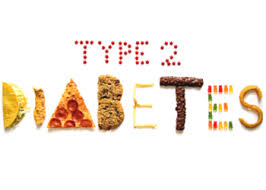
**DIABETES PREDICTION**



**Jjj**

Harshit Mehta(hm23388)

Sanjit Paliwal(Sp42626)

**Department of Operation Research and Industrial Engineering**

**University of Texas at Austin**

# TABLE OF CONTENTS

Objective .....................................................................................................................1

Description ..................................................................................................................1

Data Summary ............................................................................................................1

Data Analysis & Imputation …………......................................................................... 2

Comparison of Model Characteristics on original and imputed data …….................. 3

Total Cost Minimisation ………………….................................................................... 3

Total Cost and Lift Chart .................................…………………………………………. 4

Profit Chart ................................................................................................................ 5

Selecting the Best model for prediction ..................................................................... 5

Comparison of Model with Current Practice ……………………………………………. 6

Feature Addition ........................................................................................................ 6

Feature Selection using Wrapper Approach .............................................................. 7

Selecting the distinctive Characteristics .................................................................... 7

Decrease in area under ROC curve .......................................................................... 8

Summary …………………………………………………………………………………… 8

**Objective:**

The utmost aim of the project is to devise a method for selecting the patients for the intervention program so that the total cost savings are maximised or to minimise the total cost of the intervention program. Another task is to build a model which can predict the patient as diabetic or non-diabetic most accurately and to discuss the highly predictive attributes.

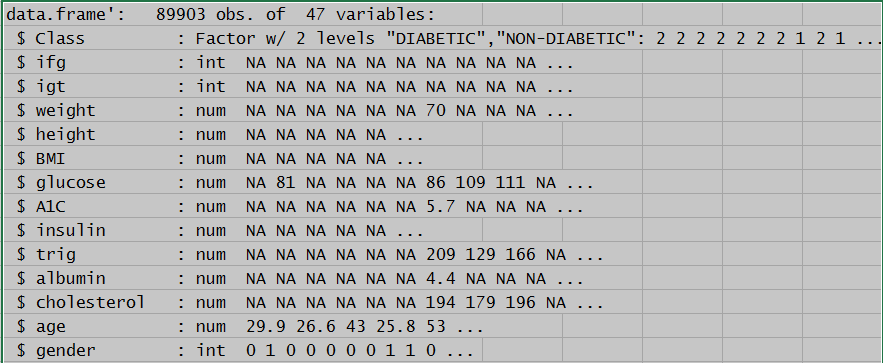
Description:

In Type II Diabetes, the Insulin hormones are not able to work effectively to convert the glucose we get from the food intake into energy. It is also known as Insulin Resistance. As the Type II Diabetes gets worse, the Pancreas makes less insulin and leads to Insulin deficiency. As per **National Diabetes Statistics Report 2014,** In USA, 86 Million people are prediabetic and among those 29.1 Million people are diabetic. 90% of the diabetic patients have Type II diabetes.

The Intervention Program is started by the Health care Organisation to reduce the likelihood of Type II Diabetes. Currently, The Organisation is selecting patients based on their Fasting Blood Glucose level for the Intervention Program. This Current practice of selecting Patients having FBC greater than 110 incurs total cost $**35,639,660.** Our first task is to build a model from the data available to us which can reduce the total cost of the intervention program.

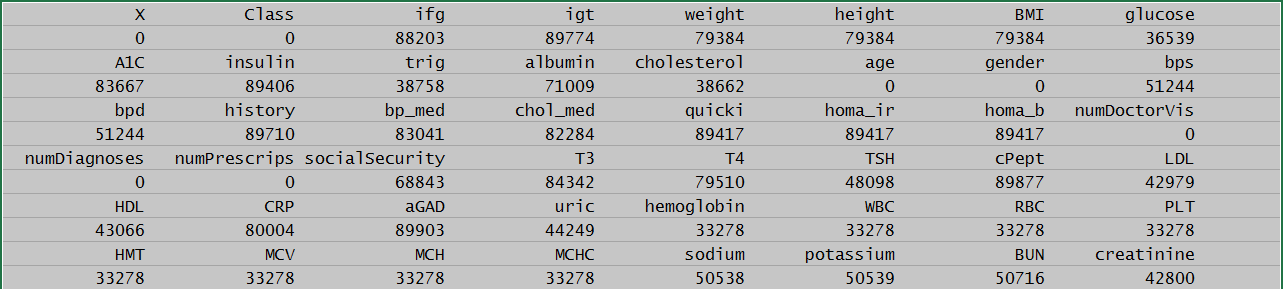
**Data Summary:**

The data available has **89903** observations and **47** variables including the response variable “Class”. Some of the variables have missing data as well as shown below. “NA” represents the missing values in the data. The structure of the original data is follows:



**Data Analysis:**

Number of missing observations per attributes are calculated to check the sparsity of the attributes.



Attribute **“aGAD”** has all instances missing, so it won’t add any information to the model, hence we removed this attributes. Furthermore, when we ran the J48 model on the original data, the following attributes did not affect the Classification Accuracy and ROC curve area of the model, so we removed them.

1). **“ifg”** (Impaired Fasting Glycaemia)

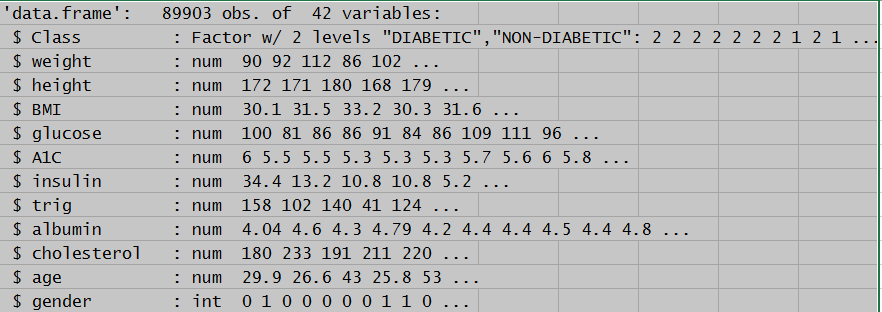
2). **“igt”** (Impaired Glucose tolerance)

3). **“Chol\_med”** (Cholesterol Medical)

4). **“history”**

**Data Imputation:**

We have done the imputation on the data available after removing the above attributes using the package **“mice”** (**Multivariate imputation by chained equation**) available in **Rstudio** software. The imputed data looks as follows:



**Comparison of Model on Original and Imputed Data:**

For the comparison between Raw data and Imputed data, we used the **cross validation with N fold=10** for different models. Following are the characteristics of the models we used:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy rate** | | **ROC curve** | |
|  | **Original Data** | **Imputed Data** | **Original Data** | **Imputed Data** |
| **J48** | **94.92%** | **94.965%** | **0.616** | **0.624** |
| **Logistic Regression** | **94.91%** | **94.914%** | **0.667** | **0.686** |
| **Naïve Bayes** | **91.58%** | **91.980%** | **0.667** | **0.672** |
| **Bagging with 70 iterations** | **94.53%** | **95.024%** | **0.752** | **0.760** |
| **Random\_forest with 100 iterations** | **94.78%** | **95.032%** | **0.748** | **0.754** |

We have witnessed that for each algorithm, the imputed data has better Classification Accuracy and ROC curve area, hence we used the Imputed data for our task.

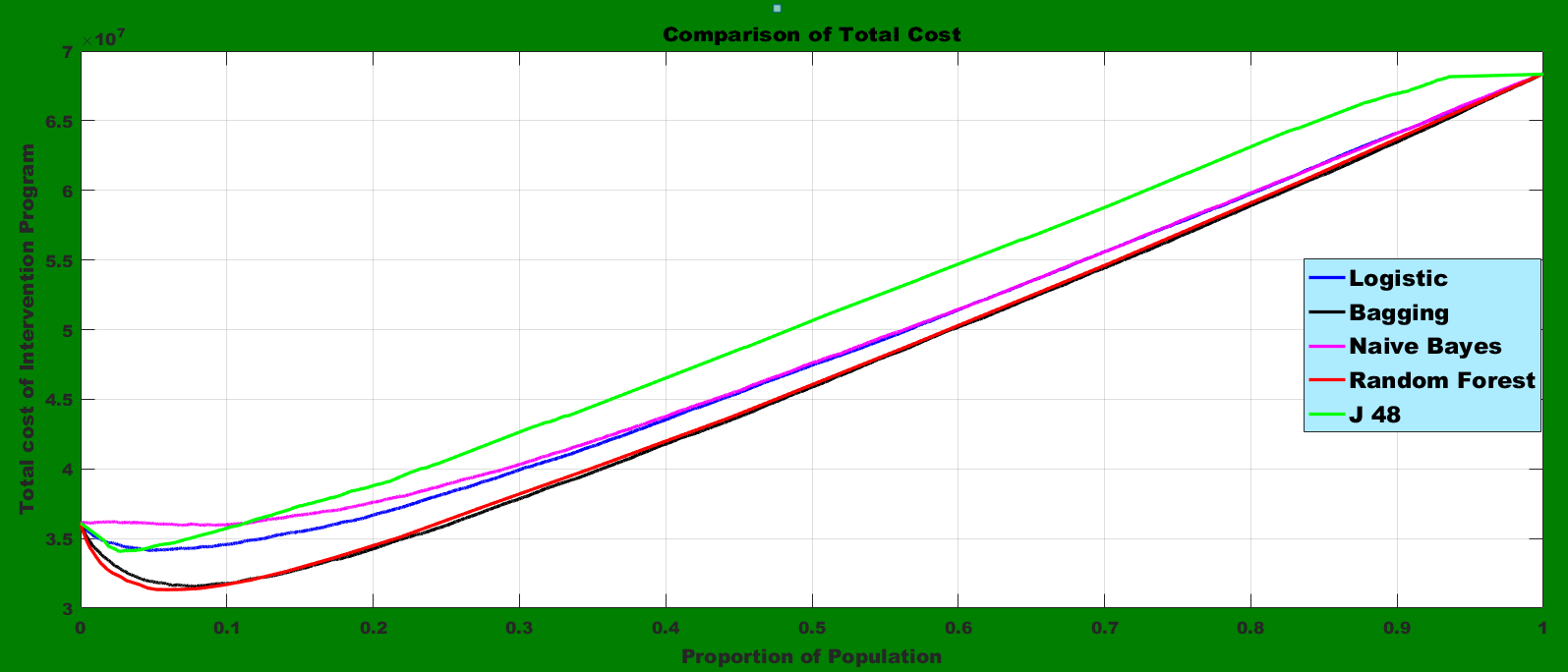
PART 1: Minimise the Total Cost of Intervention Program

We have used different algorithms like J48, Naïve Bayes, Logistic Regression, Bagging, Random Forest to estimate the Total Cost incurred to the Health Care Organisation. **Random Forest with 100 iterations** has the minimum total cost overall when targeting the **top 6% of population**. So, we compare the total cost for all the models by targeting the top 6% only.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **True Positive** | **False Positive** | **False Negative** | **True Negative** | **Threshold score** | **recall (Diabetic)** | **Total Cost ($)** |
| **J48** | **994** | **4403** | **3577** | **80929** | **0.057** | **0.217** | **34,637,540** |
| **Logistic Regression** | **1081** | **4313** | **3490** | **81019** | **0.12** | **0.236** | **34,223,360** |
| **Naïve Bayes** | **699** | **4695** | **3872** | **80637** | **0.2915** | **0.153** | **36,034,040** |
| **Bagging with 70 iterations** | **1624** | **3770** | **2947** | **81562** | **0.129** | **0.355** | **31,649,540** |
| **Random forest with 100 iterations** | **1684** | **3649** | **2887** | **81683** | **0.15** | **0.368** | **31,328,540** |
| **Current practice with Glucose** | **836** | **4983** | **3735** | **80349** |  | **0.183** | **35,639,660** |

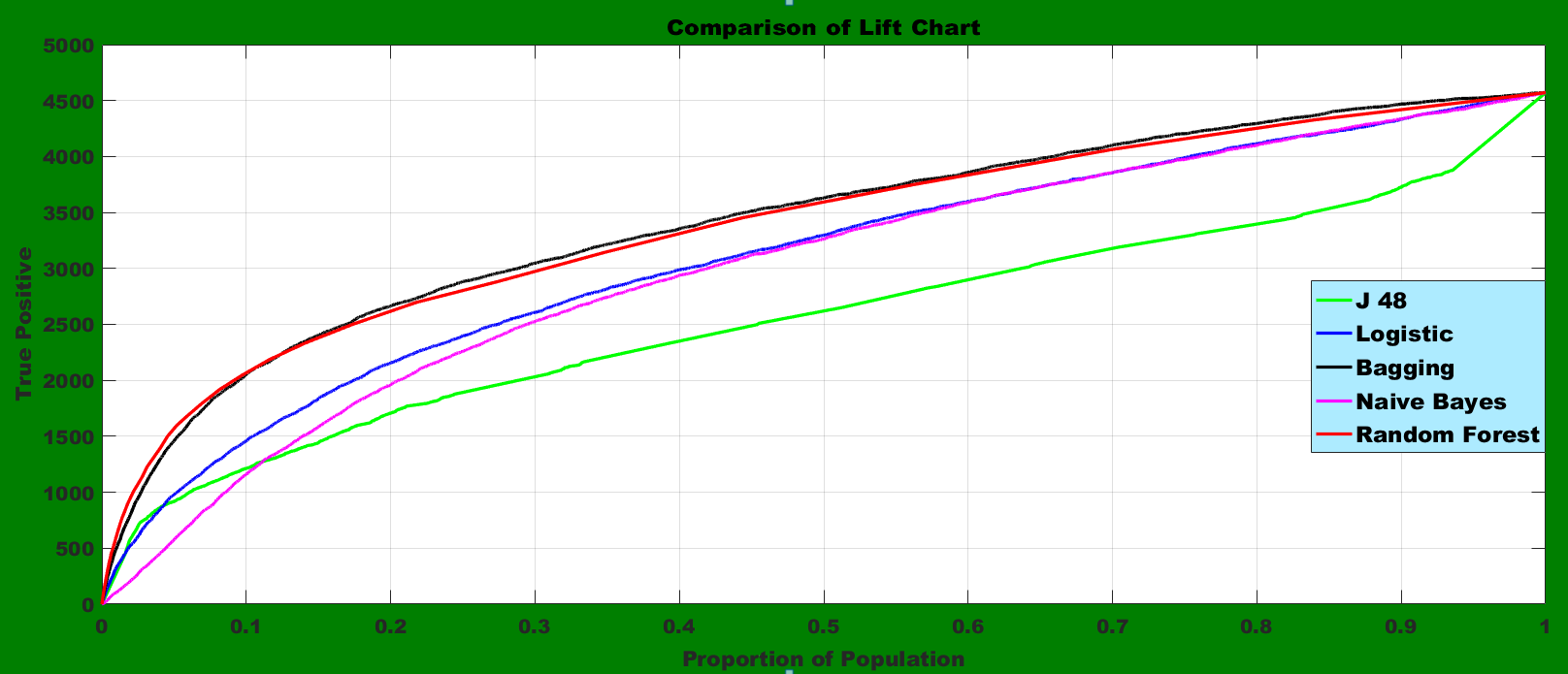
Among the above models, the best one is **Random Forest with 100 iterations** as it has the lowest estimated Total cost of the intervention program. The Total cost incurred to the Health care Organisation according to the current practice is $**35,639,660.** By implementing the Random Forest with **100 iterations model** we save the cost by **$4,311,120** and the recall improved from **0.183 to 0.368**. Saving per prediabetic patient is $ **943.15.** As per CDC **(Centers for Disease Control and Prevention)** **86 Million** people in USA are prediabetic, so by implementing the new method we can save **81 Billion.**

**TOTAL COST CHART**



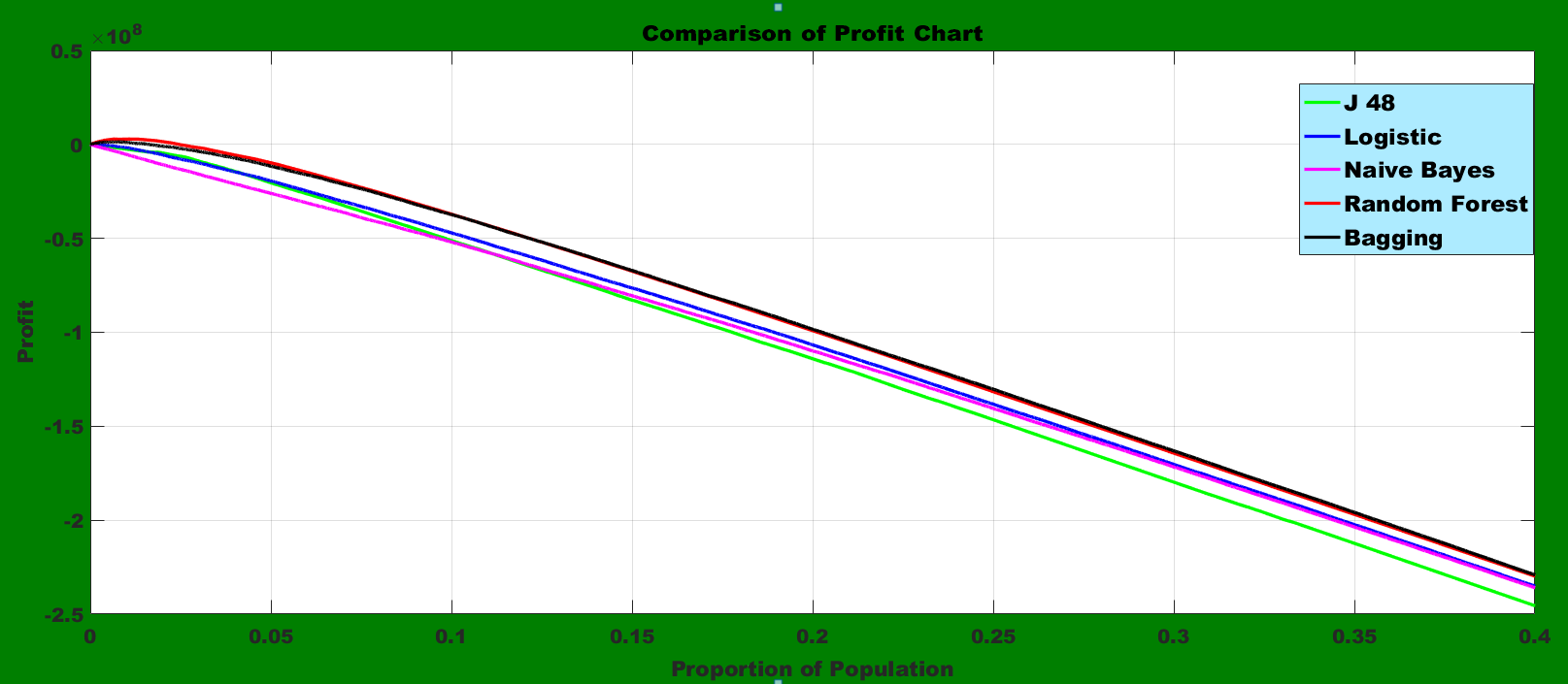
We are getting minimum cost when targeting the 6%(0.06) of Population with the best model.

**LIFT CHARTS**



Since, we are targeting the top 6%(0.06) of the population, hence Random Forest has the better lift in comparison to other models.

**PROFIT CHARTS**

****

For the top 6%(0.06) of population, Random Forest has the maximum profit value.

**PART 2: Selecting the Most Accurate Model for Prediction**

To check the out of sample accuracy of the different model we need to split the entire data into training set and testing set. Since, the number of diabetes instances are less, hence we need to split the data using **StratifiedRemoveFolds** in “weka” software. We use the number of folds equal to 3, so our training set has 70% of the entire data.

On the training set we used different models with **10-Fold Cross-Validation** and compare their Classification Accuracy. The model with the best Accuracy rate will be compared against the current practice. Each model calculates the accuracy at threshold score of 0.5. We alter the score threshold to see whether we can have higher testing accuracy at different threshold.

**Training Set:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy (Threshold =0.5)** | **Maximum Accuracy (threshold)** |
| J48 | 94.05% | 94.344% (0.93) |
| Naïve bayes | 91.69% | 94.68% (1) |
| Logistic Regression | 94.96% | 94.986% (0.36) |
| Random\_Forest with 100 Iterations | 95.05% | 95.248% (0.31) |
| Bagging with 60 iterations | 95.12% | 95.151% (0.41) |
| K Nearest Neighbors with K=5 | 94.85% | 94.959%(0.80) |

**Testing Set:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy (Threshold =0.5)** | **Maximum Accuracy (threshold)** |
| J48 | 94.02% | 94.23% (0.93) |
| Naïve bayes | 91.87% | 94.614% (1) |
| Logistic Regression | 94.82% | 94.81%(0.36) |
| Random\_Forest with 100 Iterations | 94.88% | 95.188% (0.31) |
| Bagging with 60 iterations | 94.99% | 95.018% (0.41) |
| K Nearest Neighbors with K=5 | 94.70% | 94.873%(0.80) |

Clearly, The **Random\_Forest with 100 iterations at threshold 0.31** gave the best Training and testing set accuracy rate. Since, the decline in accuracy rate from training to testing set is not very high, it implies that overfitting is absent also.

**Comparison of Best Model with Current Practice:**

Using the Current Practice, where we used only Glucose level as the classifier, the confusion Matrix of the testing set is:

|  |  |
| --- | --- |
| True Positive | 238 |
| False Positive | 1386 |
| False Negative | 1312 |
| True Negative | 27032 |
| Accuracy rate | 91.00% |

The Accuracy rate of the current practice **91%** and corresponding error rate is **9%** Hence, we recommend to use **Random\_Forest (with 100 iterations at threshold 0.31)** as it has testing accuracy of **95.188%** and error rate of **4.812%.**

**Feature Addition:**

We have added the following attributes in the training set and testing set and ran **Random Forest with 100 iterations** on the dataset.

1. **K\_5\_prediction:**

Probabilities of response variable calculated using K Nearest Neighbors.

1. **Naïve\_Bayes:**

Probabilities of response variable calculated using Naïve Bayes.

1. **Logistic\_prediction:**

Probabilities of response variable calculated using Logistic Regression.

The Classification Accuracy rate on the training set and testing set by adding the above three attributes in the dataset achieved is **95.21%** and **95.06%** at threshold score of **0.33** respectively. However, this Classification accuracy rate is still less than our Best model.

**Feature Selection using Wrapper approach:**

For features selection, we used the “**Boruta”** Package in **Rstudio**. It uses the wrapper approach to calculate the importance of the attributes. Attribute importance is calculated using the Z-Score which is proportional to the decrease in model accuracy when an attribute is removed. We keep the best **10 attributes** in our dataset and ran the **Random\_Forest with 100 iterations** algorithm to check whether we can improve the accuracy rate of the model. The 10 important features are:

1. Age
2. numDoctorVis
3. numDiagnosis
4. numPrescrips
5. glucose
6. A1c
7. Uric
8. Creatinine
9. HDL
10. Gender

The classification accuracy increased to **95.48%** on training set at threshold score of **0.47**. The testing set accuracy at the same threshold score is **95.385%.** By keeping the ten most important attributes in the dataset only we improved the accuracy by **0.232** and **0.197** on training and testing set respectively.

**Part 3: Selecting the Distinctive characteristics**

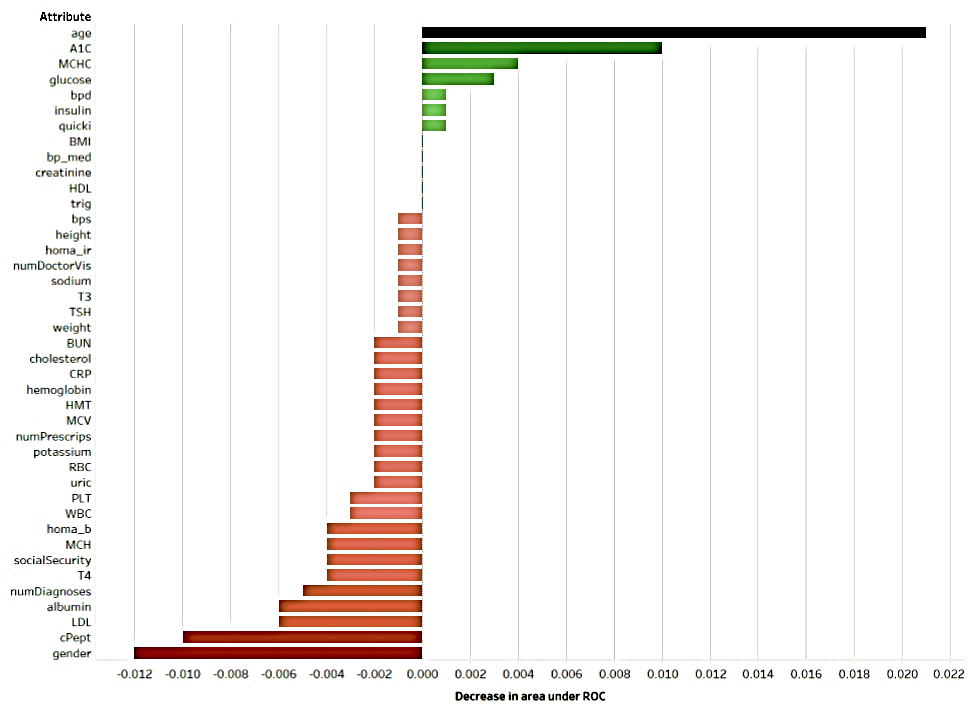
To find out the set of highly predictive characteristics, we used “**Decrease in area under the ROC curve**” as the key factor. We calculate decrease in the ROC curve area for each attribute before and after removing it. The attributes which lead to decrease in area are classified as important features. We then kept those attributes only in the dataset and ran J48 model to get the decision rules which gave us the pure nodes.

**Decision Rules:**

1. If (**A1C > 6.2** & **Age > 56.6744** & **glucose > 109** & **AIC > 6.1**) then probability of Diabetes is (**146/200**)**.**
2. If (**MCHC > 33.375** & **(6.1 < A1C < = 6.2)** & **Age > 56.6744** & **glucose > 109**) then probability of Diabetes is (**43/55**).

**Features Visualisation based on decrease in area under ROC:**

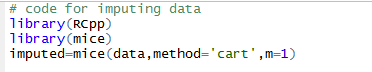
To illustrate the highly predictive attributes, we construct a histogram showing the decrease in area under the ROC curve for each attribute.



**Summary:**

There were different objectives within the project. The first one was to minimize the total cost for the intervention program. For this the best model was **Random Forest (with 100 iterations)** that provided a cost saving of **$4,311,120,** which is equivalent to **12.1%** saving. The second one was to accurately predict whether a patient will have diabetes. For this we split the dataset into training and testing set using stratified technique. The best model was **Random\_Forest** (with **100** iterations at threshold score of **0.31**) with a test accuracy rate of **95.188%.** We further improve this accuracy to **95.385% by using the wrapper approach**. The last objective was to find the most important predictors as well as finding the characteristics of patients with diabetes. To find the most important predictors we used **“Decrease in area under ROC curve”** as the deciding factor. To find the distinctive set of characteristics and decision rules of patients with diabetes, a **J48 tree** model was built by using only those attributes whose absence lead to decrease in area of ROC curve.

**Appendix**



Data is the dataset which is given to us.

