greatlearning

“Effective Treatment for Diabetes”

Submitted in Partial Fulfilment of requirements for the Award of certificate of

Post Graduate Program in Data Science and Engineering

Capstone Project Report

Submitted to

GREAT LAKES
INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

Submitted By

Amit

Gaurav Dewangan

Mohammed Salman

Thiviya SK

Tharmesh VR

Under the Guidance Of

Mr. Dipanjan Goswami

Batch- PGPDSE March-2019

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 DIABETES

Diabetes is a type of chronic disease that occurs when your blood glucose, also known as blood sugar, is very high. Glucose in blood is your main source of energy and it comes from the food that you eat. To make glucose as source of energy, hormone called Insulin plays vital role by helping glucose to enter into cells to use for energy. Sometimes, human body don't produce or use insulin well that results in increasing glucose level in blood. However, high glucose level in blood can lead to health problem. Though diabetes has no cure, we can manage glucose level using insulin and other effective medication. Around the world, millions of people live with diabetes and the count is growing at considerable growth rate. According to World Health Organization (WHO), about 422 million people are suffering from diabetes. Moreover, in 2012, approximately 1.5 million deaths were caused by diabetes.

1.2 AIM OF THE PROJECT

After diagnosis of diabetes, effective treatment for diabetes is major concern for physician as well as patients. In such conditions we are trying to prescribe the more effective treatment possible to the patients based on the database collected from 130 hospitals over the period of 10 years which includes information of patients and their treatment.

Using this information, we are predicting the effectiveness of solo insulin or others drugs treatment for patients diagnosed with diabetes.

1.3 PROBLEM STATEMENT

The hospitals are evaluating efficiency of Insulin based treatment for patients. Recommend if solo insulin treatments work well towards the above stated objective. For a new patient, given his medical history and characteristics, should we recommend solo insulin or a conjunction of other drugs/ treatment?

CHAPTER 2

DATASET DESCRIPTION

2.1 DATA SET

The dataset, “Diabetes 130-US hospitals for years 1999-2008 Data Set” was obtained from the UCI Repository. The data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data. This data is a de-identified abstract of the Health Facts database (Cerner Corporation, Kansas City, MO).

2.2 VARIABLES CONSIDERED FOR ANALYSIS

The dataset comprises of 101766 record with 50 features. Out of 50 features, 28 are drugs molecule which combine to from target variable name treatment.

Below is the detailed description of each of the variables.

VARIABLE NAMES	DATA TYPE	DESCRIPTION
Encounter ID	Integer	Unique identifier of an encounter
Patient number	Integer	Unique identifier of a patient
Race Values	Object	Caucasian, Asian, African American, Hispanic, and other
Gender Values	Object	male, female, and unknown/invalid
Age	Object	Grouped in 10-year intervals (0-10), (10-20), …,(90-100)
Weight	Object	Weight in pounds
Admission type	Integer	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
Time in hospital	Integer	Integer number of days between admission and discharge
Discharge disposition	Integer	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
Discharge disposition	Integer	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
Admission source	Integer	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
Payer code	Object	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical
Medical specialty	Object	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	Integer	Number of lab tests performed during the encounter
Number of procedures	Integer	Numeric Number of procedures (other than lab tests) performed during the encounter

Number of medications	Integer	Number of distinct generic names administered during the encounter
Number of outpatient visits	Integer	Number of outpatient visits of the patient in the year preceding the encounter
Number of emergency visits	Integer	Number of emergency visits of the patient in the year preceding the encounter
Number of inpatient visits	Integer	Number of inpatient visits of the patient in the year preceding the encounter
Diagnosis 1	Object	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
Diagnosis 2	Object	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
Diagnosis 3	Object	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
Number of diagnoses	Integer	Number of diagnoses entered to the system 0%
Glucose serum test result	Object	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	Object	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	Object	Indicates if there was a change in diabetic medications (either dosage or generic name). Values “change” and “no change”
Diabetes medications	Object	Indicates if there was any diabetic medication prescribed. Values “yes” and “no”
24 features of medications in the form of drug 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	Object	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

2.3 TARGET VARIABLE

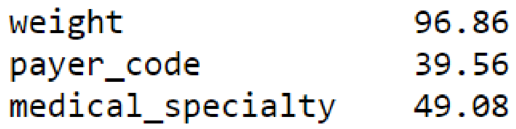
The target variable is treatment. This is a new feature created based on the medication given to each patient. In treatment, the unique values are ‘insulin’- solo insulin, ‘io’- insulin along with other drugs, ‘other’- combination of other drugs without insulin, ‘no_med’- did not receive any medication. But with respect to our we don’t consider records with ’no_med’ since our problem statement is to find the effective treatment for a given diabetes patient. This helps us to measure efficiency of solo insulin and insulin with other drugs and other drugs without insulin.

CHAPTER 3

DATA CLEANING

3.1 MISSING VALUE TREATMENT

The columns **‘weight’, ’payer_code, ’medical_speciality’ have more than 40 % missing values**. In normal cases we usually drop these variables since we have a major portion of the data missing and if we impute it would create more noise in the data.



```
weight               96.86
payer_code           39.56
medical_specialty    49.08
```

We can observe that the ‘**Race**’ Feature also has some missing values**.** We impute the missing value using MODE for Race Feature as most of the people in the Dataset are ‘Caucasian’.

3.2 FEATURE ENGINEERING

We create a feature called ‘**treatments**’. With respect to our problem statement we can find that the patients are given a solo-insulin treatment or combination of drugs treatment. There is a total of 23 different drugs given to the patients. In treatments, the unique values are:

- ‘insulin’- solo insulin,
- ‘io’- insulin along with other drugs,
- ‘other’- combination of other drugs without insulin,
- ‘no_med’- did not receive any medication.

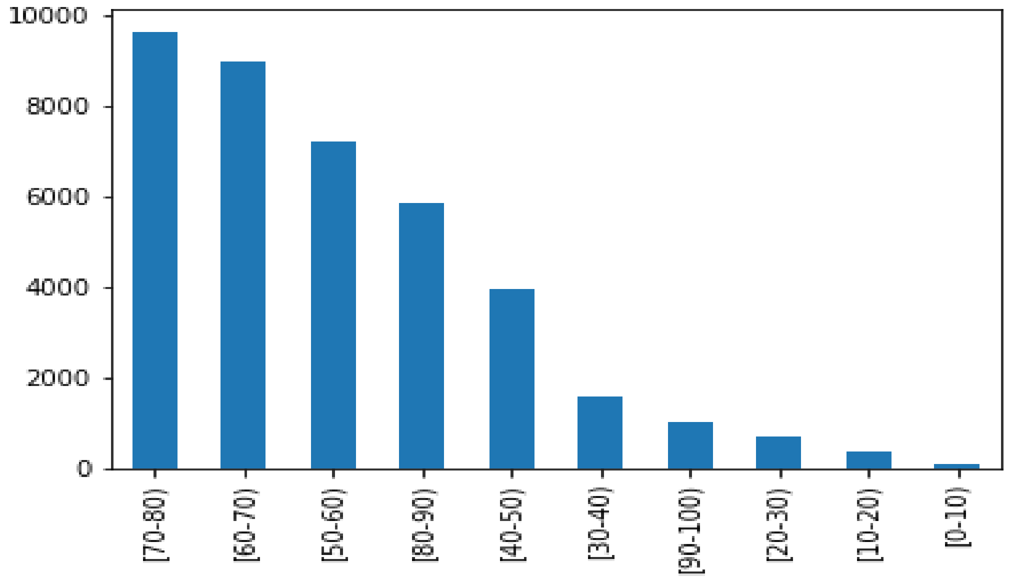
3.3OUTLIER TREATMENTS

Since the data is medical in nature, outlier treatment is not advisable as each data point is significant and we need to include all of them in our analysis.

CHAPTER 4

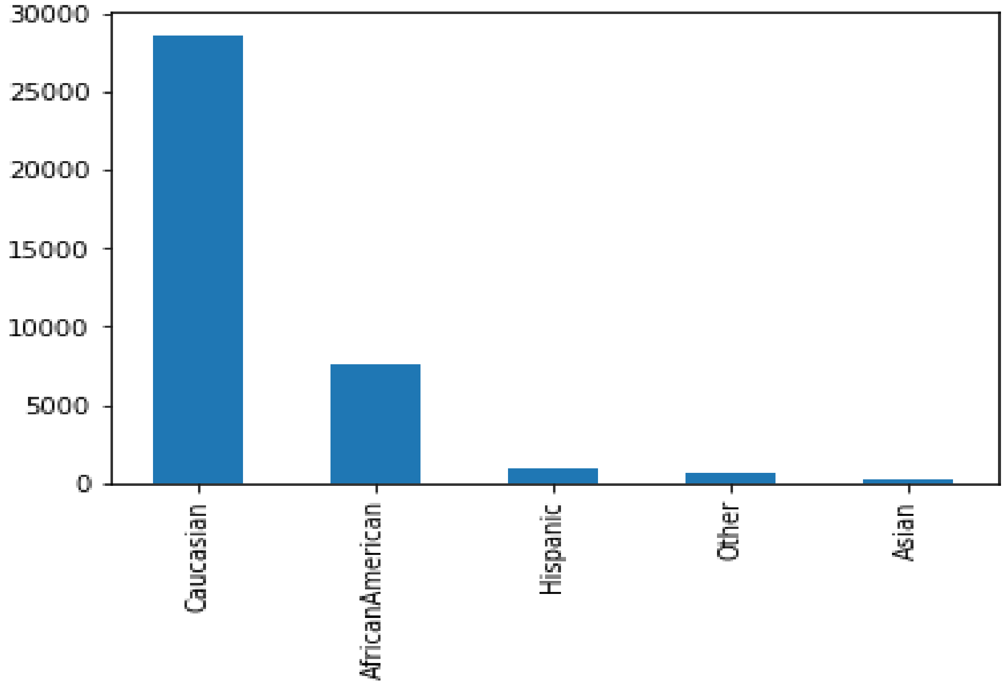
EXPLORATORY DATA ANALYSIS

4.1 DISTRIBUTION OF AGE



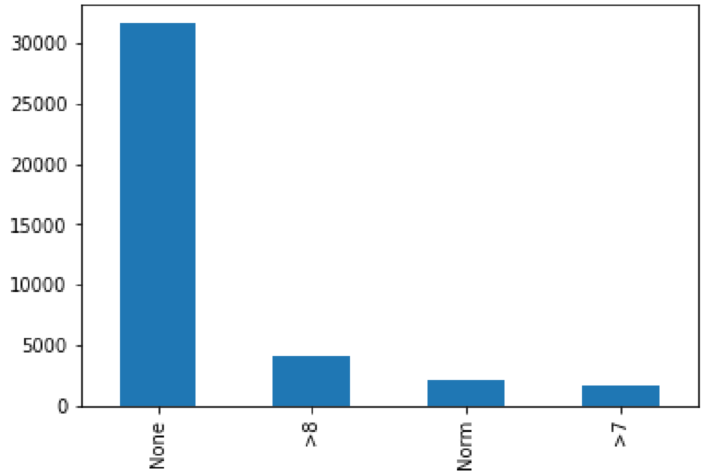
It can be inferred from the above plot that most of the diabetes are in the age groups [50-80].

4.2 DISTRIBUTION OF RACE



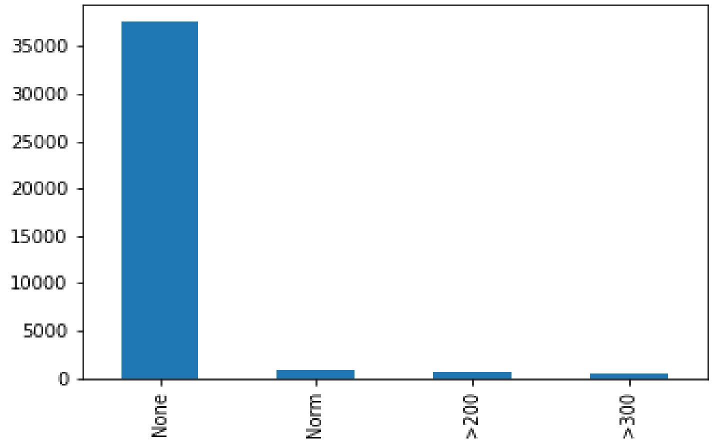
Most of the patients are of the Caucasian race.

4.3 DISTRIBUTION OF A1C RESULTS



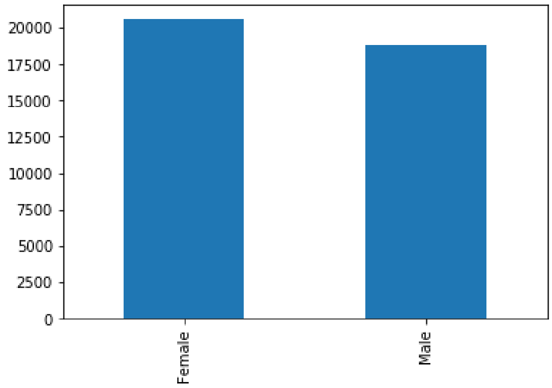
It can be inferred that most of the patients did not undergo the A1C test as most of the values are none.

4.4 DISTRIBUTION OF GLUCOSE SERUM VALUES



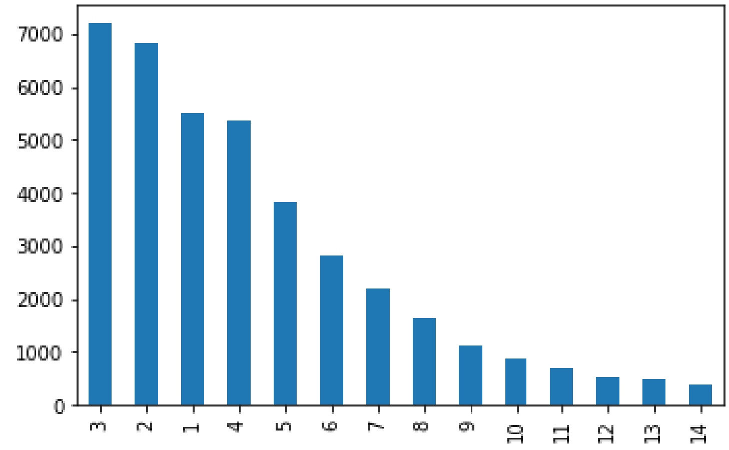
The Glucose serum values of most of the diabetes are also unknown as most of the values are none.

4.5 DISTRIBUTION OF GENDER



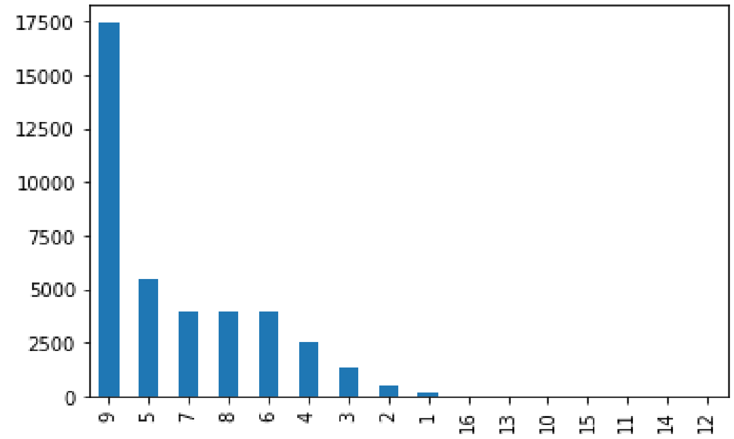
Most of the diabetes patients are female.

4.6 TIME IN HOSPITAL



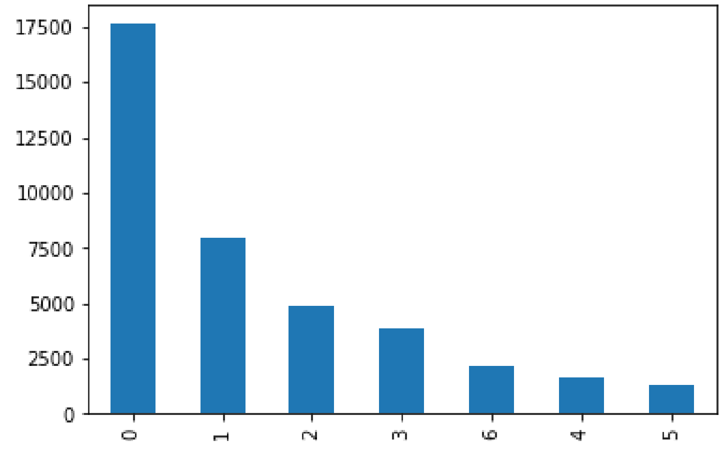
Most patients spend around 1 to 4 days in the hospital.

4.7 NUMBER OF DIAGNOSES



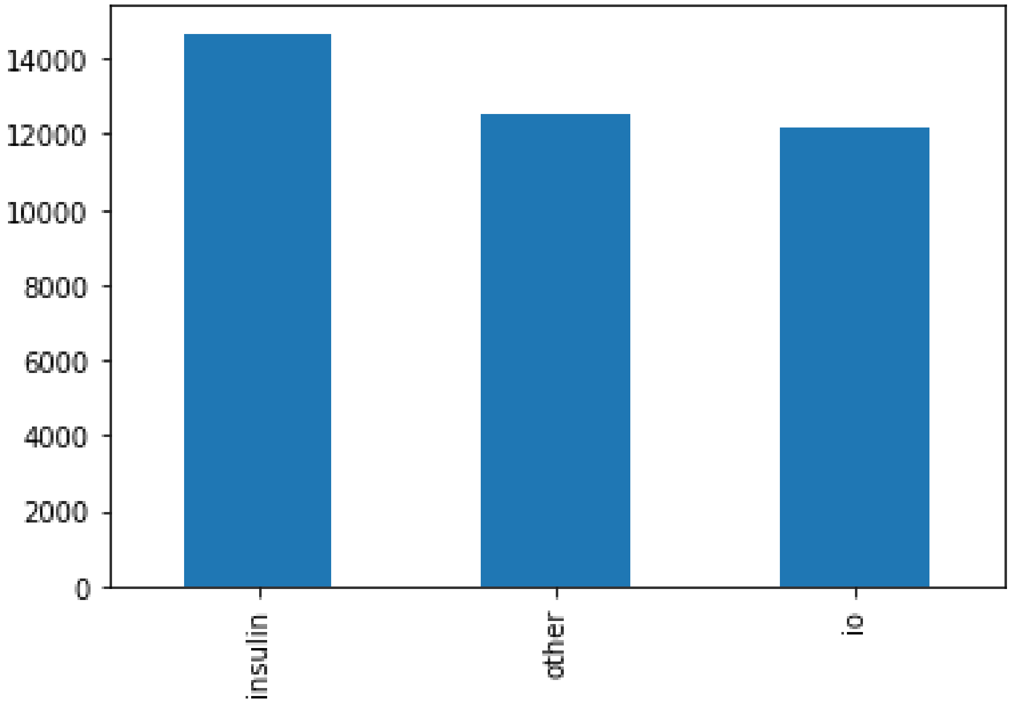
Most patients have had nine diagnoses.

4.8 NUMBER OF PROCEDURES



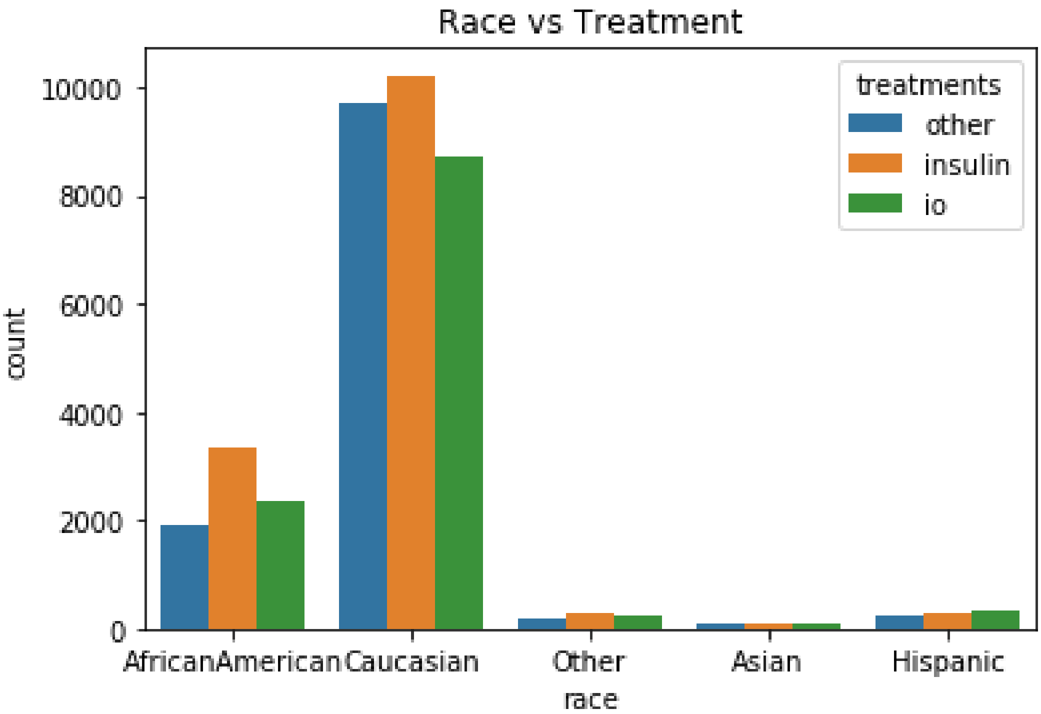
Most of the patients have not undergone any procedures

4.9 DISTRIBUTION OF TREATMENTS



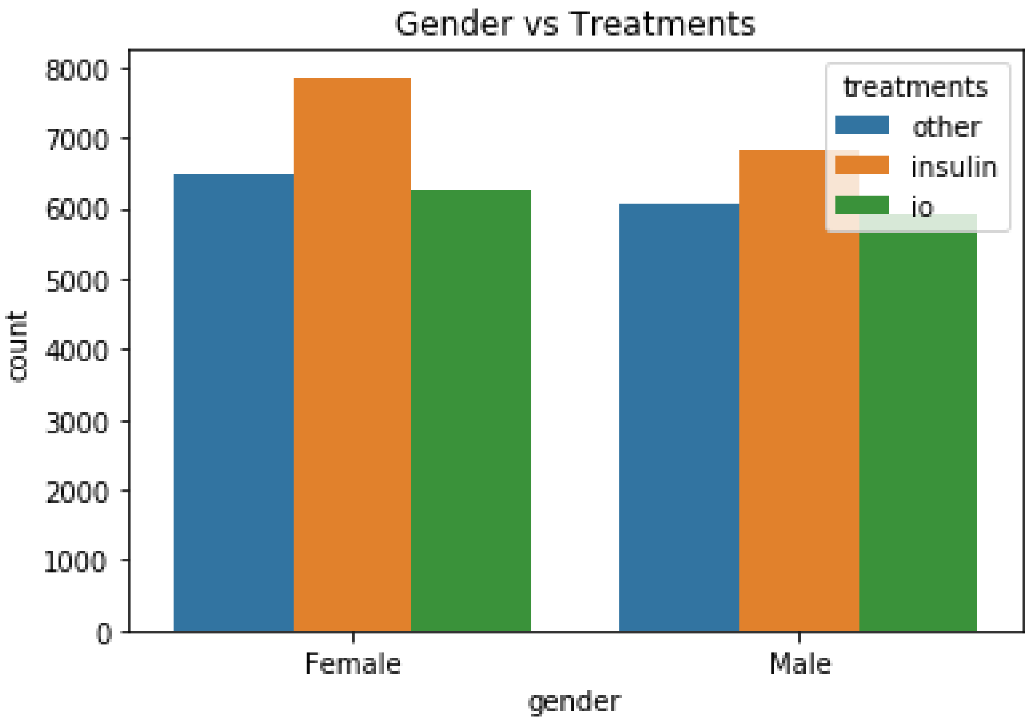
Most patients have received solo insulin treatments.

4.10 DISTRIBUTION OF TREATMENTS AMONG RACE



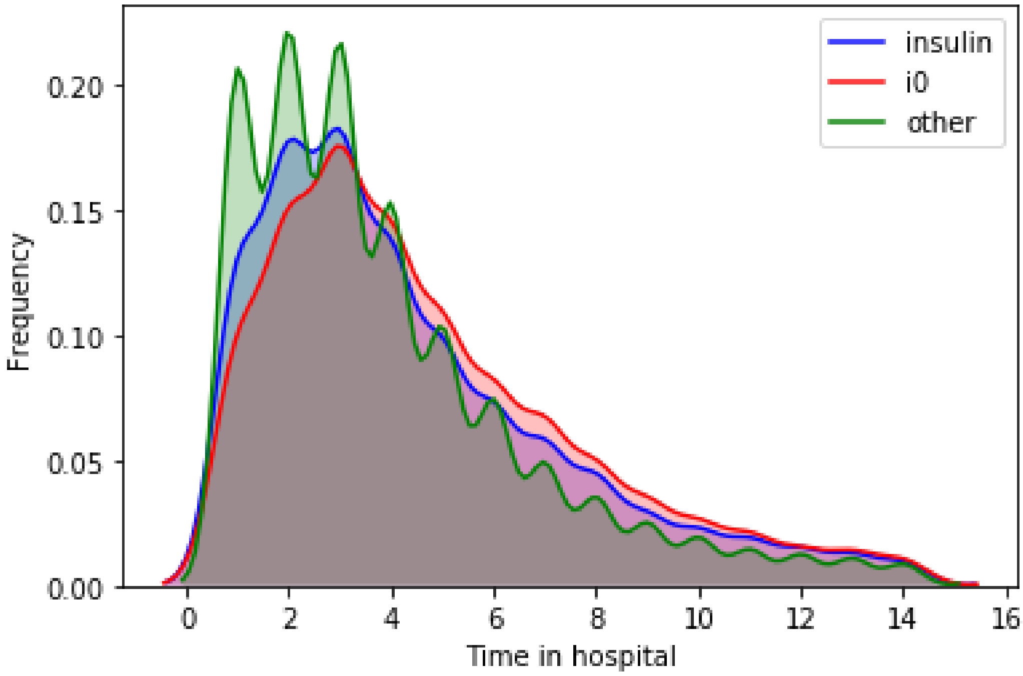
The above graph helps to conclude that race is independent to type of treatment, since in all given races the insulin treatment is contributes the major share.

4.11 GENDER VS TREATMENT



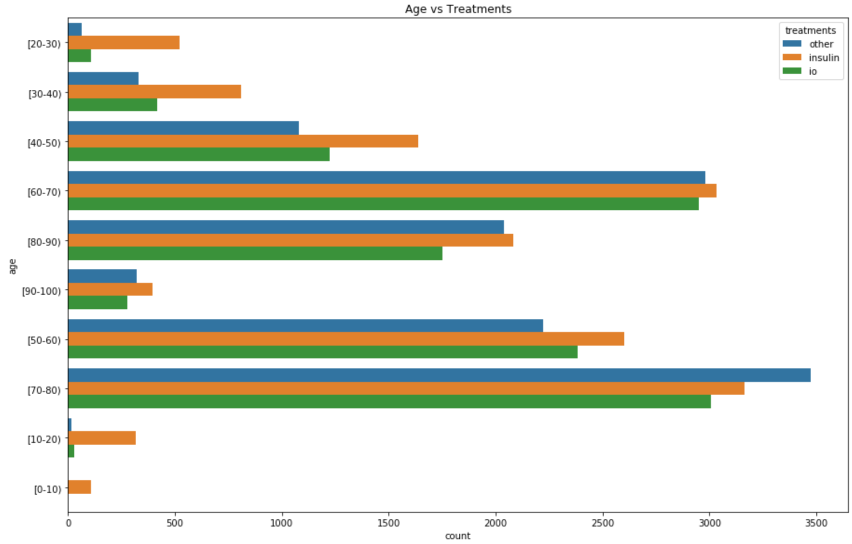
The gender of the patient doesn't matter, whether male or female insulin proved to be more effective than conjunction treatment.

4.12 TIME IN HOSPITAL VS TREATMENTS



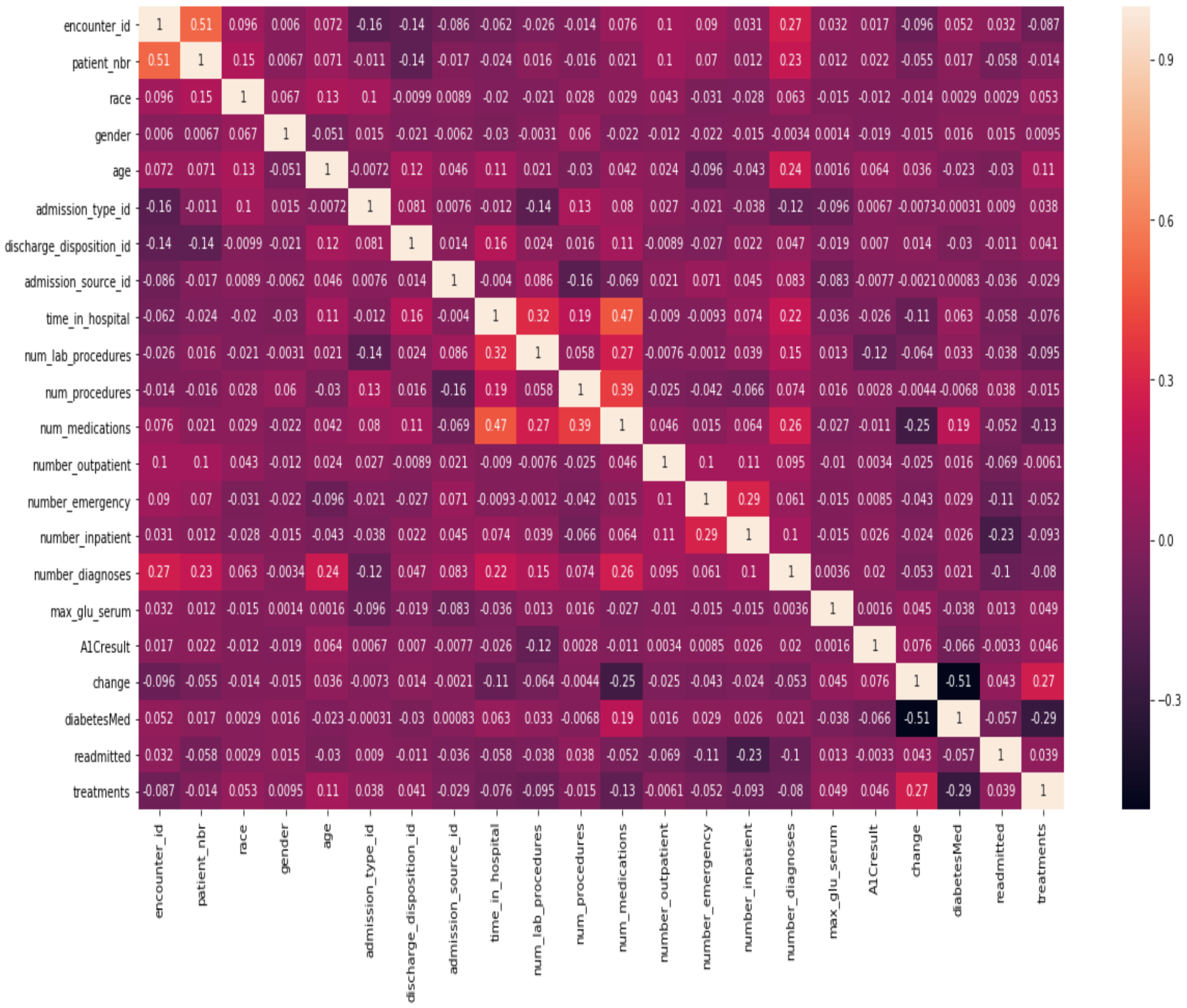
Patients who are other taking drug combinations are spending less time in hospital when compared t o the patients taking combination treatment.

4.13 AGE VS TREATMENTS



Majority of the patients suffering from diabetes is comes from age between 60 to 90.

4.14 CORRELATION AMONG GIVEN VARIABLES



	encounter_id	patient_nbr	race	gender	age	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses	max_glu_serum	A1Cresult	change	diabetesMed	readmitted	treatments
encounter_id	1	0.51	0.096	0.006	0.072	-0.16	-0.14	-0.086	-0.062	-0.026	-0.014	0.076	0.1	0.09	0.031	0.27	0.032	0.017	-0.096	0.052	0.032	-0.087
patient_nbr	0.51	1	0.15	0.0067	0.071	-0.011	-0.14	-0.017	-0.024	0.016	-0.016	0.021	0.1	0.07	0.012	0.23	0.012	0.022	-0.055	0.017	-0.058	-0.014
race	0.096	0.15	1	0.067	0.13	0.1	-0.0099	0.0089	-0.02	-0.021	0.028	0.029	0.043	-0.031	-0.028	0.063	-0.015	-0.012	-0.014	0.0029	0.0029	0.053
gender	0.006	0.0067	0.067	1	-0.051	0.015	-0.021	-0.0062	-0.03	-0.0031	0.06	-0.022	-0.012	-0.022	-0.015	-0.0034	0.0014	-0.019	-0.015	0.016	0.015	0.0095
age	0.072	0.071	0.13	-0.051	1	-0.0072	0.12	0.046	0.11	0.021	-0.03	0.042	0.024	-0.096	-0.043	0.24	0.0016	0.064	0.036	-0.023	-0.03	0.11
admission_type_id	-0.16	-0.011	0.1	0.015	-0.0072	1	0.081	0.0076	-0.012	-0.14	0.13	0.08	0.027	-0.021	-0.038	-0.12	-0.096	0.0067	-0.0073	-0.00031	0.009	0.038
discharge_disposition_id	-0.14	-0.14	-0.0099	-0.021	0.12	0.081	1	0.014	0.16	0.024	0.016	0.11	-0.0089	-0.027	0.022	0.047	-0.019	0.007	0.014	-0.03	-0.011	0.041
admission_source_id	-0.086	-0.017	0.0089	-0.0062	0.046	0.0076	0.014	1	-0.004	0.086	-0.16	-0.069	0.021	0.071	0.045	0.083	-0.083	-0.0077	-0.0021	0.00083	-0.036	-0.029
time_in_hospital	-0.062	-0.024	-0.02	-0.03	0.11	-0.012	0.16	-0.004	1	0.32	0.19	0.47	-0.009	-0.0093	0.074	0.22	-0.036	-0.026	-0.11	0.063	-0.058	-0.076
num_lab_procedures	-0.026	0.016	-0.021	-0.0031	0.021	-0.14	0.024	0.086	0.32	1	0.058	0.27	-0.0076	-0.0012	0.039	0.15	0.013	-0.12	-0.064	0.033	-0.038	-0.095
num_procedures	-0.014	-0.016	0.028	0.06	-0.03	0.13	0.016	-0.16	0.19	0.058	1	0.39	-0.025	-0.042	-0.066	0.074	0.016	0.0028	-0.0044	-0.0068	0.038	-0.015
num_medications	0.076	0.021	0.029	-0.022	0.042	0.08	0.11	-0.069	0.47	0.27	0.39	1	0.046	0.015	0.064	0.26	-0.027	-0.011	-0.25	0.19	-0.052	-0.13
number_outpatient	0.1	0.1	0.043	-0.012	0.024	0.027	-0.0089	0.021	-0.009	-0.0076	-0.025	0.046	1	0.1	0.11	0.095	-0.01	0.0034	-0.025	0.016	-0.069	-0.0061
number_emergency	0.09	0.07	-0.031	-0.022	-0.096	-0.021	-0.027	0.071	-0.0093	-0.0012	-0.042	0.015	0.1	1	0.29	0.061	-0.015	0.0085	-0.043	0.029	-0.11	-0.052
number_inpatient	0.031	0.012	-0.028	-0.015	-0.043	-0.038	0.022	0.045	0.074	0.039	-0.066	0.064	0.11	0.29	1	0.1	-0.015	0.026	-0.024	0.026	-0.23	-0.093
number_diagnoses	0.27	0.23	0.063	-0.0034	0.24	-0.12	0.047	0.083	0.22	0.15	0.074	0.26	0.095	0.061	0.1	1	0.0036	0.02	-0.053	0.021	-0.1	-0.08
max_glu_serum	0.032	0.012	-0.015	0.0014	0.0016	-0.096	-0.019	-0.083	-0.036	0.013	0.016	-0.027	-0.01	-0.015	-0.015	0.0036	1	0.0016	0.045	-0.038	0.013	0.049
A1Cresult	0.017	0.022	-0.012	-0.019	0.064	0.0067	0.007	-0.0077	-0.026	-0.12	0.0028	-0.011	0.0034	0.0085	0.026	0.02	0.0016	1	0.076	-0.066	-0.0033	0.046
change	-0.096	-0.055	-0.014	-0.015	0.036	-0.0073	0.014	-0.0021	-0.11	-0.064	-0.0044	-0.25	-0.025	-0.043	-0.024	-0.053	0.045	0.076	1	-0.51	0.043	0.27
diabetesMed	0.052	0.017	0.0029	0.016	-0.023	-0.00031	-0.03	0.00083	0.063	0.033	-0.0068	0.19	0.016	0.029	0.026	0.021	-0.038	-0.066	-0.51	1	-0.057	-0.29
readmitted	0.032	-0.058	0.0029	0.015	-0.03	0.009	-0.011	-0.036	-0.058	-0.038	0.038	-0.052	-0.069	-0.11	-0.23	-0.1	0.013	-0.0033	0.043	-0.057	1	0.039
treatments	-0.087	-0.014	0.053	0.0095	0.11	0.038	0.041	-0.029	-0.076	-0.095	-0.015	-0.13	-0.0061	-0.052	-0.093	-0.08	0.049	0.046	0.27	-0.29	0.039	1

There is low multi collinearity between independent features.

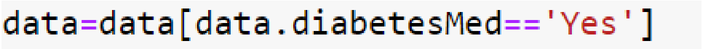
CHAPTER 5

DATA PREPARING

In this project we are trying to create a model that recommends the effective treatment for a given diabetes patient. So, it is necessary that we do our analysis only on diabetes patients who have received effective treatments.

5.1 FILTERING DIABETES PATIENTS

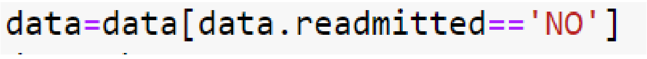
We can filter out diabetes patients using the '**diabetesMed**' feature. We consider only those records for which 'diabetesMed'='Yes'.



```
data=data[data.diabetesMed=='Yes']
```

FILTERING EFFECTIVE TREATMENTS

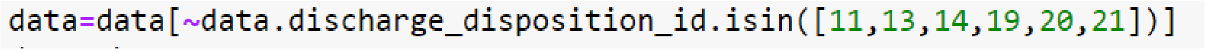
We filter out patients all those who received effective treatments which can be identified using the 'readmitted' column. Hence, we further filter the records and consider only those records for which 'readmitted'='NO'.



```
data=data[data.readmitted=='NO']
```

5.2 REMOVING DEAD PATIENTS

We also exclude patients who are dead and are in hospice condition (patients whose condition is such that a doctor would not be surprised if the patient died within the next six months). It is done by removing records for which 'discharge_disposition_id'= 11,13,14,19,20,21.



```
data=data[~data.discharge_disposition_id.isin([11,13,14,19,20,21])]
```

ENCODING CATEGORICAL VARIABLES

We encode the nominal categorical values using one hot encoder.

5.3 FEATURE SELECTION

With respect to the problem statement given, the output variable is observed to be the "treatments" feature.The input variables are both Discrete Quantitative and Categorical and our output variable is Categorical. Since we have a combination of Discrete Quantitative Variables and Categorical Variables, we cannot perform general Correlation test. So, we must perform Chi-Square test of independence to find the important features.

5.4 CHI-SQUARE TEST FOR INDEPENDENCE

A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each other.

In our case, we need to determine whether there is indeed a relationship between a predictor variable and any of the target variables to a significant degree. We only need to consider these features for our further analysis.

The null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables being tested. I.e. they are independent.

The p-value will tell us if our test results are significant or not. In order to perform a chi square test and get the p-value, you need two pieces of information:

- Degrees of freedom. That's just the number of categories minus 1.
- The alpha level(α). The usual alpha or significance level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

We reject the null hypothesis when the P-value is less than the set significance level.

These are the attributes we got as significant from this chi-square method:

- age
- admission_type_id
- discharge_disposition_id
- admission_source_id
- time_in_hospital
- num_lab_procedures
- num_procedures
- num_medications
- number_outpatient
- number_emergency
- number_inpatient
- number_diagnoses
- race_AfricanAmerican
- race_Asian
- race_Caucasian
- race_Hispanic
- gender_Female
- gender_Male
- max_glu_serum_>200
- max_glu_serum_>300
- max_glu_serum_None
- max_glu_serum_Norm
- A1Cresult_>7
- A1Cresult_>8
- A1Cresult_None
- A1Cresult_Norm
- change_Ch
- change_No
- level1_diag1
- level2_diag1
- level1_diag2
- level2_diag2
- level1_diag3
- level2_diag3

CHAPTER 6

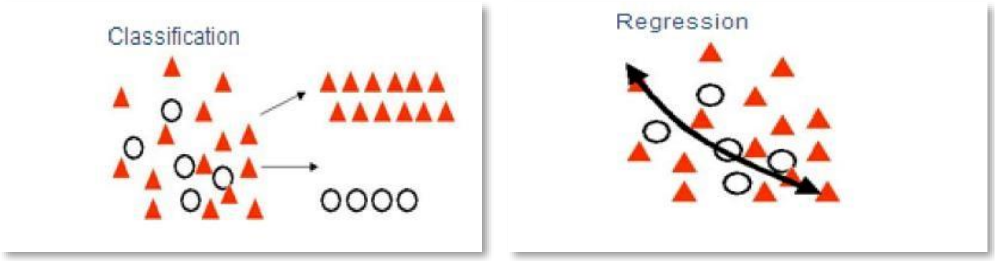
ALGORITHMS USED

6.1 DECISION TREE (CART)

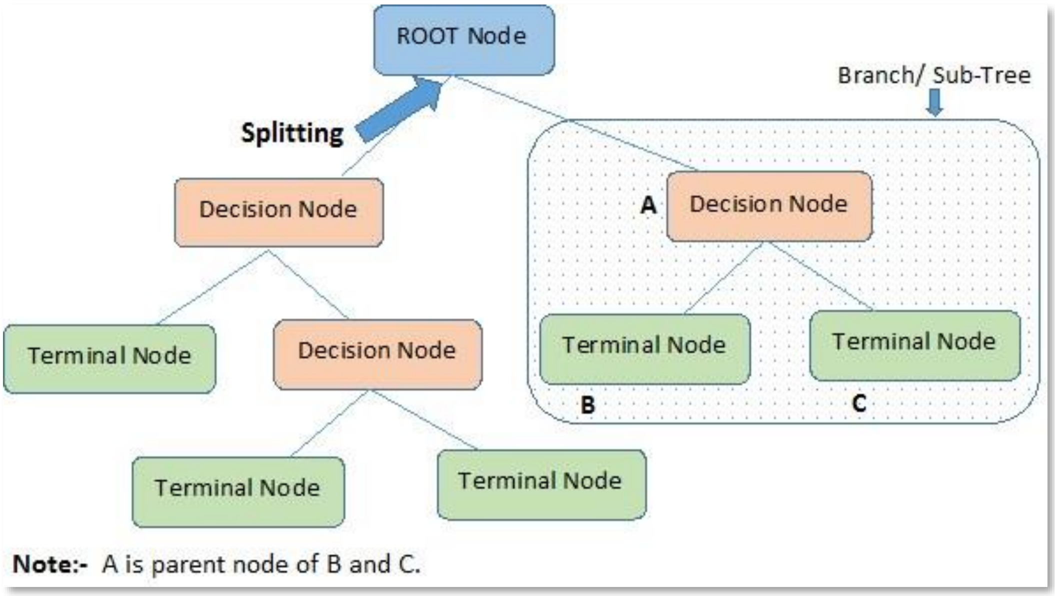
A Decision tree (CART) is a schematic, tree-shaped diagram used to determine a course of action or show a statistical probability. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision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

Types of Decision Tree

- **Classification Trees:** where the Dependent variable is categorical and the tree is used to identify the "class" within which a Dependent variable would likely fall into.
- **Regression Trees:** where the Dependent variable is continuous and tree is used to predict its value. (e.g. the price of a house, or a patient's length of stay in a hospital).



Layout / flow of Decision Tree



Advantages of CART

- Simple to understand, interpret, visualize.
- Decision trees implicitly perform variable screening or feature selection.
- Can handle both numerical and categorical data. Can also handle multi-output problems.
- Decision trees require relatively little effort from users for data preparation.
- Nonlinear relationships between parameters do not affect tree performance.

Disadvantages of CART

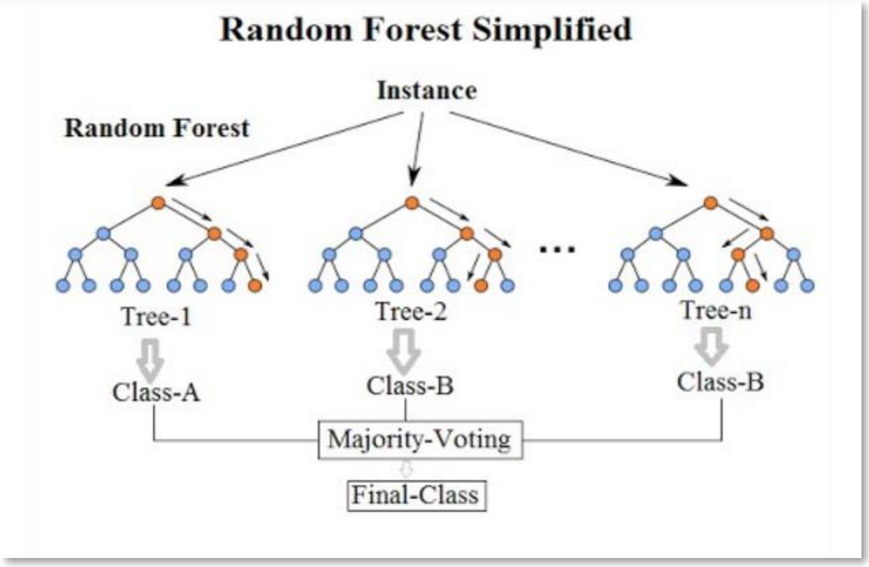
- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting.
- Greedy algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement.
- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree.

6.2 RANDOM FOREST

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

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

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



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

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

Advantages of Random Forest

- There is no need for feature normalization
- Individual decision trees can be trained in parallel
- Reduced overfitting
- Require almost no input preparation
- Performs implicit feature selection
- It’s very quick to train

Disadvantages of Random Forest

- No interpretability

6.3 K NEAREST NEIGHBORS (KNN)

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

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

Advantages of KNN

- No Training Period
- Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm
- KNN is very easy to implement

Disadvantages of KNN

- Does not work well with large dataset
- Does not work well with high dimensions
- Need feature scaling
- Sensitive to noisy data, missing values and outliers
- We need to manually impute missing values and remove outliers

6.4 CATBOOST ALGORITHM

CatBoost is a recently open-sourced machine learning algorithm from Yandex. It can easily integrate with deep learning frameworks like Google's TensorFlow and Apple's Core ML. It can work with diverse data types to help solve a wide range of problems that businesses face today. To top it up, it provides best-in-class accuracy.

It is especially powerful in two ways:

- It yields state-of-the-art results without extensive data training typically required by other machine learning methods, and
- Provides powerful out-of-the-box support for the more descriptive data formats that accompany many business problems.

"CatBoost" name comes from two words "**Cat**egory" and "**boost**ing". As discussed, the library works well with multiple **Cat**egories of data, such as audio, text, image including historical data. **"Boost**" comes from gradient boosting machine learning algorithm as this library is based on gradient boosting library. Gradient boosting is a powerful machine learning algorithm that is widely applied to multiple types of business challenges like fraud detection, recommendation items, forecasting and it performs well also. It can also return very good result with relatively less data, unlike DL models that need to learn from a massive amount of data.

Advantages of CatBoost

- Performance
- Handling Categorical features automatically
- Robust
- Easy-to-use

Disadvantages of CatBoost

- High training and optimization times

CHAPTER 7

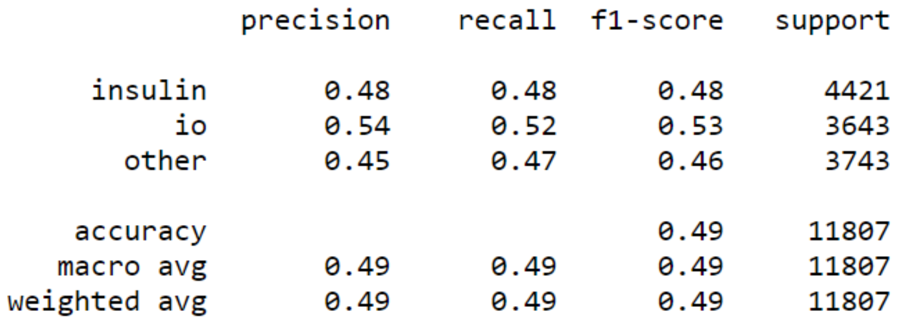
MODELLING AND RESULTS

Modelling using the listed algorithms was done and results were compiled for overview.

7.1 PREDICTIVE MODEL DEVELOPMENT - ITERATION 1

In this iteration the base models for each algorithms is built without any hyper parameter tuning.

DECISION TREE (CART):

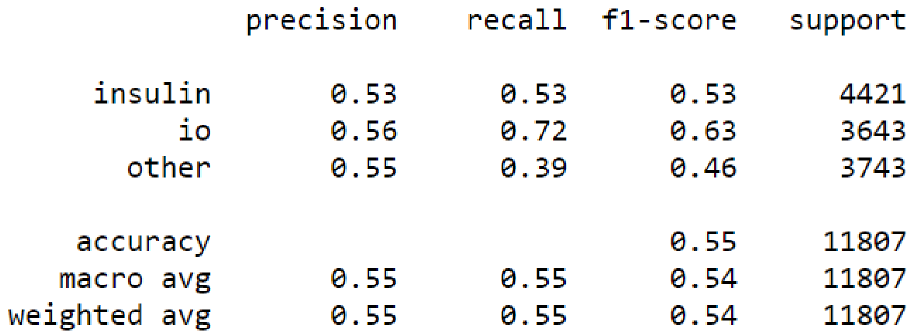


```
              precision    recall  f1-score   support

     insulin       0.48      0.48      0.48      4421
          io       0.54      0.52      0.53      3643
       other       0.45      0.47      0.46      3743

    accuracy                           0.49     11807
   macro avg       0.49      0.49      0.49     11807
weighted avg       0.49      0.49      0.49     11807
```

Training Accuracy: 1.000
Testing Accuracy: 0.487

INFERENCE: The model is not able to correctly identify the effective treatment for given patients at a reasonable rate. We can see that the model is overfitted as the training accuracy is 100%. In the next iteration we must prune the Decision Tree to get better predictions.

RANDOM FOREST:

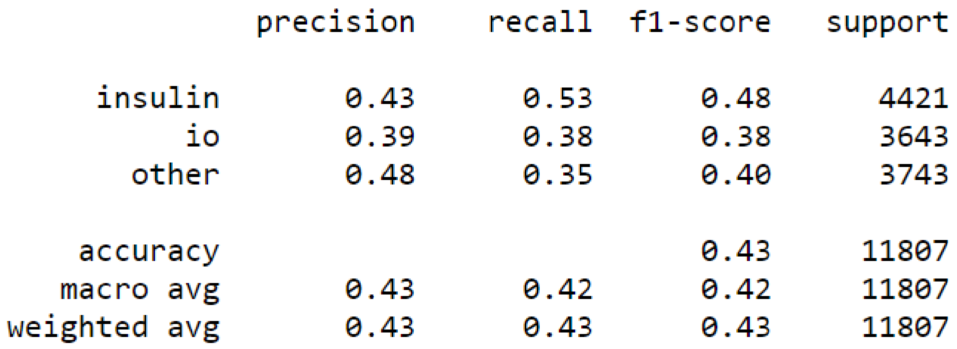


```
              precision    recall  f1-score   support

     insulin       0.53      0.53      0.53      4421
          io       0.56      0.72      0.63      3643
       other       0.55      0.39      0.46      3743

    accuracy                           0.55     11807
   macro avg       0.55      0.55      0.54     11807
weighted avg       0.55      0.55      0.54     11807
```

Training Accuracy: 0.985
Testing Accuracy: 0.545

INFERENCE: The model is not able to correctly identify the effective treatment for given patients at a reasonable rate. We can see that the model is overfitted as the training accuracy is 100%. In the next iteration we must optimise it.

KNN:



```
              precision    recall  f1-score   support

     insulin       0.43      0.53      0.48      4421
          io       0.39      0.38      0.38      3643
       other       0.48      0.35      0.40      3743

    accuracy                           0.43     11807
   macro avg       0.43      0.42      0.42     11807
weighted avg       0.43      0.43      0.43     11807
```

Training Accuracy: 0.620
Testing Accuracy: 0.428

INFERENCE: The model is not able to correctly identify the effective treatment for given patients at a reasonable rate. We can see that the model accuracy is very less.

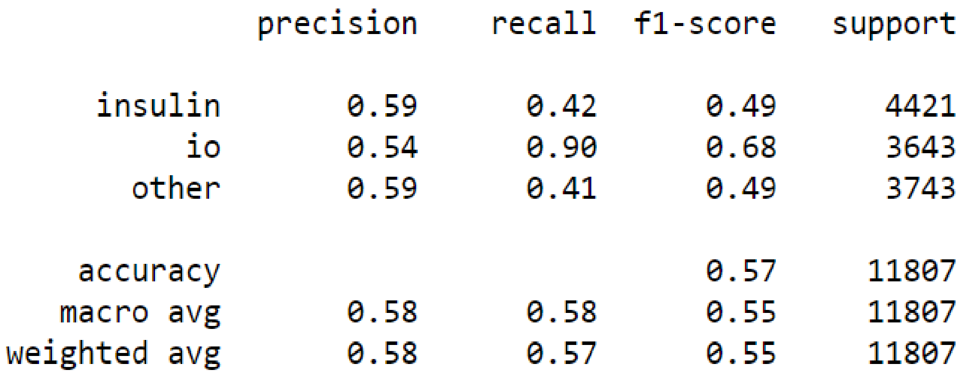
7.2 PREDICTIVE MODEL DEVELOPMENT - ITERATION 2

In this iteration the best parameters for each model is found using GridSearchCV.

TUNED DECISION TREE (CART):

Using GridSearchCV the best hyper parameters found were:

- 'criterion': 'entropy'
- 'max_depth': 10
- 'min_samples_leaf': 100
- 'min_samples_split': 50



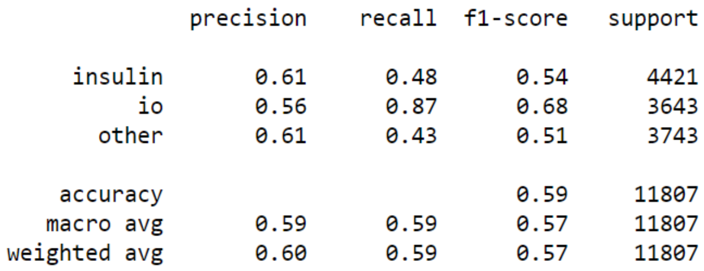
```
              precision    recall  f1-score   support

     insulin       0.59      0.42      0.49      4421
          io       0.54      0.90      0.68      3643
       other       0.59      0.41      0.49      3743

    accuracy                           0.57     11807
   macro avg       0.58      0.58      0.55     11807
weighted avg       0.58      0.57      0.55     11807
```

Training Accuracy: 0.590
Testing Accuracy: 0.566

INFERENCE: After pruning overfitting has been removed but the maximum accuracy that we could get is 57%.

TUNED RANDOM FOREST:

Using GridSearchCV the best hyper parameters found were: 'n_estimators': 96



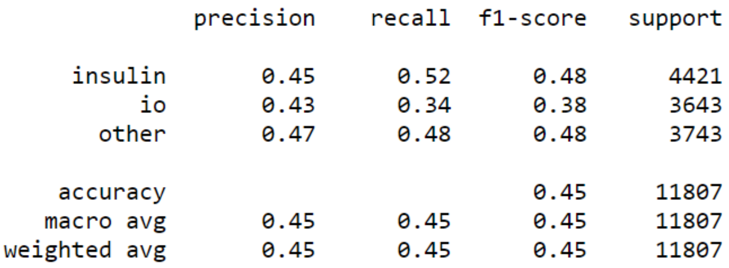
```
              precision    recall  f1-score   support

     insulin       0.61      0.48      0.54      4421
          io       0.56      0.87      0.68      3643
       other       0.61      0.43      0.51      3743

    accuracy                           0.59     11807
   macro avg       0.59      0.59      0.57     11807
weighted avg       0.60      0.59      0.57     11807
```

Training Accuracy: 1.000
Testing Accuracy: 0.585

INFERENCE: Even after tuning the hyper parameters we can see that the model is overfitted as the training accuracy is 100% and maximum accuracy that we could get is 59%

TUNED KNN:

Using GridSearchCV the best hyper parameters found were: 'n_neighbors': 39

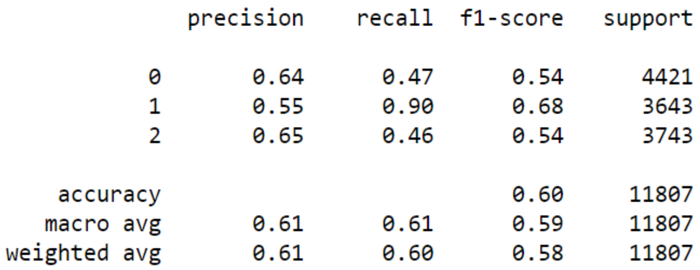


```
              precision    recall  f1-score   support

     insulin       0.45      0.52      0.48      4421
          io       0.43      0.34      0.38      3643
       other       0.47      0.48      0.48      3743

    accuracy                           0.45     11807
   macro avg       0.45      0.45      0.45     11807
weighted avg       0.45      0.45      0.45     11807
```

Training Accuracy: 0.503
Testing Accuracy: 0.451

INFERENCE: After tuning the hyper parameters the maximum accuracy that we could get is 45%.

CATBOOST:

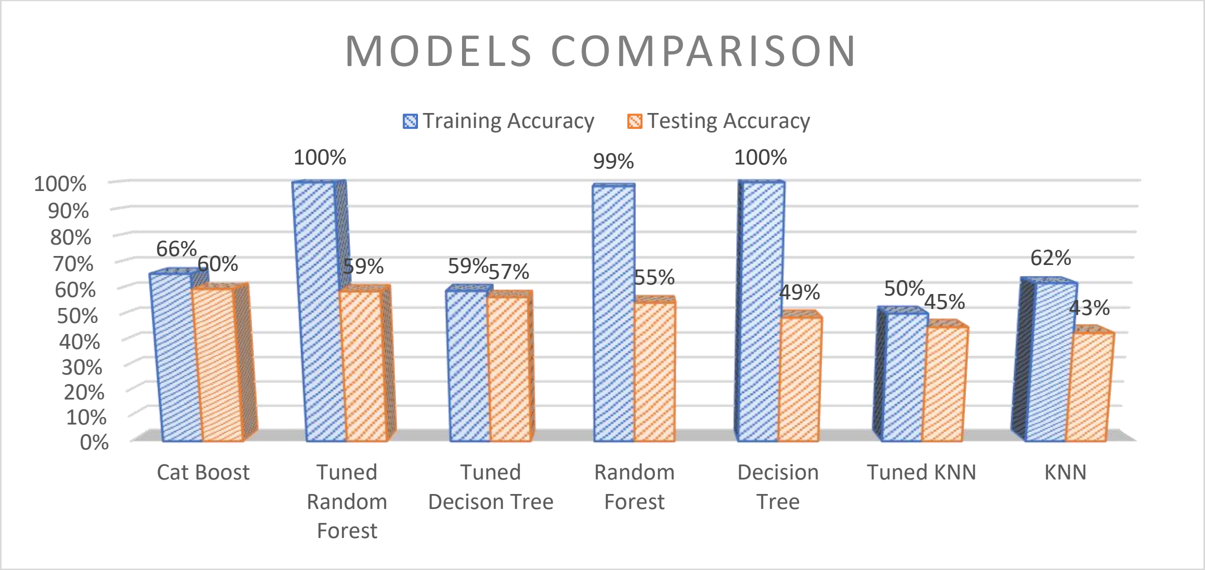


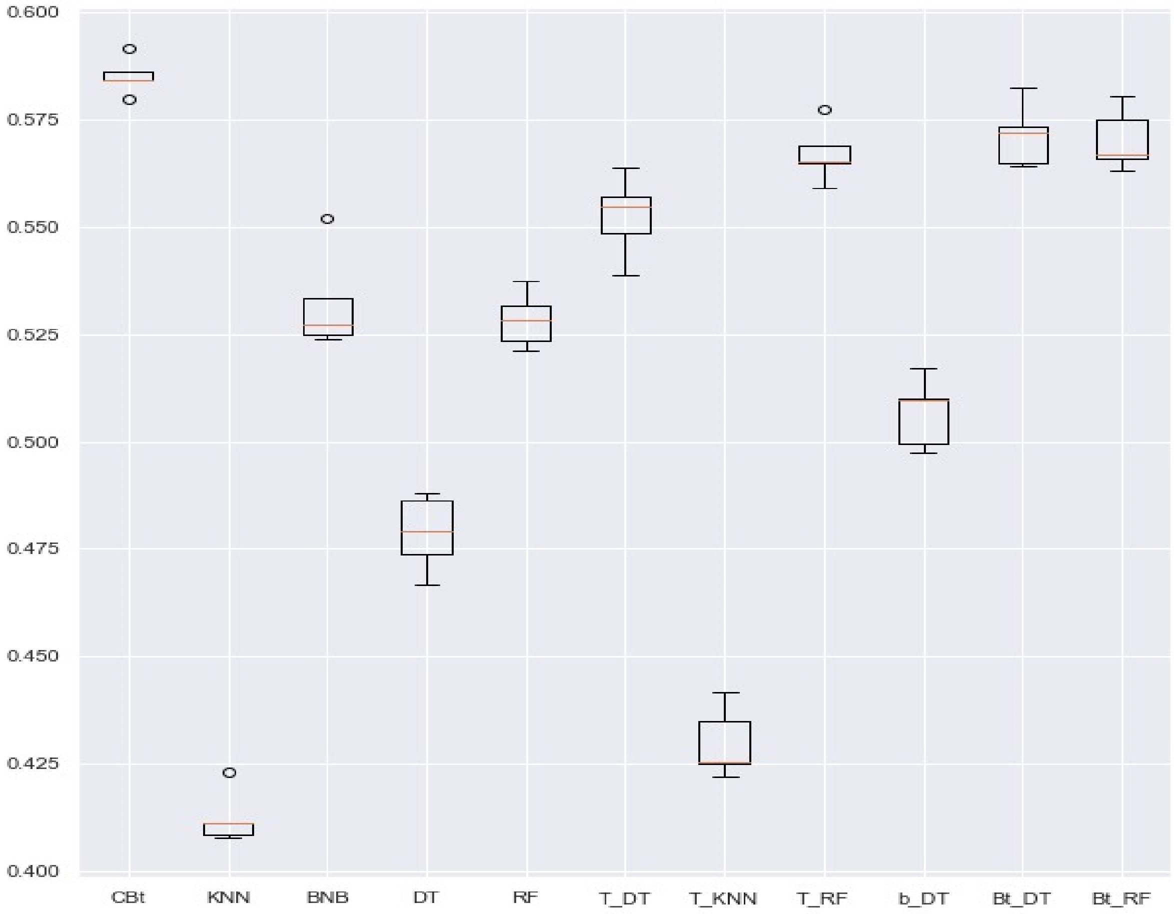
```
              precision    recall  f1-score   support

           0       0.64      0.47      0.54      4421
           1       0.55      0.90      0.68      3643
           2       0.65      0.46      0.54      3743

    accuracy                           0.60     11807
   macro avg       0.61      0.61      0.59     11807
weighted avg       0.61      0.60      0.58     11807
```

Test Accuracy:0.598
Train Accuracy: 0.664

INFERENCE: The best iteration gave a maximum accuracy of 60%.

7.3 MODELS COMPARISON





INFERENCE: We can infer from the above comparisons that CatBoost model gives us the best fit among the models and has a better bias-variance trade-off compared to the other models.

CHAPTER 8

CONCLUSION

- From the models built and the tests performed, CatBoost is the best method to predict the effective treatment for diabetes for patients from a Machine Learning perspective.

- But from our study, the original independent variables are not enough for the prediction of the treatments to a reasonable accuracy just by themselves as some of the important variables in relation to diabetes had no values.

- The patients can receive effective treatments only if they are diagnosed correctly at the first level of diagnosis. This diagnosis can be effectively done when the patients undergo important tests like HbA1C test and Glucose serum value tests. Also, the weight of the patients also needs to be recorded for better diagnosis and analysis.

- Thus, by recommending effective treatments the hospitals can reduce the readmission rates which can save them millions of dollars while also improving the quality of care.