**Project Summary**

|  |  |
| --- | --- |
| **Batch details** | PGP-DSE PUNE JUL’20 |
| **Team members** | Praful Bhoyar  Soyeb Kapasi  Varun Kukday  Vinay Deokar  Yagjna Kurra |
| **Domain of Project** | Healthcare |
| **Proposed project title** | Diabetes Patient Readmission Prediction  Analysis of 100,000 Clinical Database Patient Records |
| **Group Number** | Group 3 |
| **Team Leader** | Yagjna Kurra |
| **Mentor Name** | Srikar Muppidi |

Date: 30 / 03 / 2021

Signature of the Mentor Signature of the Team Leader

Table of Contents

|  |  |  |
| --- | --- | --- |
| Sr. NO | Topic | Page No |
| 1 | Overview | 3 |
| 2 | Business problem goals | 3 |
| 3 | Topic survey in depth | 5 |
| 4 | Critical assessment of topic survey | 6 |
| 5 | Methodology to be followed | 6 |
| 6 | References | 31 |

Project Details

# OVERVIEW:

Hospital readmissions pose additional costs and discomfort for the patient and their occurrences are indicative of deficient health service quality, Hence, efforts are generally made by medical professionals in order to prevent them. These endeavors are especially critical in the case of chronic conditions, such as diabetes. Recent developments in machine learning have been successful at predicting readmissions from the medical history of diabetic patients. These approaches rely on a large number of clinical variables and machine learning models achieving superior prediction performance. In this project, the relationship between diabetes and the various patient attributes is examined. Further, several prediction models are built based on different sets of attributes of the patient.

# Business Problem Understanding:

**Business Understanding:**

It is increasingly recognized that the management of Hyperglycemia (High Blood Sugar) in the hospitalized patient has a significant bearing on the outcome, in terms of both morbidity and mortality. This recognition has led to the development of formalized protocols in the intensive care unit (ICU) setting with rigorous glucose targets in many institutions.

In particular, we examined the use of HbA1c as a marker of attention to diabetes care in a large number of individuals identified as having a diagnosis of diabetes mellitus. We hypothesize that measurement of HbA1c is associated with a reduction in readmission rates in individuals admitted to the hospital.

**Business Objective:**

A hospital readmission is when a patient who is discharged from the hospital, gets re-admitted again within a certain period of time. Hospital readmission rates for certain conditions are now considered an indicator of hospital quality, and also affect the cost of care adversely. Determining which factors are the strongest predictors of hospital readmission in diabetic patients & the efficiency and accuracy of the model in predicting hospital readmission with limited features.

**Approach:**

According to the problem definition we have to build a classification model as our target

variable is a categorical variable. After applying various techniques of EDA,

Feature Engineering into a single data source we would build our machine learning model and check and compare the performance of the different models by evaluating it through the evaluation metrics used in the classification model. Post the model building phase we would try to further improve the efficiency of our models through feature selection and hyper parameter tuning.

**Conclusions:**

The hospital data of in-patients having diabetes as an existing condition in conjunction with other medical illnesses were analyzed to build a predictive model to identify patients who had a higher likelihood of being readmitted. Some of the key factors that drove readmission based on the tuned RF model are discharge\_disposition\_id, number\_diagnoses, race, diag1\_complications Features, A1C result, and glipizide. Using the model, Clinical results, and medication details may be helpful for physicians in the diagnosis in some way but they may provide redundant information we needed for prediction as we have the health service records already. Based on the information provided with these three variables, it can be stated that patients in the top decile have 58 % (According to Random Forest Algo.) Higher likelihood of being readmitted.

# 

# TOPIC SURVEY IN BRIEF

1. **Problem understanding:**

For any healthcare organization, patient readmissions present a major challenge. Currently, one in every five patients who are discharged from the hospital is readmitted in less than 30 days. Hospital readmissions are expensive and more often than not they are avoidable, but avoiding them is still a major challenge for healthcare organizations. Also, the legislation penalizes healthcare organizations with comparatively higher readmission rates, so reducing the readmission rate becomes a necessity. While the reasons behind readmissions of patients are manifold, they are the outcome of inadequate follow-up care or poor discharge procedures. Big data analytics can be taken into account to eliminate unnecessary readmissions that can be avoided by proper post-discharge care. Machine learning can provide doctors with daily updates on patients’ status, predict which patients are more likely to need readmission, and how they might be able to reduce the risk of readmission.

1. **Current solution to the problem:**

At this point, Hospitals can make sure that each patient is provided the highest quality care, keep patients under observation in their homes or do follow up appointments to make sure the patients have been given the proper care and are not at risk of being readmitted. Apart from this there is no clear-cut way for Hospitals to lower the rates of patient readmissions.

1. **Proposed solution to the problem:**

With this Machine Learning Model, we aim to provide Hospitals with a tool to narrow down the number of patients the hospital needs to keep their watch on and make sure that they can mitigate the risk of patient readmission.

# CRITICAL ASSESSMENT OF TOPIC SURVEY

**Find the key area, gaps identified in the topic survey where the project can add value to the customers and business:**

We can solve a problem for Hospitals which reduces the man hours that need to be put into follow ups and observation of patients with a prediction model, which can narrow down the patients the Hospital needs to pay more attention to.

# 

**Data Description:**

**Encounter ID** - Unique identifier of an encounter

**Patient number** - Unique identifier of a patient

**Race** – Race of patient (Caucasian, Asian, African American, Hispanic, and other)

**Gender** - Male, Female, and unknown/invalid

**Age** - Grouped in 10-year intervals: 0-10, 10-20, etc.

**Weight** - Weight in pounds

**Admission type** - Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, new-born and not available

**Discharge disposition** - Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available

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

**Time in hospital** - Integer number of days between admission and discharge

**Payer code** - Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay

**Medical specialty** - 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** - Number of lab tests performed during the encounter

**Number of procedures** - Number of procedures (other than lab tests) performed during the encounter

**Number of medications** - Number of distinct generic names administered during the encounter

**Number of outpatient visits** - Number of outpatient visits of the patient in the year preceding the encounter

**Number of emergency visits** - Number of emergency visits of the patient in the year preceding the encounter

**Number of inpatient visits** - Number of inpatient visits of the patient in the year preceding the encounter

**Diagnosis 1** - The primary diagnosis (coded as first three digits of ICD 9); 848 distinct values

**Diagnosis 2** - Secondary diagnosis (coded as first three digits of ICD 9); 923 distinct values

**Diagnosis 3** - Additional secondary diagnosis (coded as first three digits of ICD 9); 954 distinct values

**Number of diagnoses** - Number of diagnoses entered to the system

**Glucose serum test result** - Indicates the range of the result or if the test was not taken.

**A1c test result** - Indicates the range of the result or if the test was not taken.

**Metformin -** sold under the brand name Glucophage among others, is the first-line medication for the treatment of type 2 diabetes, particularly in people who are overweight.

**Repaglinide -** isused alone or with other medications to control high blood sugar along with a proper diet and exercise program. It is used in people with type 2 diabetes.

**Nateglinide -** is a drug for the treatment of type 2 diabetes

**Chlorpropamide -** isan oral antihyperglycemic agent used for the treatment of non-insulin-dependent diabetes mellitus (NIDDM).

**Glimepiride -** isan oral diabetes medicine that is used together with diet and exercise to improve blood sugar control in adults with type 2 diabetes mellitus.

**Acetohexamide -** is a first-generation sulfonylurea medication used to treat diabetes mellitus type 2, particularly in people whose diabetes cannot be controlled by diet alone.

**Glipizide** - sold under the brand name Glucotrol among others, is an anti-diabetic medication of the sulfonylurea class used to treat type 2 diabetes

**Glyburide -** is a diabetes medicine used to help control blood sugar levels and treat type 2 diabetes.

**Tolbutamide -** is a first-generation potassium channel blocker, sulfonylurea oral hypoglycemic medication. This drug may be used in the management of type 2 diabetes if diet alone is not effective.

**Pioglitazone -** is a diabetes drug (thiazolidinedione-type, also called "glitazones") used along with a proper diet and exercise program to control high blood sugar in patients with type 2 diabetes.

**Rosiglitazone** - is an insulin sensitizing agent and thiazolidinedione that is indicated for the treatment of type 2 diabetes.

**Acarbose -** is an anti-diabetic drug used to treat diabetes mellitus type 2 and, in some countries, prediabetes.

**Miglitol** - is an oral anti-diabetic drug that acts by inhibiting the ability of the patient to break down complex carbohydrates into glucose.

**Troglitazone -** is an antidiabetic and anti-inflammatory drug, and a member of the drug class of the thiazolidinediones. It was prescribed for people with diabetes mellitus type 2.

**Tolazamide -** is an oral blood glucose lowering drug used for people with Type 2 diabetes. It is part of the sulfonylurea family.

**Citoglipton (Sitagliptin)** - Sitagliptin is a diabetes drug that works by increasing levels of natural substances called incretins. Incretins help to control blood sugar by increasing insulin release, especially after a meal. They also decrease the amount of sugar your liver makes.

**Glyburide-Metformin** - The combination of glyburide and metformin is used to treat type 2 diabetes (condition in which the body does not use insulin normally and therefore cannot control the amount of sugar in the blood) in people whose diabetes cannot be controlled by diet and exercise alone.

**Glipizide-Metformin** - Glipizide and Metformin combination is used to treat high blood sugar levels that are caused by a type of diabetes mellitus or sugar diabetes called type 2 diabetes

**Glimepiride-Pioglitazone** - Pioglitazone and glimepiride combination is used with proper diet and exercise to treat high blood sugar levels caused by type 2 diabetes. Pioglitazone works by helping your body use insulin better. Glimepiride stimulates the release of insulin from the pancreas which will help your body turn food into energy

**Metformin-Rosiglitazone** - Rosiglitazone and metformin combination is used to treat a type of diabetes mellitus called type 2 diabetes. It is used together with a proper diet and exercise to help control blood sugar levels.

**Metformin-Pioglitazone** - Metformin/pioglitazone is used to improve blood sugar control in adults with type 2 diabetes. It's used along with diet and exercise. Metformin/pioglitazone isn't used to treat type 1 diabetes.

**24 features for medications** - 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

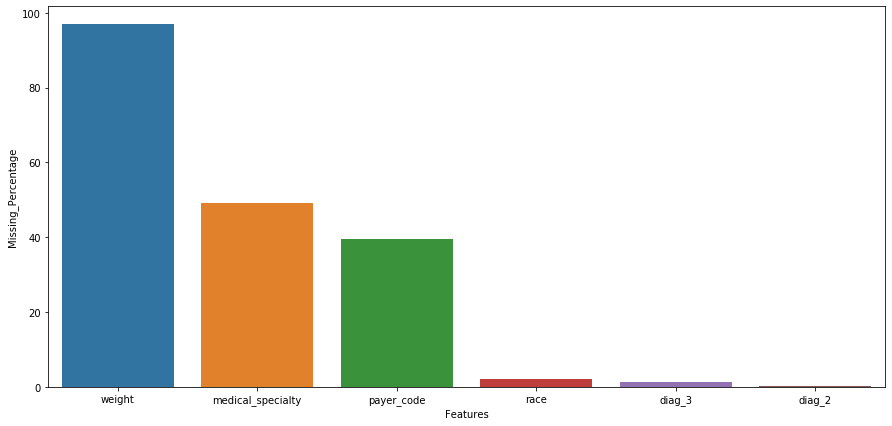
**Change of medications** - Indicates if there was a change in diabetic medications (either dosage or generic name).

**Diabetes medications** - Indicates if there was any diabetic medication prescribed

**Readmitted** - Whether the Patient was Readmitted or not OR whether the Patient was readmitted within 30 days or not.

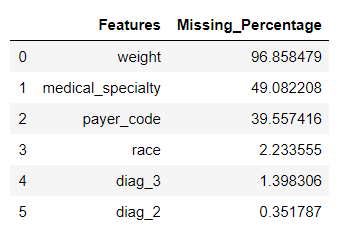
**Data Pre-Processing and Preparation:**

**NULL Handling:**

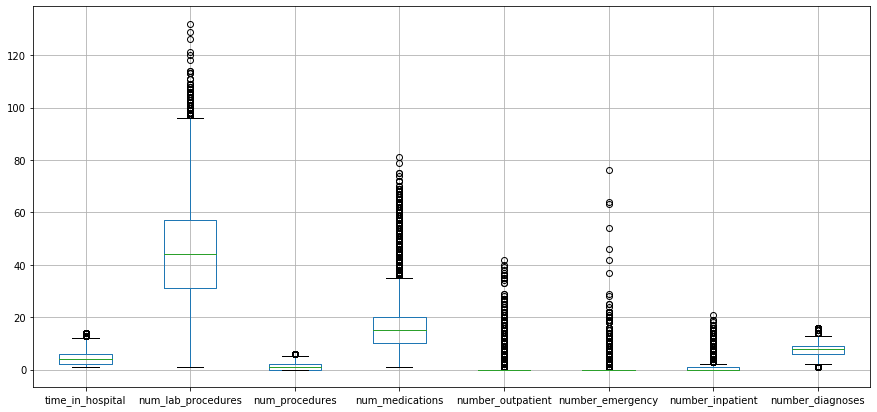
****

There are no traditional “**Null Values**” in the data rather there are some values which are missing that have been filled with “**?**”. We have treated these values in following ways:

* In case of the “**race**” variable - Categorized the missing values to the already existing “**Others**” category
* In case of the “**weight**” variable - dropping the variable altogether where 97% values were missing
* In case of the “**payer\_code**” variable - dropped the variable as here were too many existing classes making accurate imputation of missing values almost impossible
* In case of the “**medical\_speciality**” variable - dropped the variable as here were too many existing classes making accurate imputation of missing values almost impossible.

In the case of diag\_2 and diag\_3 variables we have been able to impute those values with the Feature Engineering.

**Outlier Handling:**

****

As this is real world patient data, the outliers are the main patients who are affecting our Target.

For example, patients with a greater number of inpatient records are getting readmitted, patients who have gone through more lab procedures are less likely to get readmitted.

Thus, we have chosen not to do outlier treatment so as to preserve the integrity of the dataset which is linked to real world outcomes.

**Exploratory Data Analysis & Business Insights:**

**UNIVARIATE:**

We can see from the data that on average, a patient spends about 4.3 days in the hospital, and around 76% of the patients we have in the data are of the Caucasian race i.e., White, with African Americans patients in the data being around 19%. As for the “**Gender**” of the patients we have a more even distribution with around 53% Female patients and 47% Male Patients.

The data categorizes the “**Age”** of the patients in the 10-year intervals, i.e., 10-20, 20-30, etc. with the majority of patients in the data are between the ages of 50 and 90. We have around 97% Missing values in the “**Weight**” column, deeming it unfit for consideration.

We have 71518 unique patients with some of them having multiple visits. We were able to create a new feature **‘Patient\_Visits’** by calculating based on the unique patient numbers how many visits the patient has had over the span of the 10 years in which the data was collected.

We have 3 variables “**Admission\_type\_id**”, “**Admission\_source\_id**” and “**Discharge\_dispostion\_id**” which have been pre-encoded with unique classifications for the type circumstance of the Patient’s Admission and Discharge from the Hospital. We know from the data that the most common admission types are Emergency & Urgent admissions along with elective admission, the most common sources of admission is again Emergency or through Referrals. The most common types of discharge granted to patients being discharged to their homes or transferred to another facility.

We changed the mapping of these labels to a more manageable number of classes for the three variables, especially in the case of “**Admission\_source\_id**” and “**Discharge\_dispostion\_id**” where there were around 30 classes which we were able to reduce down to 4 and 7 categories respectively.

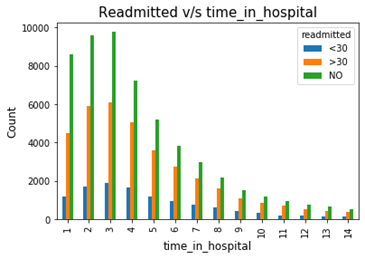
The “**payer\_code**” is the identifier of the mode of payment by the patients i.e through insurance provider, blue cross or self-pay, etc with MC or Medicaid being the most common mode of payment amongst the patients.

We can see from the analysis of the Drugs in the data, the majority of the patients are being prescribed one or more of the drugs with a steady dosage, with the exception of the drugs **‘Examide’** and **‘Citoglipton’** where there are no patients being prescribed either drug.

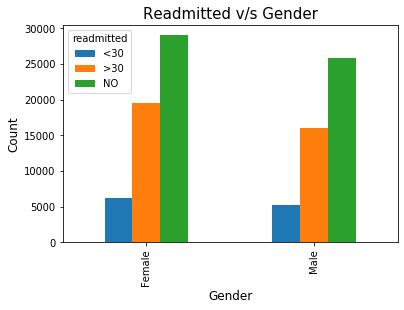
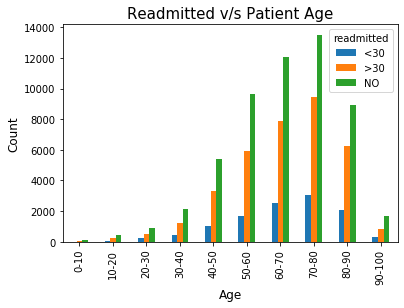
As for the target variable “**Readmitted**”, we have 54864 patients who were not readmitted while 35545 patients having to be readmitted after 30 days and 11357 patients having to be readmitted before 30 days from date of discharge.

**BIVARIATE:**

From the bi-variate analysis done with the target variable, the main inferences we can draw are that the higher the “**time\_in\_hospital**” i.e., the duration the patients were admitted in the hospital for, the lower the likelihood of their re-admission, meaning that the patients who were under hospital care for longer, are less likely to be readmitted.



Another inference or insight we have gained is that the “**Age**” and “**Gender**” are also critical factors with respect to the likelihood of readmission for the patients, the higher **Age** increases the likelihood of readmission, and in case of **Gender**, Females have more likelihood of being readmitted.

**Feature Engineering:**

Feature engineering is about creating new input features from your existing ones. These features can be used to improve the performance of machine learning algorithms.

In our analysis we have implemented feature engineering techniques on a few features in order to gain more insights from our data.

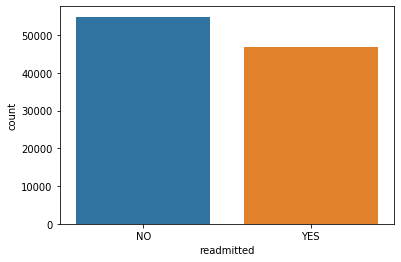
New Features Created / Transformed from existing Features for making Interpretations:

* **Readmitted (Target Variable)**
* **Admission\_type\_id**
* **Admission\_source\_id**
* **Discharge\_disposition\_id**
* **Diagnosis Kind**
* **Diagnosis Complications**

We made the decision after building our first base models, to change the Target Variable “**Readmitted**” from a multi-class variable into a binary **‘Yes’** or **‘No’** i.e. whether a patient was readmitted or not.

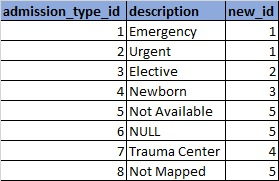
This was done to increase the clarity of our primary objective which was the prediction of patient readmission and also solve the problem of class distribution in the Target Variable.

**Readmitted:**

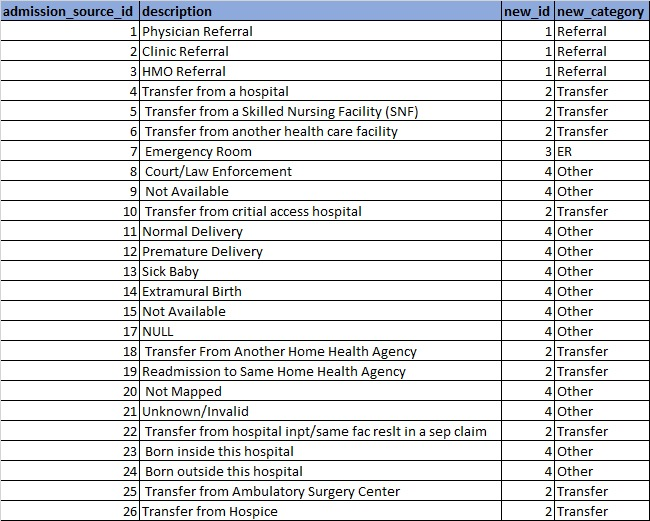


We changed the mapping of the labels in the “**Admission\_type\_id**”, “**Admission\_source\_id**” and “**Discharge\_dispostion\_id**” by clubbing together similar classifications into one main classification label and were able to reduce the number of labels in each of these variables to more interpretable and more accurate representation.

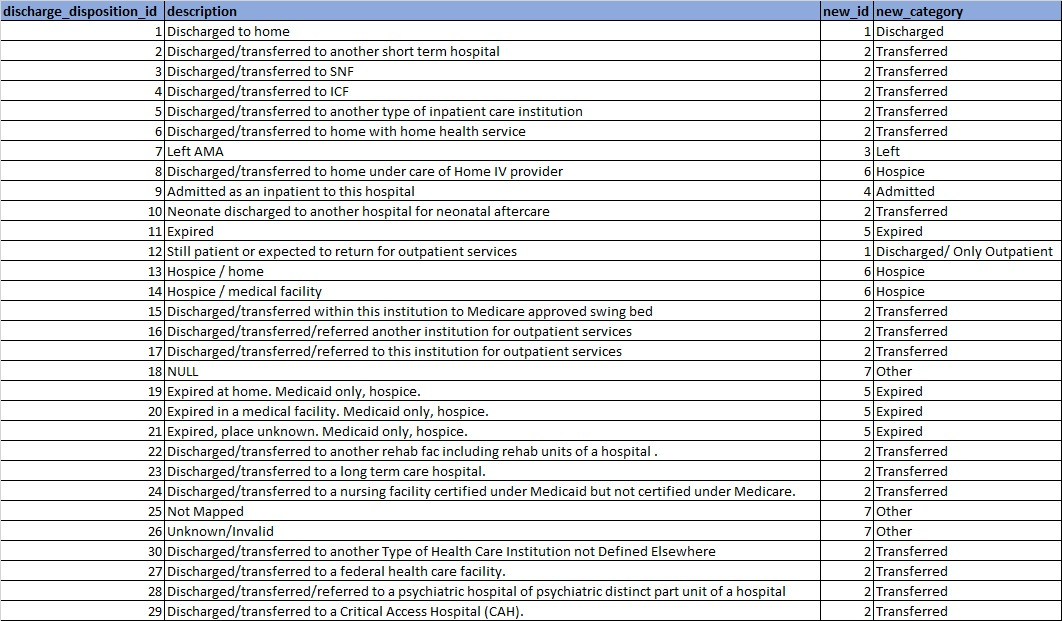
**Admission\_type\_id:**



**Admission\_source\_id:**



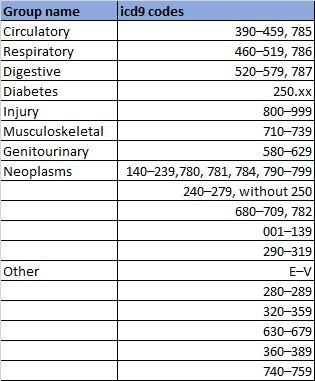
**Discharge\_dispostion\_id:**



“**Diagnosis Kind**” , “**Diagnosis Complications**” Features are created based on the ICD-9 (International Classification of Diseases ) Codes which are shared under the diagnoses noted down. These ICD-9 Codes are the method of capturing the diagnoses in US Hospitals.

**Diagnosis Kind:**

The code-mapping for Diagnosis Kind is described below:



We could see here how the Diagnosis Kind was affecting my Target of Readmission:





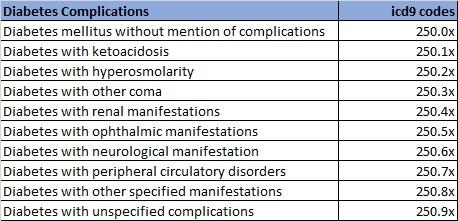


Based on the above graphs we could see that the people getting diagnosed are getting detected by the 3rd diagnosis, as the number of diabetic kind diagnoses are increasing from 1st to 2nd to 3rd Diagnostic Test.

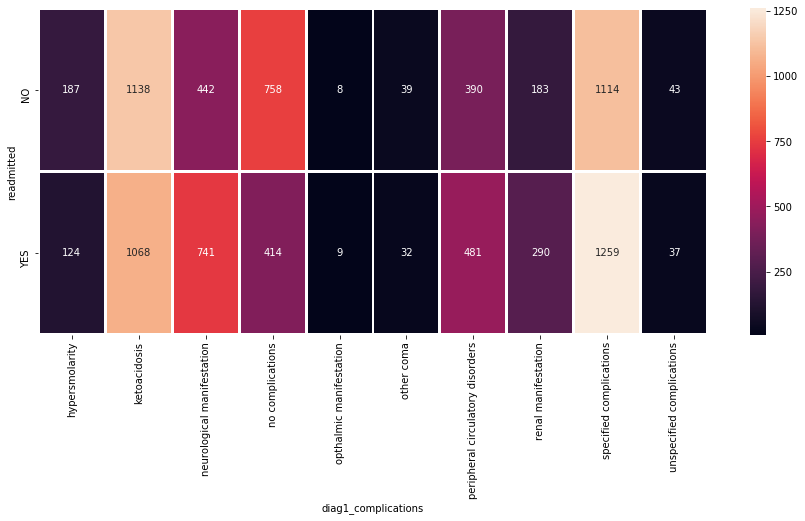
This might be a reason for the increase in readmission percentage of the patients!

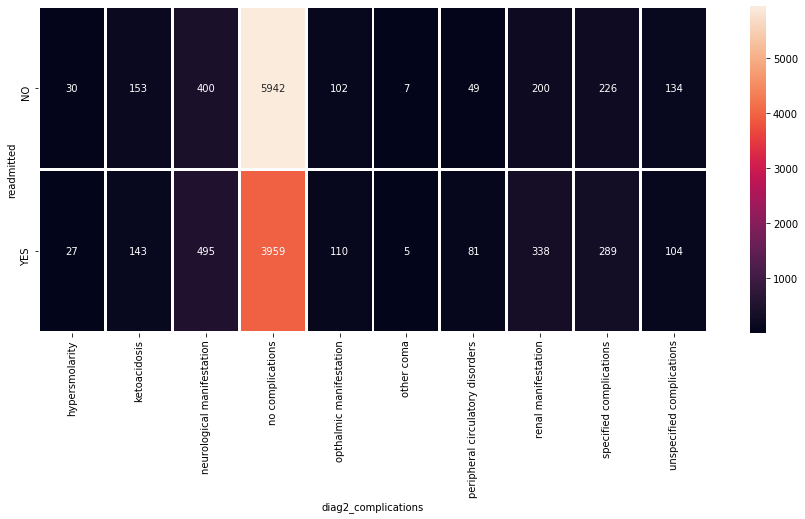
**Diabetes Complications:**

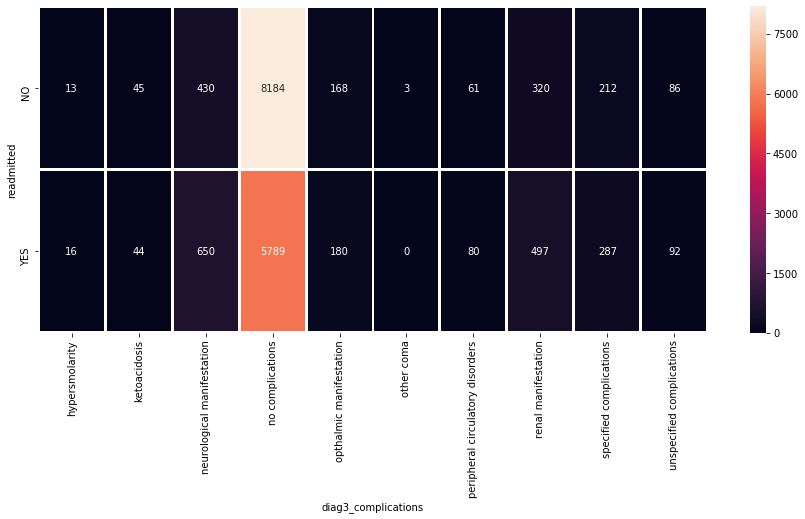
The Codes mapped for Diagnosis Complications are only considered here towards Diabetes Patients and None was assigned for others. The Mapping is as given below:



We could see here how the Diagnosis Kind was affecting my Target of Readmission:







We could see that major diagnoses of Diabetic people are related to Ketoacidosis and Neurological manifestations.

**Ketoacidosis** is a serious complication of diabetes that occurs when your body produces high levels of blood acids called ketones. The condition develops when your body can't produce enough insulin. Insulin plays a major role for Glucose (Sugar) to enter the cells which in turn provides Energy!

**Neurologic** disorders are a common and often disabling aspect of diabetes mellitus. Pain and sensory disturbances, weakness and paralysis and symptoms of autonomic dysfunction may be experienced by the diabetic patient.

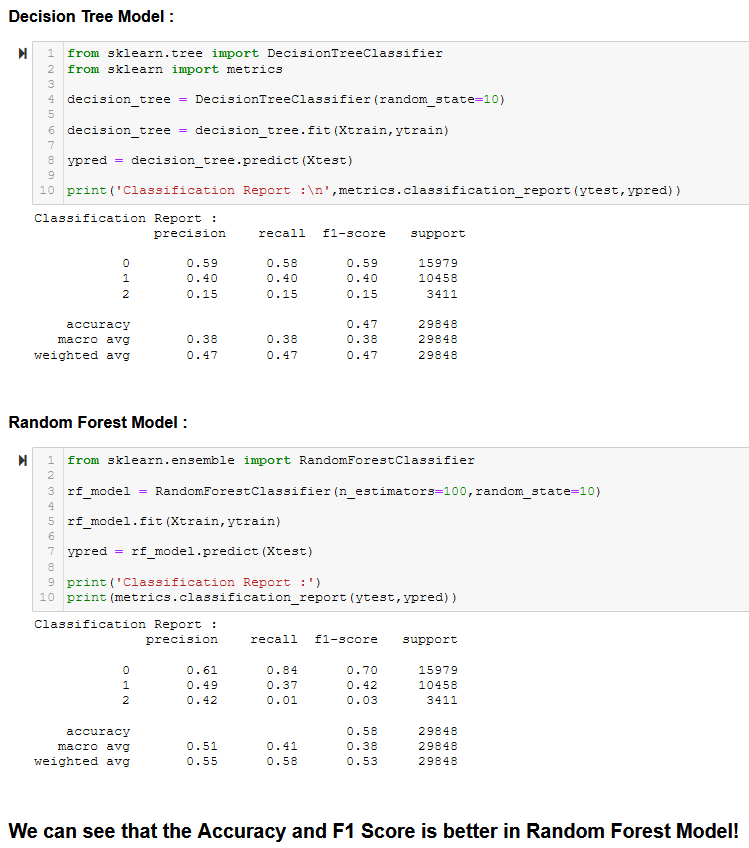
These might be the Major Focus Areas for the Hospital for preemptive care Measures to be taken which would in turn reduce the Patient Readmission.

**Patient Visits**: A new feature of Patient Visits was manufactured from the patient number. As this data is around 9 years from 1999-2008, the first occurrence of individual patients is given as count 1 and then the count was given in an incremental fashion for each patient.



Based on this new information gained from the **codes** and **patient\_nbr**, we made new features of Kind and Complications for each diagnosis 1, 2 and 3 and patient\_visits.   
Post this, we chose to drop the ICD codes and patient\_nbr which were not giving any other information.

**Basic Model:**

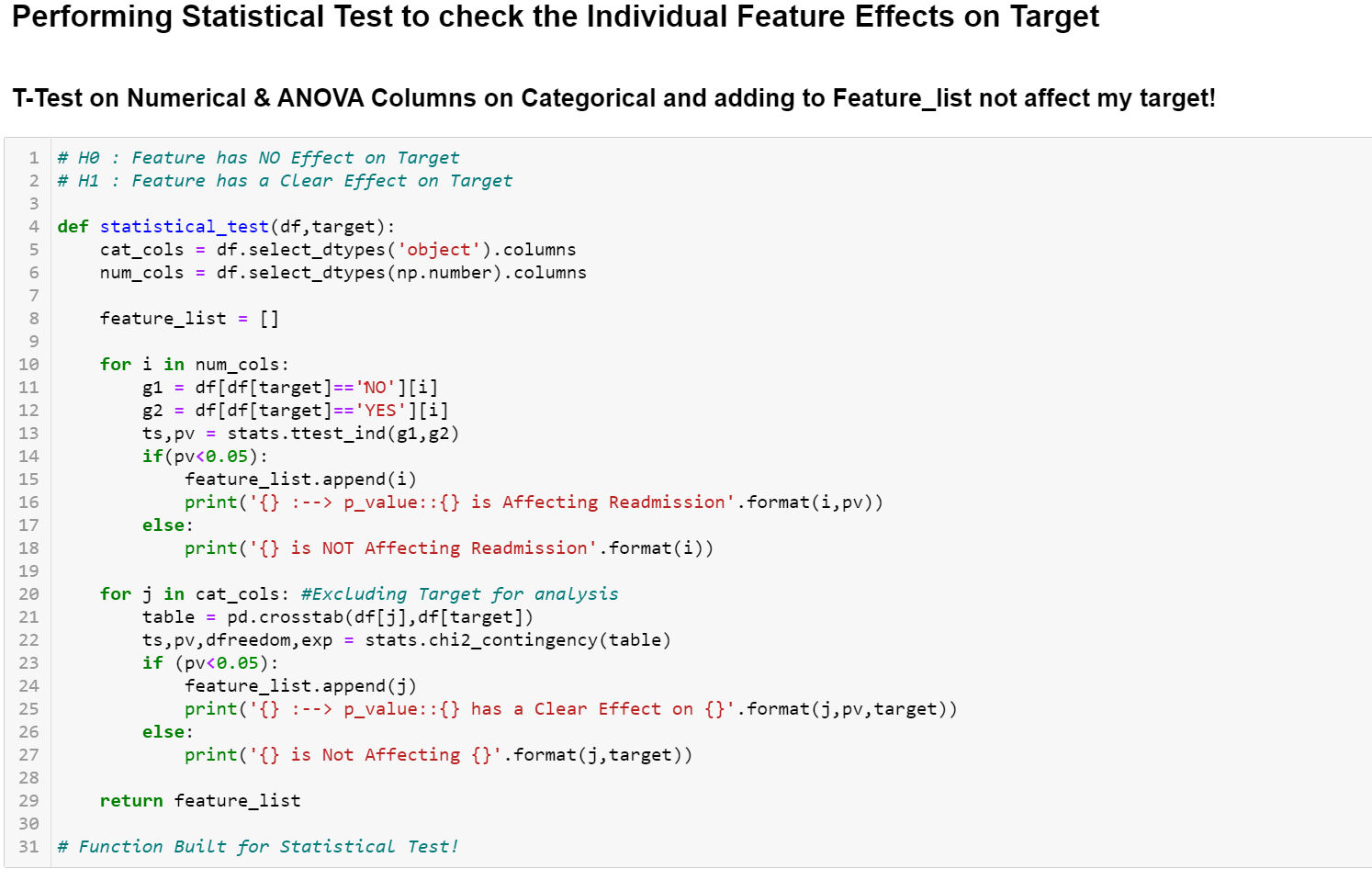
****

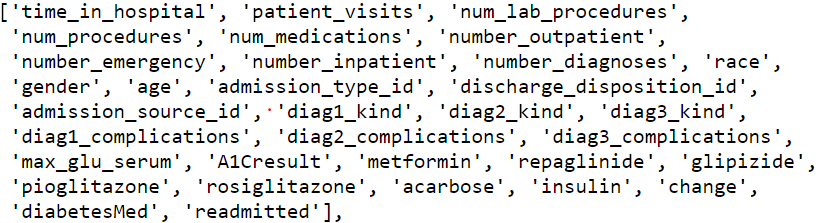
# 

**MODEL TUNING:**

After the Data Preparation, Exploratory Data Analysis and Feature Engineering, Data Encoding and building our base models which were the “**Decision Tree Classifier**” and the “**Random Forest Classifier**”, we now further tune the model to optimize our key metrics which are **Precision**, i.e. the proportion of positive instances that were correctly predicted and **Recall,** i.e the proportion of actual positive cases that were correctly predicted, also sometimes called ‘*True Positive Rate (TPR*)’ or *‘Sensitivity’*. We will do the Model Tuning in the following ways:

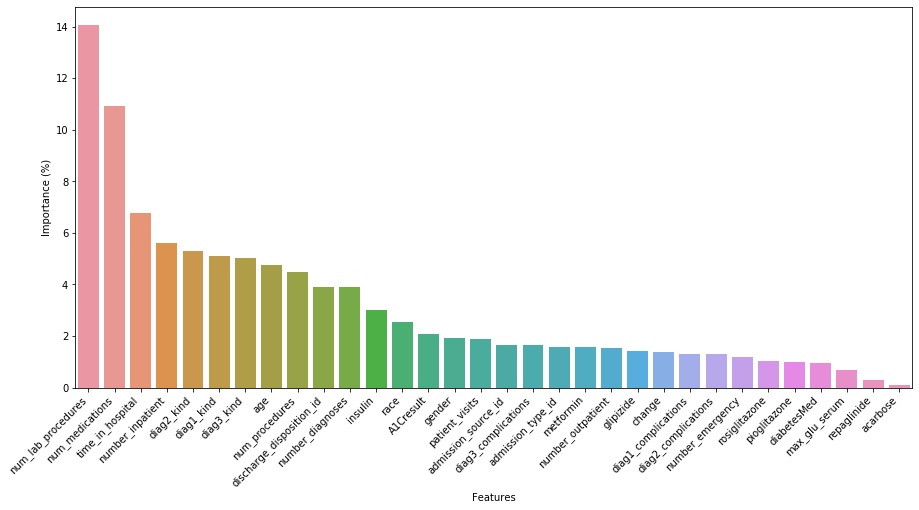
1. Run Statistical Tests to find out the Significant Features in our Data.

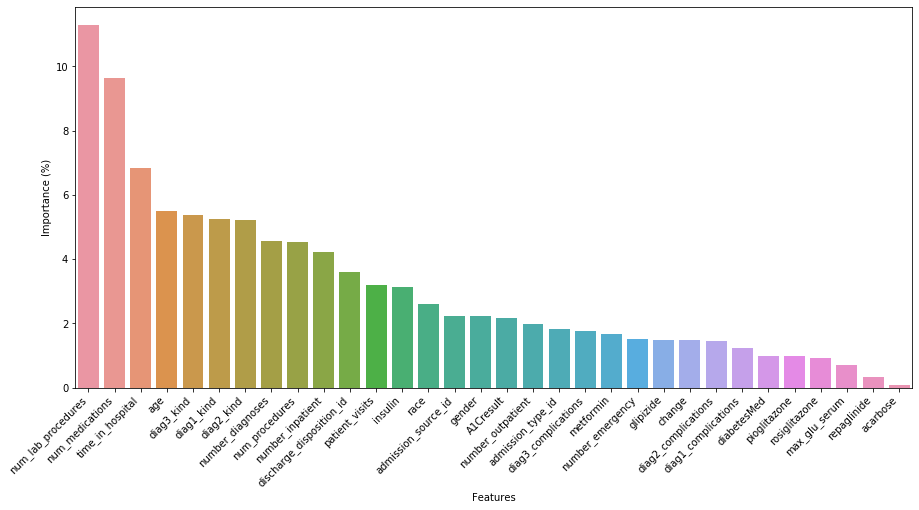




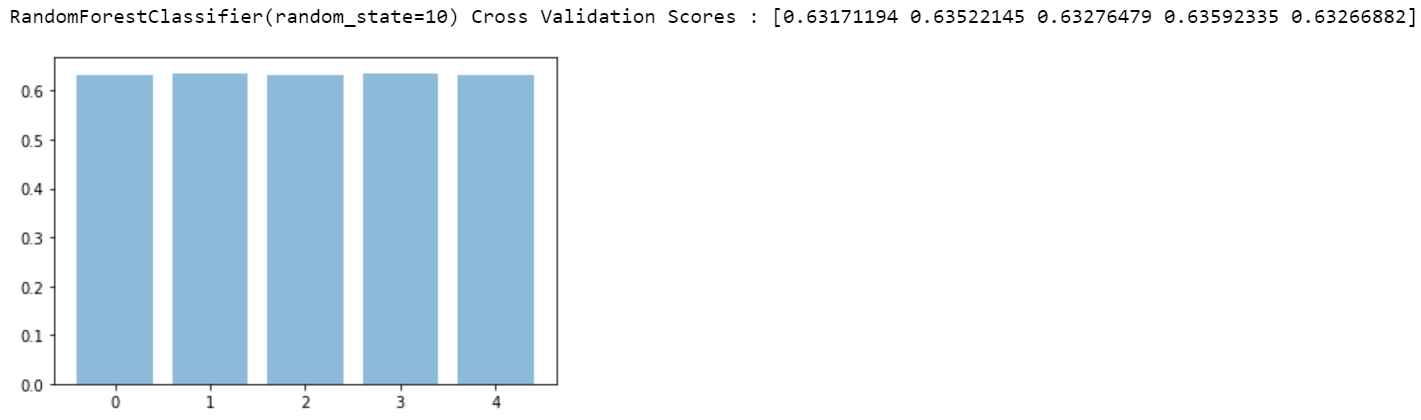
The Following Features are obtained as Significant.

1. Finding out the Feature Importance of our Base Models.

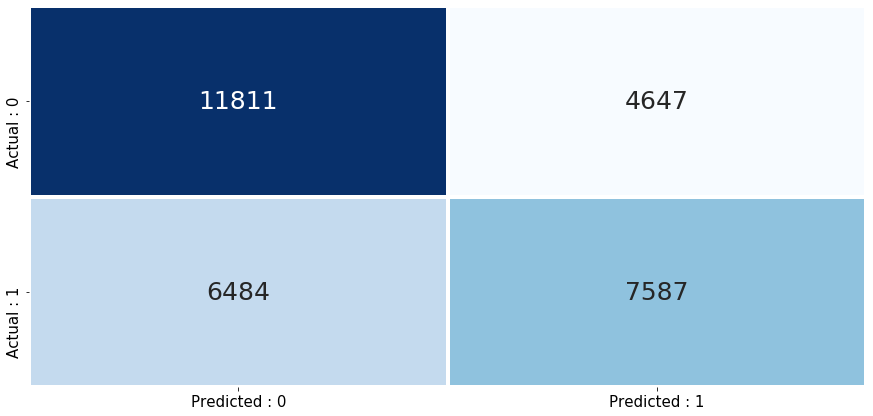




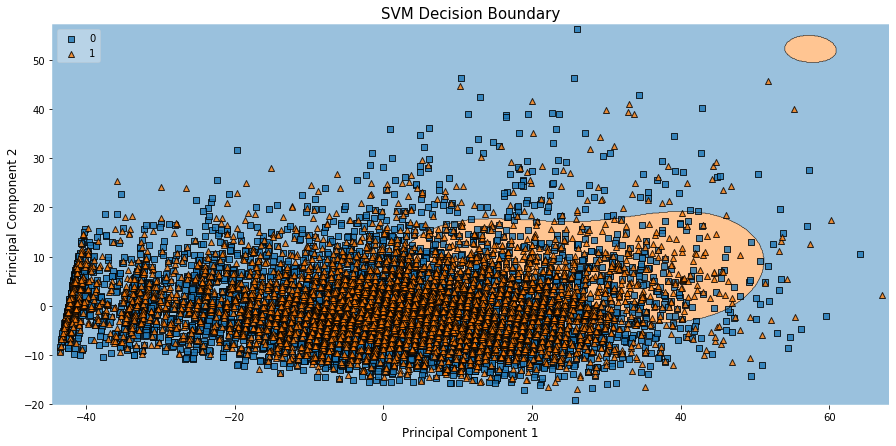
1. Creating a New Data frame with only the Significant Features and Build new Models on this Data frame.
2. Making sure the data is not Overfitting or Underfitting with use of “**Stratified K-Fold**” and “**Cross Validation Score**”.



1. Plotting a **Confusion Matrix** to see the Distribution of Correctly and Incorrectly Classified Data Points for each Model.

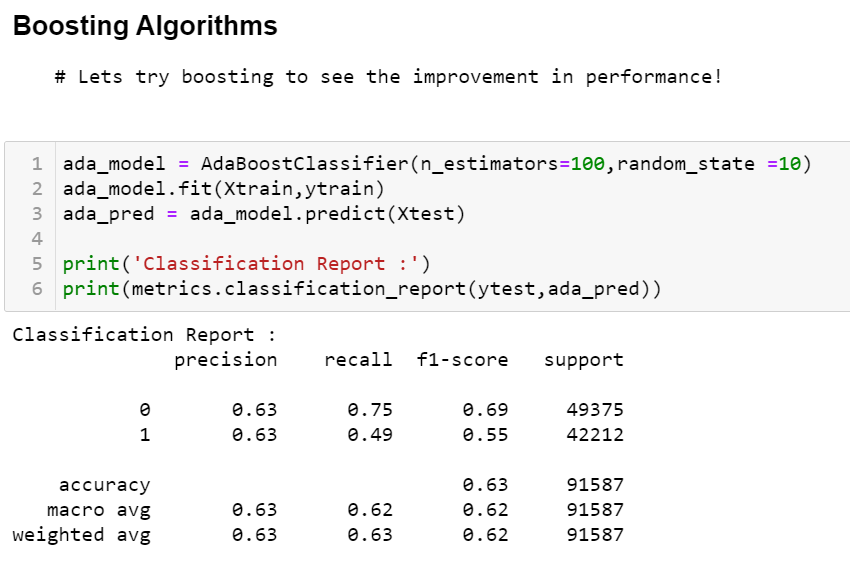


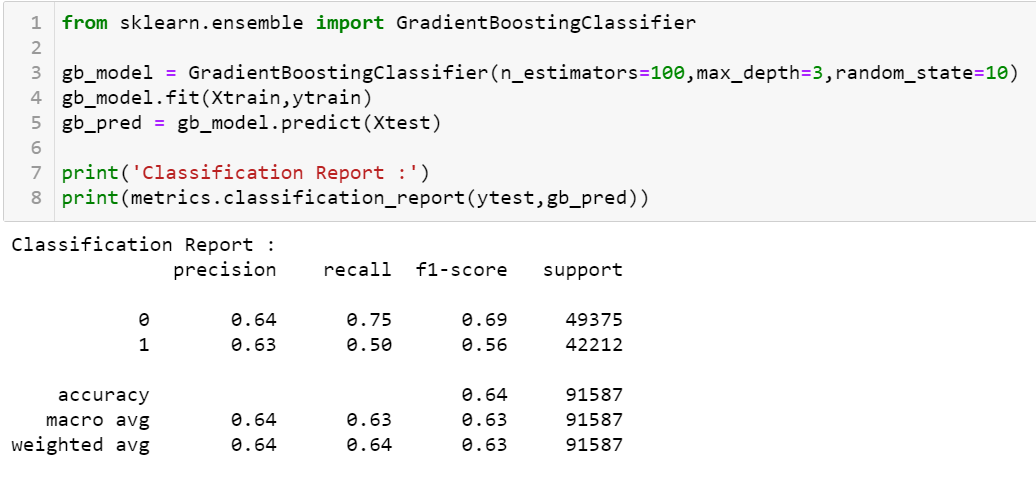
1. Using **PCA** (Principal Component Analysis) and a **SVM** (Support Vector Machine) on the data to gauge the Decision Boundary of the SVM Classifier.

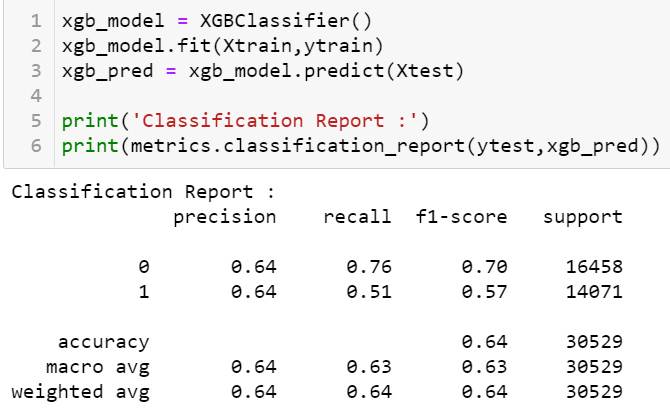


As the Data is not Linear, we have not pursued PCA.

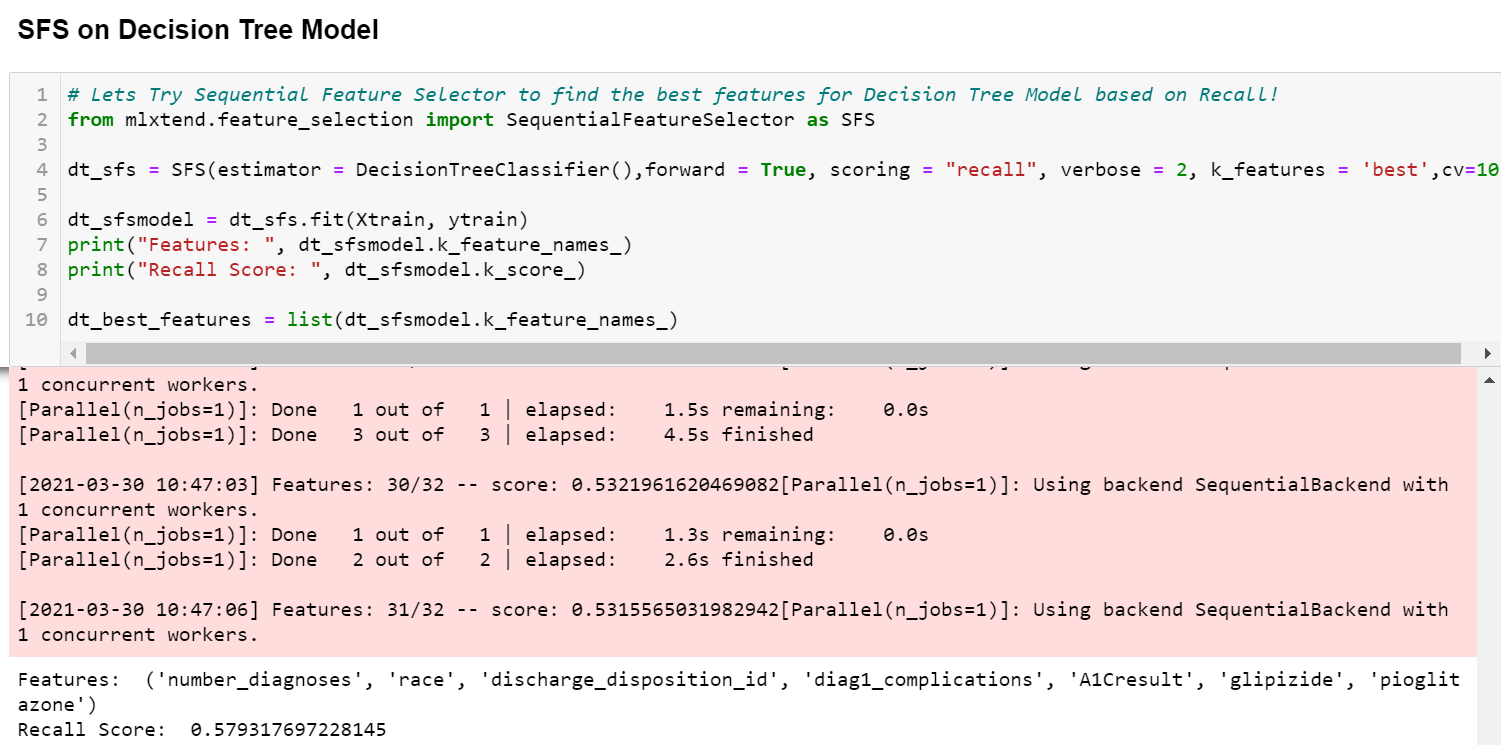
1. Implementing Boosting Algorithms like “**Ada Boost**”, “**Gradient Boosting**” and “**XGBoost**” to try and improve on key metrics.



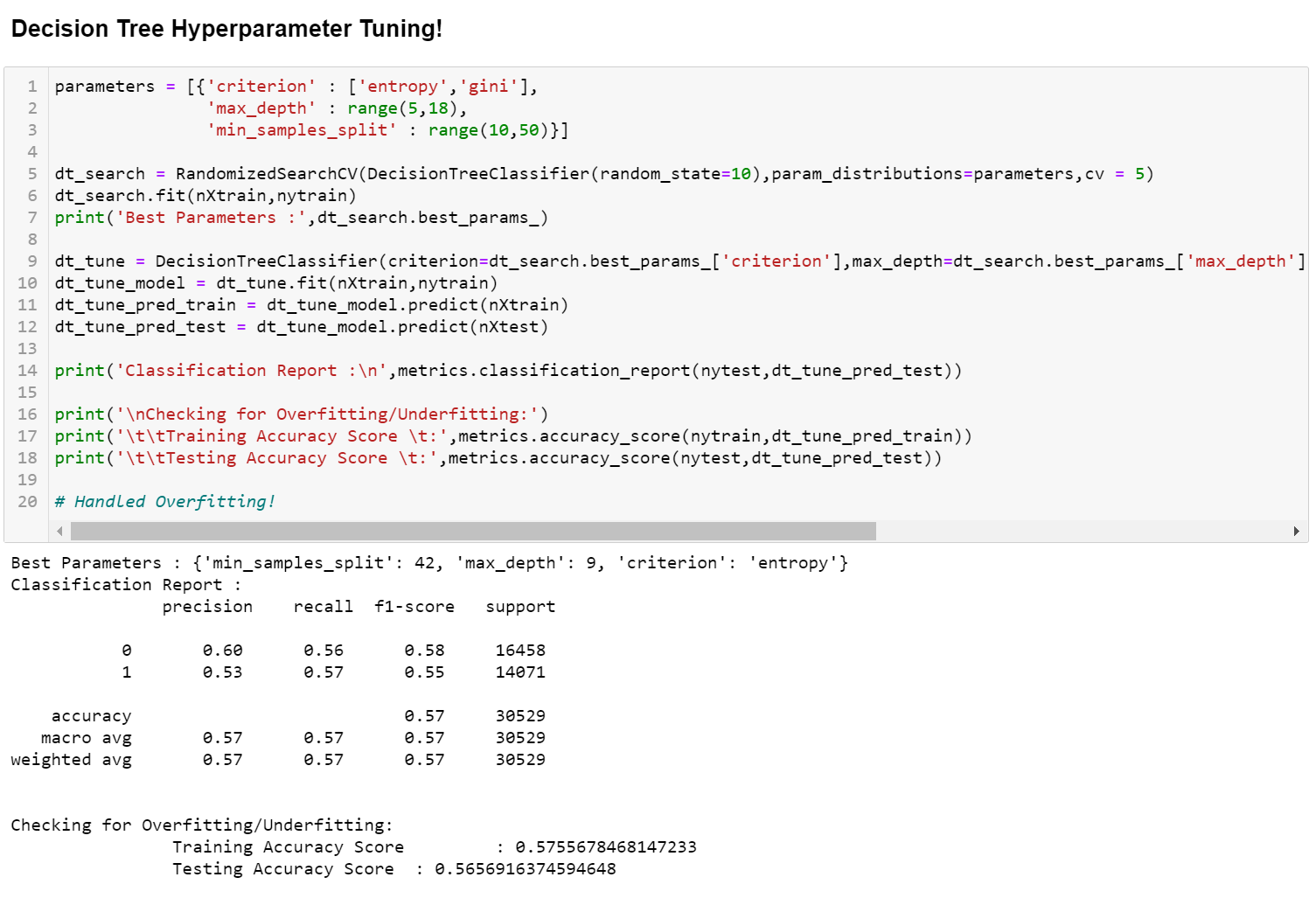


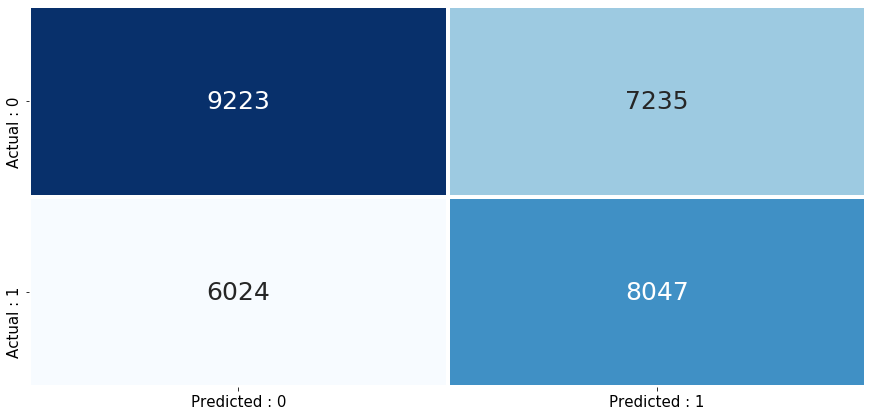


1. Implementing Feature Selection Techniques on the Data, such as **SFS** (Sequential Feature Selection) to further optimise the key metrics.



1. Conducting Hyper Parameter Tuning on the data with “**Randomized Search CV**” to get the best parameters for the models and deploying the tuned models.





# REFERENCE Notes for Project Team:

# 

|  |  |
| --- | --- |
| Original owner of data | Center for Clinical and Translational Research, Virginia Commonwealth University |
| Data set information | The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. |
| Any past relevant articles using the dataset | Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records |
| Reference | <https://www.hindawi.com/journals/bmri/2014/781670/> |
| Link to web page | <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008> |

### 

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*