhanakul, B. (2008, March). Factors Affecting Health-Care Costs and Hospitalizations among Diabetic Patients in Thai Public Hospitals. https://doi.org/10.1111/j.1524-4733.2008.00369.x.
- Gale, E., & Gillespie, K. (2001, January). Diabetes and Gender. https://doi.org/10.1007/s001250051573.
- Comino, E. J., Harris, M. F., Islam, M. D. F., Tran, D. T., Jalaludin, B., Jorm, L., Flack, J., & Haas, M. (2015, January). Impact of diabetes on hospital admission and length of stay among a general population aged 45 year or more: a record linkage study. https://doi.org/10.1186/s12913-014-0666-2.
- Desarda, A. (2019, January 17). *Understanding AdaBoost - Towards Data Science*. Medium. https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe
- Brownlee, J. (2021, April 26). *How to Develop an AdaBoost Ensemble in Python*. Machine Learning Mastery. https://machinelearningmastery.com/adaboost-ensemble-in-python/
- *Online ICD9/ICD9CM codes*. (2009). Http://Icd9.Chrisendres.Com/. http://icd9.chrisendres.com/