Text

Description automatically generated

The University of Texas at Arlington

College Of Business

INSY 5339 001 Principles of Business Data Mining

Jayarajan Samuel

Submitted by:

Komal Keshao Panchputre

# Contents

[Contents 4](#_Toc121073521)

[Summary 5](#_Toc121073522)

[Motivation 5](#_Toc121073523)

[Section 1: Project Proposal 5](#_Toc121073524)

[1.1 Source Of Dataset 5](#_Toc121073525)

[1.2 A Brief Description of Dataset 5](#_Toc121073526)

[1.2.1 Attributes Columns: 6](#_Toc121073527)

[1.2.2 Target Columns: 6](#_Toc121073528)

[1.2.3 Issues With Data: 6](#_Toc121073529)

[1.3 Objectives of Mining This Data 7](#_Toc121073530)

[1.4 Business Problem That We Are Tackling 7](#_Toc121073531)

[1.5 Data Visualization and Prediction Techniques 8](#_Toc121073532)

[1.5.1 Prediction Techniques 8](#_Toc121073533)

[1.5.2 Visualization Techniques 11](#_Toc121073534)

[1.5.2.1 Tableau Results: 11](#_Toc121073535)

[1.5.2.2 Stacked Bars: 12](#_Toc121073536)

[1.6 Result Pondering and Observation Table 14](#_Toc121073537)

[1.7 Data Cleaning and Preprocessing 14](#_Toc121073538)

[1.7.1 Cleansing 14](#_Toc121073539)

[1.7.2 Data Partitioning 15](#_Toc121073540)

[1.7.3 Model comparison 15](#_Toc121073541)

[Section 2: Analysis using an 80:20 split of the data 16](#_Toc121073542)

[2.1 Data Visualization and Prediction Techniques 16](#_Toc121073543)

[Section 3: Conclusion 19](#_Toc121073544)

# Summary

* 96 million adults in America, or more than one in three, have prediabetes. More than 80% of people with prediabetes are ignorant of their disease.
* Prediabetes is a serious health condition where blood sugar levels are higher than normal but not yet high enough to be diagnosed as type 2 diabetes.

# Motivation

* As an insurance company, we are targeting the diabetes problem. We can basically determine a person's diabetes status based on the data.
* The insurance provider will suggest a high premium improvement plan to the impacted individuals after learning of such instances.

# Section 1: Project Proposal

In this project we want to utilize an open-source dataset related to Diabetes and perform extensive mining on the dataset. The overall goal is to get enough information out of the data so that we can make a business model by creating a prediction model out of it.

The information will be mainly focused on predicting if the subject may potentially show signs of early diabetes in their future. This will tremendously help with the diagnostic efforts.

## Source Of Dataset

**Kaggle**: <https://www.kaggle.com/kandij/diabetes-dataset>

## A Brief Description of Dataset

The dataset that we are borrowing from Kaggle has tremendous information. The dataset sports few medical predictors. The dataset has independent variables and one target variable also called as the dependable variable. The information was gathered and provided by the "National Institute of Diabetes and Digestive and Kidney Diseases". The selection of these examples from a broader database was subject to several restrictions. Every patient at this facility is a female patient who is at least 21 years old and belong to Pima Indian heritage (a subset of Native Americans).

Restricting our dataset to a smaller subset enables us to make a good training set. The extracted rules and models can easily be extended to create general prediction models for other datasets to further prove our base.

Let’s us take a deep dive into our dataset. Following is the list of Attribute columns we have,

## Attributes Columns:

1. Pregnancies: Number of times the subject was pregnant
2. Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. BloodPressure: Diastolic blood pressure (mm Hg)
4. SkinThickness: Triceps skin fold thickness (mm)
5. Insulin: 2-Hour serum insulin (mu U/ml)
6. BMI: Body mass index (weight in kg/(height in m)^2)
7. DiabetesPedigreeFunction: Diabetes pedigree function
8. Age: Age (years)

## Target Columns:

1. Outcome: Class variable (0 or 1)

## Issues With Data:

* Missing Data – It was represented with 0.
* Missing Data in Few Columns
* We also felt that few more columns such as blood type might have been useful.

## Objectives of Mining This Data

The dataset has 9 columns in all, including 8 attribute columns and 1 target column. Based on the diagnostic measurements provided in the dataset, the main goal is to forecast whether the patient has diabetes or not.

## Business Problem That We Are Tackling

Prediabetes is a significant health condition where blood sugar levels are higher than normal but not yet high enough to be diagnosed as type 2 diabetes. More than one in three American adults—or 96 million people—have prediabetes. Over 80% of people with prediabetes are unaware of their condition. However, research also shows that it can be prevented or delayed with lifestyle changes. For example, weight loss, reduced glucose level, low blood pressure, low BMI and Insulin.

Medical Insurance in the US is few of the companies that thrive on extracted information from the medical data. We are approaching this problem from an Insurance Company’s perspective. The data essentially tells us whether a person has diabetes or not. We want to extract one more abstract layer of information from the dataset and that is identifying the people with borderline probability of Diabetes. Borderline probability will help us determine cases that will have a high probability of falling prey of Diabetes in near future with their current lifestyle.

Once we identify such cases, we as the Insurance Company will suggest them a high premium improvement plan. The incentive for the patients is going to be steering away from potential risks of Diabetes in near future and incentive for the company will be the money saved on patient’s potential expenditure on Diabetes cure and prevention efforts. The patient’s enrolled in this plan will furnish further data that will not only help us improve our model but also help us cater to improve patient’s suggested plan and incentives. After monitoring the patient’s curve on the probability of Diabetes, the positive scenario where a patient successfully follows plan and data suggests their probability reduces, we will incentivize the behavior with discounts on the premium and advise on the modified lifestyle.

## Data Visualization and Prediction Techniques

## Prediction Techniques

We have done out visualization through SAS EM, Excel, and Tableau. First, the data was filtered with missing values of 0 for all other variables except for pregnancies, skin thickness, insulin and outcome variables. Then the model was created in SAS Enterprise Miner using decision tree and linear regression. For both of the model, data partion is set to 50% training and 50% validation. We will be giving the insurance in three categories, high premium, low premium, and no insurance.

We have 724 people in this dataset, out of them, we are declining insurance to 284 people because they are predicted to be diabetic, 286 are given insurace because they are predicted to be non-diabetic. Our model has also determined 154 people to be borderline diabetic. Among the declined ones, 246 are declined due to high glucose level and high BMI level. Another 21 are declined due to high **diabetespedigreefunction** which is genetic and very likely to be diabetic and 17 are declined due to higher BMI (>=45).

From the borderline diabetic people, that are higher than 28.5 of age and BMI of higher than 26.35,

will be given insurance at high premium at first, and when their conditions are met, the insurance will be lowered. What this means is that we will provide the initial health support with a very high premium plan support. We are going to give the people in this category certain goals to attain, once they are attained we will be suggesting a low premium plan as the threat of impending Diabetes would have been subsided. Similarly, for people with no borderline issue of diabetes, the insurance will be given at low premium.

Looking at the decision tree from SAS EM, we are focusing on the group of people of age greater or equal to 28.5 and BMI of equal to or greater than 26.5 to 29.5. These are the group of potential people to improve their diabetic condition, who will be given high premium at first and if they reduce their BMI to less than 26.5, their insurance will be lowered.

**Diagram

Description automatically generated**

From linear regression, we found out the probability of 62% borderline diabetic people are greater than 28.5 of age and BMI of higher than 26.35. We took this targeted group and divided into two groups, one as declined insurace and one for reapply insurance once BMI is less than 29.85.

Table

Description automatically generated

From the borderline patients of 352, with age > 28.5 and BMI higher than 26.5, we took the 58 people that are higher than 28.5 of age and have BMI between 26.35 to 29.5, and ask them to reduce their BMI to less than 26.5 by changing their lifestyle to get a low premium.

This is our predictive borderline diabetic patients that we are interested in giving insurace

## Visualization Techniques

## Tableau Results:

**Bar chart:** Class distribution of outcome variable.

Chart, bar chart

Description automatically generated

Prior to data reduction we had 768 patients. Number of diabetic patients= 268 and number of non-diabetic patients is 500.

Chart, bar chart

Description automatically generated

After reduction in Excel, total patients in our dataset are 724, out of them 250 are diabetic and 474 are non-diabetic.

Chart, bar chart

Description automatically generated

We find 82% of the data is concentrated on pregnancy numbers between 0 to 8.

## Stacked Bars:

Chart, bar chart

Description automatically generated

51% of patients in the dataset are of age below 30 (372 records).

Also, we can notice that the range on age is 21 – 81.

A computer screen capture

Description automatically generated with low confidence

The diagram shows the comparison of BMI against age.

A screenshot of a computer

Description automatically generated with medium confidence

The graph shows the pregnancies against age. This visualization tells us how pregnancies and age relate to each other. As we can see that the tallest graph is when the age is less than 31. The age category of 29-31 can be seen to have the highest number of pregnancies. We wanted to relate this data with our regression model.

## Result Pondering and Observation Table

We were performing excel manipulations on the dataset to understand and grasp the reach of the result we may get, following table is the initial result that we may got from the dataset.

|  |  |
| --- | --- |
| **Question** | **Answer** |
| How many observations in the dataset? | 768 observations |
| How many binary/categorical variables? | 1 Binary |
| 0 Categorical |
| How many continuous variables? | 8 Continuous variables |
| What is the outcome / target variable? | Outcome Variable |
| If binary or categorical: What percentage of the variables belong to each class. | 65.10% of 0 |
| 34.89% of 1 |
| If continuous:  What is the mean value of the target variable? | N/A |
| Before doing any further processing, what would your prediction of the target variable be? | The outcome of test samples is anticipated to be non-diabetic. |

## Data Cleaning and Preprocessing

## Cleansing

1. Missing data for Glucose, Blood Pressure and BMI was removed from the dataset in Excel.



In diabetes dataset 3 columns of the dataset have missing values. We removed those values since we need data with legitimate values. Please find the tally of removed values as follows

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Glucose** | **Blood Pressure** | **BMI** |
| **Range** | 0 - 199 | 0- 122 | 0 - 67.1 |
| **Missing Values** | 5 | 35 | 10 |

1. Based on the linear regression summary in SAS EM, we can see remove the factors like blood pressure, skin thickness and Insulin and keep the borderline probability of Age with BMI to predict our model.

Table

Description automatically generated

## Data Partitioning

To reduce bias and increase model prediction accuracy, we decided to use the stratified partitioning method to make sure that the input predictor variables and outcome variable are represented in the same way throughout all training and validation of 50% each.

## Model comparison

Graphical user interface, application

Description automatically generated

We did a model comparison between linear regression and decision tree and found that the sum of squared error is less for linear regression and thus it is better fit for our model.

# Section 2: Analysis using an 80:20 split of the data

The fundamental idea is that, in most circumstances, 20% of causes result in 80% of outcomes. That is how we decided 80% of the dataset goes into the training set and 20% of the dataset goes into the testing set.

## Data Visualization and Prediction Techniques

As you can see in the decision tree graphic, glucose is the root node, or highest decision node, in a tree that corresponds to the best predictor. We were concentrating on the group with a BMI of equal to or greater than and an age greater or equal to 28.5 while looking at the prior decision tree from SAS EM.

In an 80/20 split decision tree, we learned more about the borderline group, which is defined as those with a BMI larger than or equal to 32.7, an age less than 28.5, a BMI less than 45.4, and a diabetes pedigree function greater than or equal to 0.75.

Diagram

Description automatically generated

A picture containing letter

Description automatically generated

We discovered that Glucose, BMI, Age, DiabetesPedigreefunction are the important factors in the decision tree. This information was also furnished by our previous premise as well. The 80:20 added a further deep level of information that the insurance company can leverage. The information again showed that the BMI is the most significant factor that seems to be the culprit in most borderline cases. As we can see, the left side of the decision tree gives us peek into the candidates which our initial analysis gave a clean cheat. Basically, the people that have age less than 28.5 are also susceptible to future Diabetes if they don’t keep a tab on their BMI. Here the DibetesPedigreeFunction also plays a significant role. To conclude the analysis from the perspective of the Insurance company this does not change much, as we are still going to give incentive to people to keep their BMI under acceptable limits.

We discovered via linear regression that those who are older than 28.5 and have a BMI higher than 26.35 are more likely to be 62% borderline diabetic. The BMI was discovered to be a major effect in the model.

Table

Description automatically generated with medium confidence

After dividing the results between the two models, we found that the linear regression with the smaller sum of squares fit our dataset better. We obtained sum of squared errors 16% for the train dataset and 14% for the test dataset.

# Section 3: Conclusion

From our model, we found that the outcome variable has strong correlation with Glucose and moderate correlation with BMI and Diabetes Pedigree Function among other variables in the dataset. BMI, Pregnancies and Age also exhibit significant correlations with the outcome variable.

Following data partitioning in an 80/20 approach, we gained a clearer understanding of the correlations between the outcome variable and BMI, Age, and DiabetesPedigreefunction.