[Year]

Ray, Sharmila

[Company name]

[Date]

**Diabetes Prediction Amongst Pima Indian Women**

SHARMILA RAY

Spring 2021

PROJECT REPORT

|  |  |  |
| --- | --- | --- |
| **SR.NO** | **CONTENTS** | **PG.NO** |
| **1** | **Executive Summary** | 2 |
| **2** | **Project Motivation - Business / Social Question** | 3 |
| **3** | **Data Description** | 4 - 5 |
| **4** | **Data Cleaning and Preprocessing** | 6 |
| **5** | **Data Analysis Using Visualization - Findings** | 7-10 |
| **6** | **Data Partitioning** | 11 |
| **7** | **Prediction Models - Findings** | 11-14 |
| **8** | **Conclusions** | 15 |
| **9** | **Recommendations** | 15 |
| **10** | **Acknowledgments and References** | 16 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**INDEX**

## EXECUTIVE SUMMARY

**Background**

The Pima are a group of Native Americans living in Central and Southern Arizona, as well as Northwestern Mexico. Despite being a very homogeneous community, the Pima Indians in America have the highest rates of diabetes not only in comparison to their cousins across the border, but anywhere in the world. Their genetical homogeneity thus makes them a perfect group to study the predispositions of the disease.

In 1965, the National Institute of Health initiated a long-term observational study of the Pima tribe that has now lasted about 56 years. To inform planning and better expanded NIH health services, the NIH commissioned us to use modern Machine Learning methodologies to come up with new insights on diabetes. The objectives of the assessment were:

* To discover the determinants of diabetes and shed new light on older studies.
* To propose a set of recommendations on how to best address existing gaps in healthcare policies considering the insights from the Pima Indian case study.
* To assess the range, scope, and scale of diabetes amongst Pima women
* To be able to effectively predict diabetes in Pima women with measurable accuracy.

**Methodology**

The data scientist did an extensive exploratory data analysis of the dataset. The data scientist identified the variables and data types to determine how best to handle them. The data scientist then used various visualization methods to get an overall picture of what story the data was trying to tell. This was also used to identify outliers, missing values, skewness and come up with an initial guess of what models could better handle the data. The data was then divided into training, validation and testing for modelling. The data scientist then ran various models, tested each, transformed certain variables before retesting and coming up with the most accurate model.

**Summary of Key Findings**

It was discovered that blood glucose concentration, Body Mass Index, and age, in that order, were the greatest risk factors for potentially developing diabetes.

Factors such as pregnancy, blood pressure and insulin are also significant contributors in this prediction.

Diabetes is not an old people disease as it is commonly assumed. We will see those middle-aged women, in between their 1st and 8th pregnancies are the most at risk of any segregated group.

## PROJECT MOTIVATION

Diabetes is a disease in which your blood glucose, or [blood sugar](https://medlineplus.gov/bloodsugar.html), levels are too high. Glucose comes from the foods you eat. Insulin is a hormone that helps the glucose get into your cells to give them energy. Symptoms often include [frequent urination](https://en.wikipedia.org/wiki/Frequent_urination), [increased thirst](https://en.wikipedia.org/wiki/Polydipsia) and [increased appetite](https://en.wikipedia.org/wiki/Polyphagia). If left untreated, diabetes can cause [many health complications](https://en.wikipedia.org/wiki/Complications_of_diabetes_mellitus). About 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.6 million deaths are directly attributed to diabetes each year.

## Business Question

As consultants hired by the National Institute of Health, we seek to identify individuals that are predisposed to diabetes. This will enable the agency to focus their resources on specific groups of individuals and reduce wastage of resources and money.

**Part 1** **Determinants of Diabetes Risk to advice PIMA community policy makers**

All the variables collected in the dataset are determinants of diabetes risk. However, Skin Thickness could not be used conclusively for model building. The R2 value of skin thickness was significantly small and was thus rejected when running both models. Skin Thickness also showed a negative linear correlation with Age; when it increases, age decreases. This makes sense as older people tend to have loose skin. On the other hand, overweight individuals also tend to have loose skin. This threw into disarray an earlier hypothesis of skin thickness being an indication of body weight.

**Part 2** **Predicting diabetics with accuracy to take preemptive measures.**

Hence, we are predicting if women belonging to the Pima native Indian tribe are diabetic or not, based on the attributes of the dataset.

Using logistic regression and decision tree, predict whether the subject has diabetes or not.

## DATA DESCRIPTION

## Dataset Source

**Kaggle:** [**https://www.kaggle.com/kandij/diabetes-dataset**](https://www.kaggle.com/kandij/diabetes-dataset)

The data was collected and made available by **“National Institute of Diabetes and Digestive and Kidney Diseases” as part of the Pima Indians Diabetes Database**. Several constraints were placed on the selection of these instances from a larger database. All patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above. The Pima Indian Diabetes Dataset contains information of 768 women from a population near Phoenix, Arizona, USA. The outcome tested was Diabetes. 268 tested positive and 500 tested negatives. Therefore, we have regarded Outcome as our target (dependent) variable with the following attributes (independent) variables:

## Variable Summary

* **Outcome** (Diabetes or No Diabetes) **1 & 0**
* **Pregnancies** (number of times pregnant) **0 – 17 range**
* **Oral glucose tolerance test** - OGTT (two-hour plasma glucose concentration after 75g anhydrous glucose in mg/dl), **0 – 199 range**

It stands for milligrams per deciliter. Two hours after you finish the **glucose** drink, this is what your results **mean**: Below 140 mg/dL: normal **blood sugar**. Between 140 and **199**: impaired **glucose tolerance**, or prediabetes. 200 or higher: diabetes.

* **Blood Pressure** (Diastolic Blood Pressure in mmHg), **0 – 122 range**

The **diastolic** reading, or the bottom number, is the **pressure** in the arteries when the heart rests between beats. This is the time when the heart fills with blood and gets oxygen. A normal **diastolic** blood **pressure** is lower than 80. A reading of 90 or higher means you have high blood **pressure**.

* **Skin Thickness** (Triceps skin fold thickness in mm), **0 – 99 range**

Triceps - The back of the upper arm

Triceps skinfold thickness is done to determine body fat percentage.

Triceps Skin Fold Thickness: Normal 23mm

* **Insulin** (2 h serum insulin in mu U/ml), **0 – 846 range**

Greater than 150 mu U/ml relates to insulin therapy.

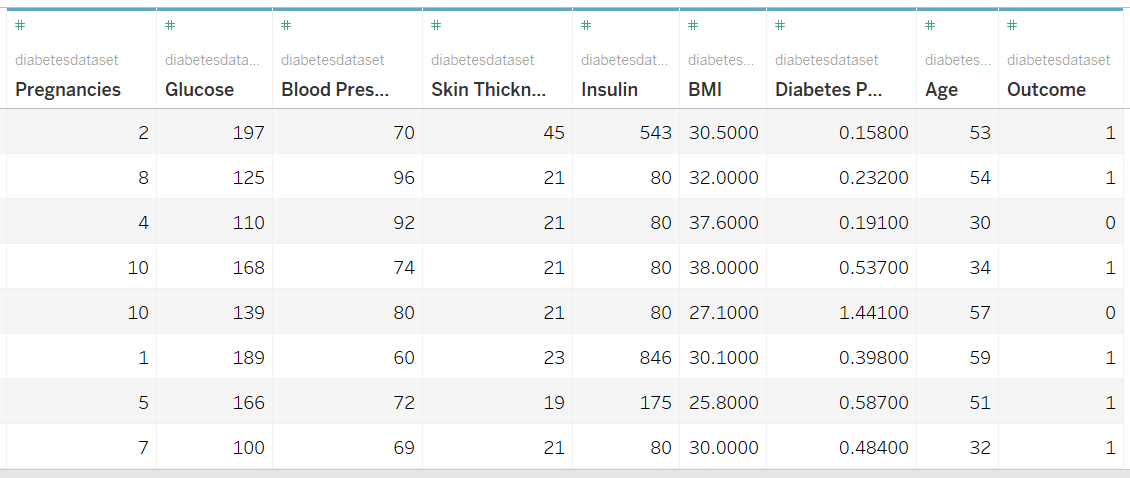
* **BMI** (Body Mass Index in kg/m2), **0 – 67.1 range**

Doctors consider a healthy **BMI** for **women** to be 18.5–24.9. A **BMI** of 30 or above may indicate **obesity**. **BMI** measurements can help someone understand whether they are underweight or **overweight**. However, **BMI** for **women** has some limitations, as it does not measure body fat specifically.

* **Age** (years), **21 yrs. to 81 yrs. (All Women)**
* **Pedigree Diabetes Function** ('function that represents how likely they are to get the disease by extrapolating from their ancestor’s history') **0 .078 – 2.42 range.**

1. If = 0.5 for parent, full sibling
2. If = 0.25 half sibling, grandparent, aunt, or uncle
3. If = 0.125 half aunt, half uncle, or first cousins

## Dataset Excerpt



## DATA CLEANING AND PREPROCESSING

1. We had 5 columns in the dataset with missing values we replaced them with the mean of those columns.

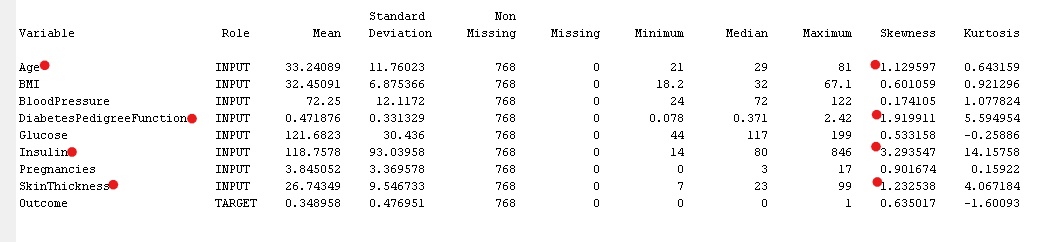
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Glucose** | **Blood Pressure** | **Skin Thickness** | **Insulin** | **BMI** |
| **Range** | 0 - 199 | 0- 122 | 0 - 99 | 0 - 846 | 0 - 67.1 |
| **Missing Values** | 5 | 35 | 227 | 374 | 10 |
| **Mean** | **120.89** | **69.1** | **20.53** | **79. 79** | **31.99** |
| **Median** | 117 | 72 | 23 | 30.5 | 32 |

1. Based on the variable summary in SAS EM Stat Explore, we found skewness over 1 in Variables – Age, Diabetes Pedigree Function, Insulin and Skin Thickness and thus decided to transform -

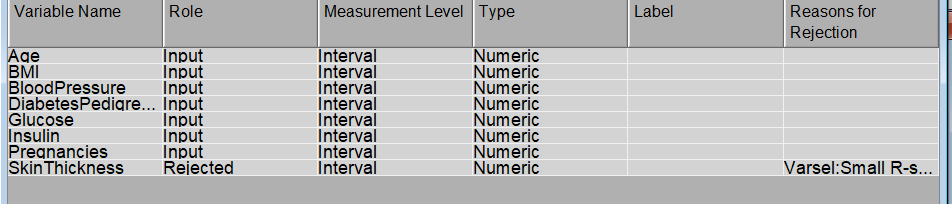
Age – **Optimal Binning**

Diabetes Pedigree Function and insulin – **Log 10**

**Variable Summary:**

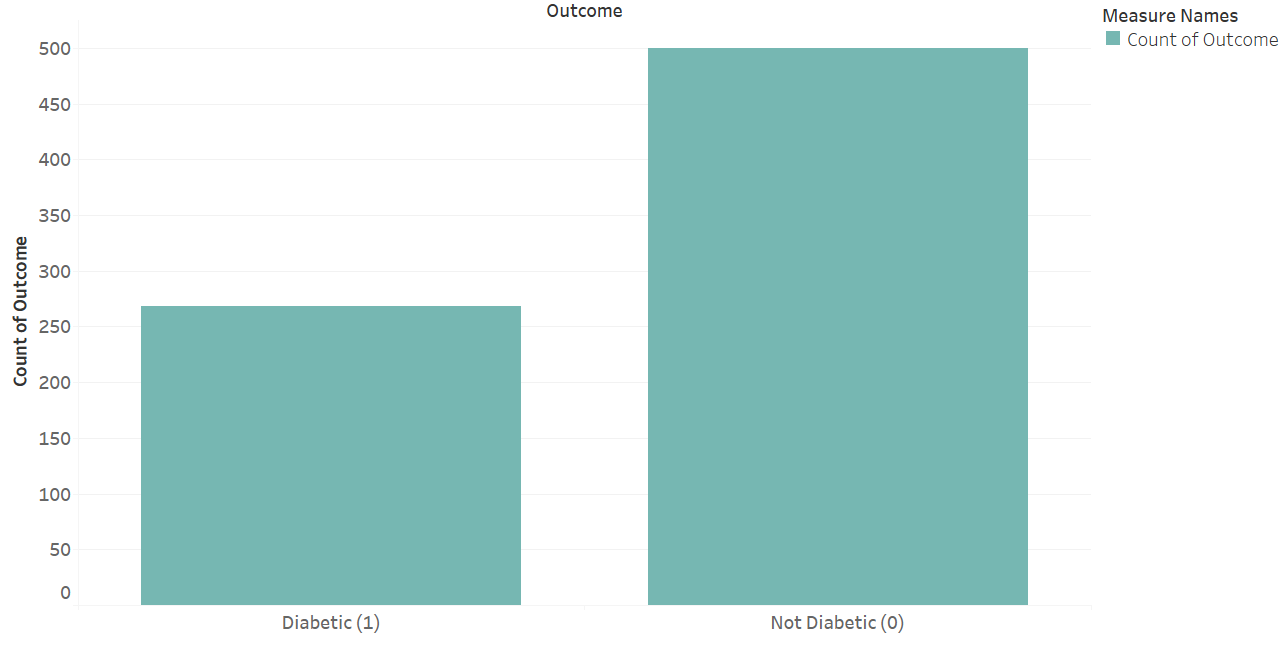


1. Dimension Reduction was done based on small R - Squared found in variable Skin Thickness, we thus **rejected** variable Skin Thickness before running our Model



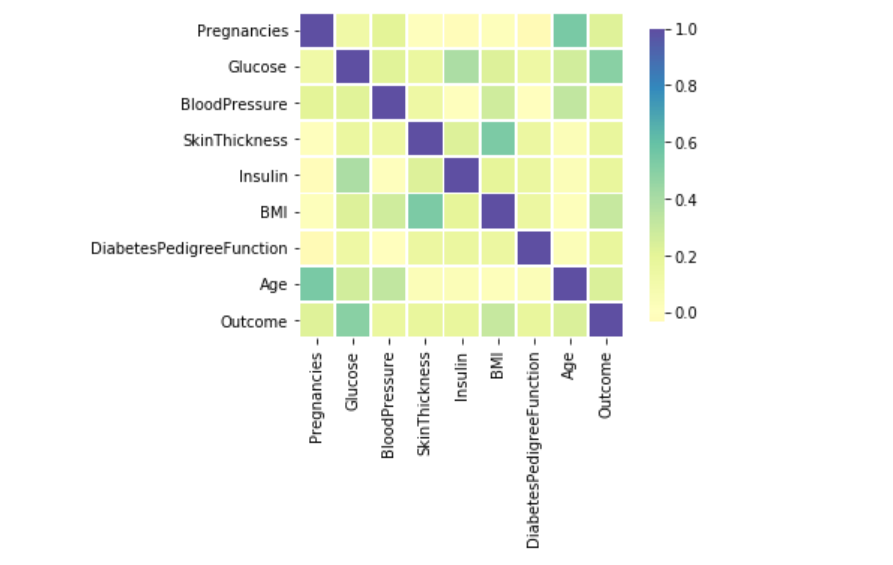
## DATA ANALYSIS USING VISUALIZATION (With Tableau, Python, SAS EM)

**Figure 1: Bar Chart** - Class distribution of target variable **(Tableau)**



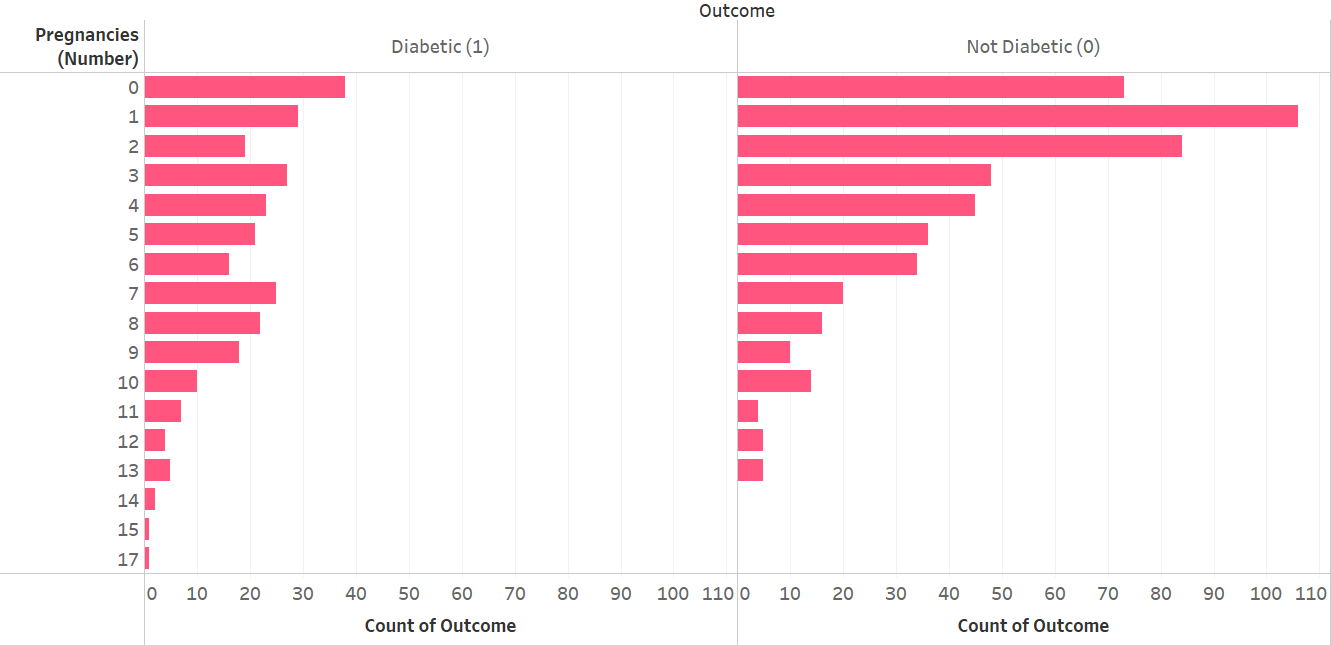
Number of diabetic records are 268 (35%) and number of not-diabetic records are 500(65%)

**Figure 2: Correlation Matrix –** Heat Map **(Python)**



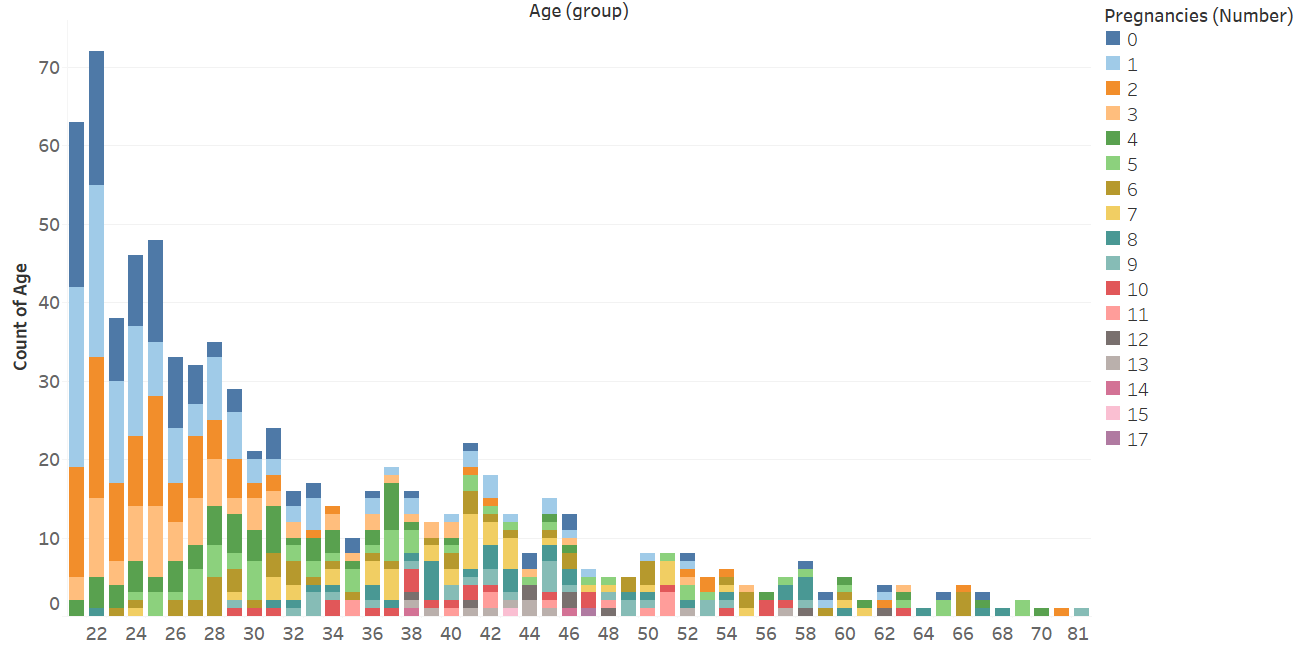
Here, the correlation matrix shows that outcome has relatively strong correlation with Glucose and moderate correlation with BMI, Age and Pregnancy amongst the attributes in the dataset. Also, Insulin and Glucose, BMI and Skin Thickness, Pregnancy and Age exhibits some strong significant correlations.

**Figure 3: Horizontal Bars -** Outcome Vs Pregnancy **(Tableau)**



We found that 14.18% of Diabetic and 14.6% of Non-Diabetic records have pregnancies as zero. And 55.6% of Diabetic and 29.94% Non-Diabetic records have pregnancies between 1 and 17.

**Figure 4: Stacked Bars -** Pregnancy Vs Age **(Tableau)**

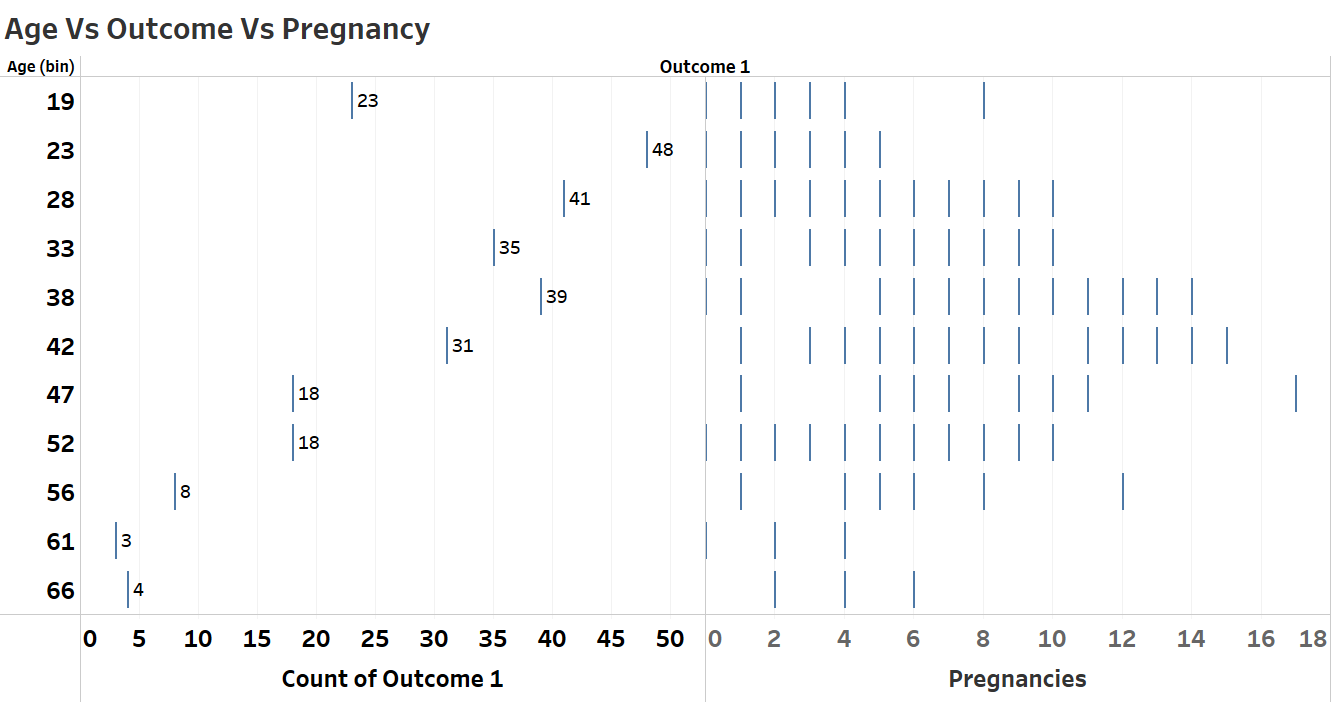


We found that most records are women in the age group 22 with number of pregnancies 0 to 5

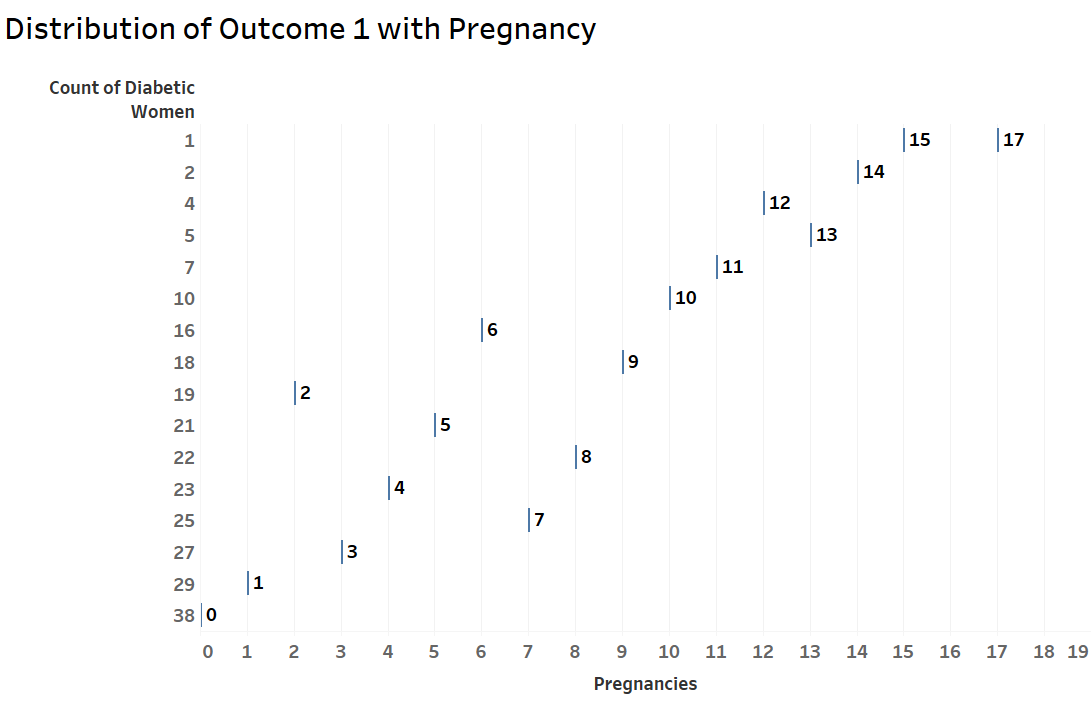
**Figure 5:** Histograms – Significant Variables by Outcome **(SAS EM)**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

### Findings based on Visualization



**At 72.4%, the largest number of people with diabetes are between the ages of 23 and 42.**



**Diabetes is also significantly higher amongst women in their 1st to 8th pregnancies.**

## DATA PARTITIONING

We decided to proceed with stratified partitioning method to ensure that the input variables along with target variable are represented in same manner in all of training, validation, and test datasets to avoid any bias and to improve prediction accuracy of the model.

Partition the data into:

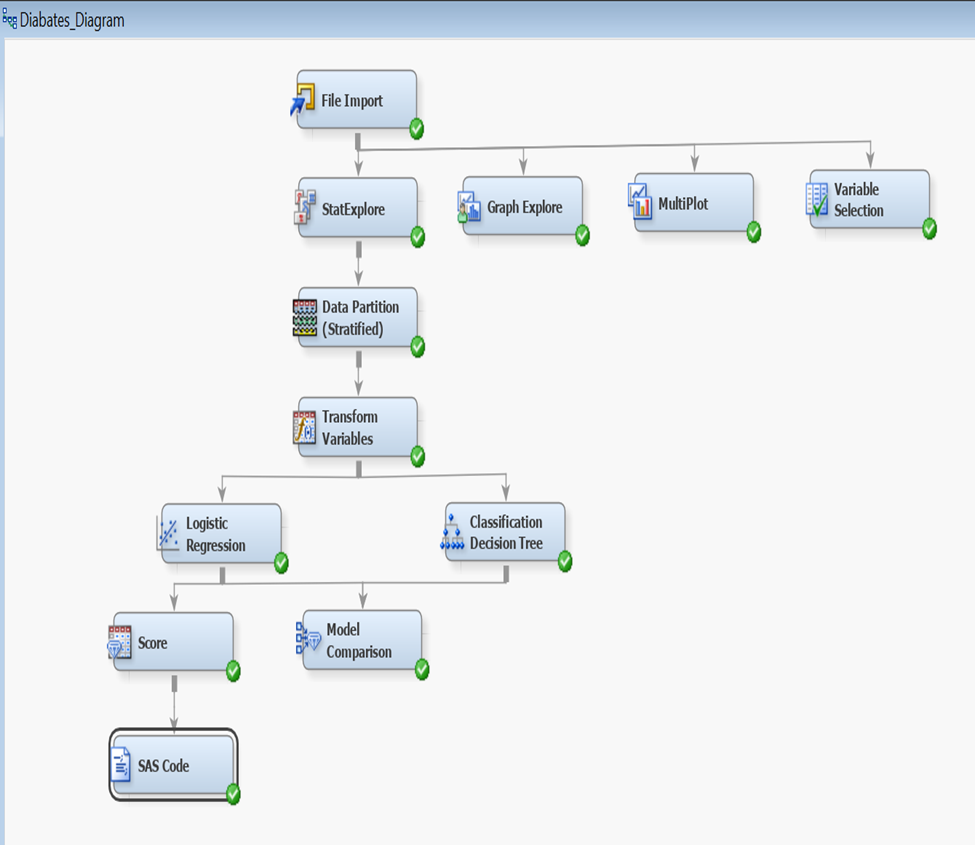
1. Train (70%) – for training model
2. Validate (20%) - for validating best fit model performance
3. Test (10%) - for scoring model against new data

## PREDICTION MODELS (WITH SAS ENTERPRISE MINER)

Our target variable (Outcome) is of binary type. Hence, we selected these two data mining Models which are better suited to predict the binary outcome of presence of diabetes.

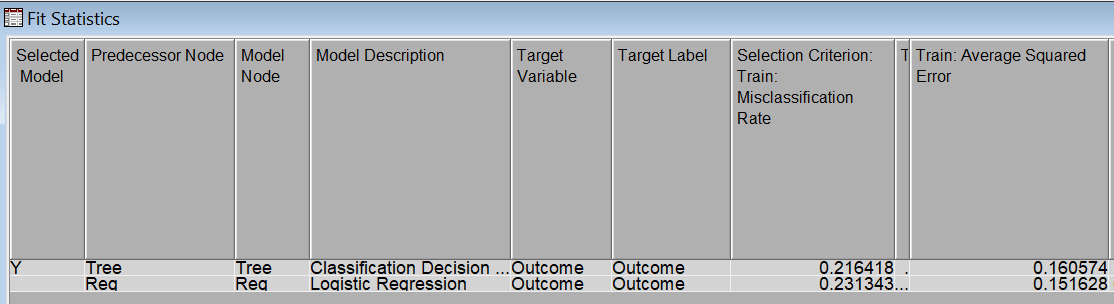
* Logistic Regression
* Classification Decision Tree

**Enterprise Miner Diagram**

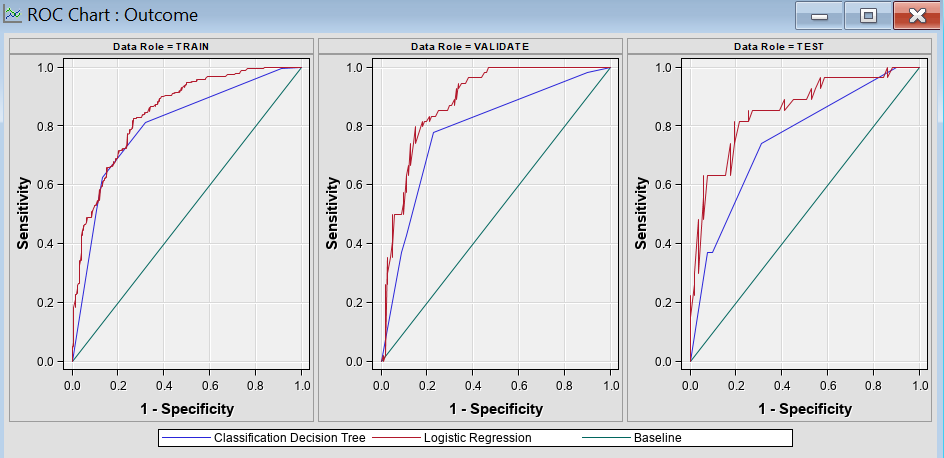


## Findings based on Modelling

**Assessing Model Performance:** We did a model comparison between Logistic Regression and Classification Decision Tree in SAS EM and found that based on the Misclassification Rate statistic, SAS chose the Classification Decision Tree as the best fit model for prediction.



The comparison of the ROC curves however shows that Logistic Regression performs better in terms of the Area Under the Curve from the ROC chart as shown below:



**Assessing model performance Metrics:**

We calculated the performance metrics based on,

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Accuracy of Prediction or the correct classification rate is calculated as:

**Accuracy = (TP + TN) / (FN + TN + FP + TP) \* 100 %**



Recall or true positive rate also called sensitivity is calculated as:

**Recall = TP / (TP + FN) \* 100 %**



Precision or positive predicted value is calculated as:

**Precision = TP / (TP + FP) \* 100%**



**Evaluating best fit classifier Model:**

Based on the ROC curve as well as analysis of the Accuracy, Recall and Precision values obtained from both Logistic Regression and Classification Decision Tree classifier we conclude that Logistic Regression performs best amongst the two.

This can be seen from the table below on validation data

|  |  |  |  |
| --- | --- | --- | --- |
| **Model performance on validation data** | **Accuracy %** | **Recall %** | **Precision %** |
| Logistic Regression | 77.27% | 53.70% | 74.35% |
| Classification Decision Tree | 72.72% | 42.59% | 67.64% |

On running the Logistic Regression Model on the Test data, we achieved an accuracy of 82%



# 

# Dataset used for data mining model building



# Decision Tree



Prediction results run on test data (Extracted from Test data using Logistic Regression model)



**CONCLUSION**

According to the World Health Organization, in 2012 diabetes was the direct cause of 1.5 million deaths globally. Untreated diabetes can also lead kidney failure, nerve damage and blindness. There are 2 major forms of diabetes. Type 1 is characterized by a lack of insulin production and Type 2 results from the body’s ineffective use of insulin. While Type 2 is potentially preventable, the causes and risk factors for Type 1 remain unknown. Gestational diabetes on the other hand, mainly affects women during pregnancy. Those who develop it are at a higher risk for developing Type 2 later in life. Pima Indians are mainly afflicted by Type 2 diabetes. This is the primary reason why the number of pregnancies was significant in this data set.

In this paper, we proposed a Machine Learning method to analyze the PIMA data set and predict diabetes based on available features. We performed a Logistic Regression Model and a Classification Decision Tree. We demonstrated that the former is the most accurate of the two models.

Using the Logistic Regression Model, we can predict with 82% accuracy whether a PIMA Indian Woman is diabetic or not.

### RECOMMENDATIONS

The following are the recommendations being put forth to the National Institute of Health to better handle diabetes nationwide. They are mainly policy changes that the NIH can propose to the United States Department of Health and Human Services and they are as follows:

1. These results give the NIH the ability to target narrower subpopulations of people to better utilize money and resources.
2. The NIH should create prediabetes awareness campaigns that target commonly held misconceptions highlighted by the findings in this study.
3. Offering lifestyle counselling facilities that promote healthy and non-sedentary living like building playgrounds and gyms in communities.
4. They should also offer diabetes testing facilities in poor communities where healthy eating is not common. This coupled with continuous glucose monitoring in pregnant women during both pre- and post-natal care.
5. Educate health care providers on the provisions of the Affordable Care Act that are geared towards obesity reduction, as it is greatly straining the entire healthcare system.
6. The NIH should also strive to continuously search and plug policy gaps that have failed to stem the tide of the disease.

## REFERENCES AND ACKNOWLEDGEMENTS

* <https://www.webmd.com/diabetes/guide/oral-glucose-tolerance-test#:~:text=It%20stands%20for%20milligrams%20per,200%20or%20higher%3A%20diabetes>
* <https://www.webmd.com/hypertension-high-blood-pressure/guide/diastolic-and-systolic-blood-pressure-know-your-numbers#1>
* [http://mercury.webster.edu/aleshunas/Support%20Materials/C4.5/Lamb%20- %20FINALDIABETESPPT.pdf](http://mercury.webster.edu/aleshunas/Support%20Materials/C4.5/Lamb%20-%20%20%20%20%20%20%20%20%20FINALDIABETESPPT.pdf)
* <https://www.medicalnewstoday.com/articles/bmi-for-women#:~:text=Doctors%20consider%20a%20healthy%20BMI,not%20measure%20body%20fat%20specifically>.
* [Transform Variables: Getting Started with SAS(R) Enterprise Miner (TM) 13.1](https://support.sas.com/documentation/cdl/en/emgsj/66018/HTML/default/viewer.htm#n1mbycbcb5e5i0n1ayc2icfmx5u3.htm)
* <https://support.sas.com/kb/24/170.html>
* <https://communities.sas.com/t5/SAS-Data-Mining-and-Machine/How-to-get-a-Classification-Chart-for-the-TEST-SET-not-just/td-p/402113>
* <https://communities.sas.com/t5/SAS-Data-Mining-and-Machine/Testing-Confusion-matrix-in-Enterprise-Miner/td-p/364432>
* <https://www.who.int/features/factfiles/diabetes/en/>