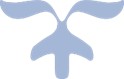


Final Project

Dataset Diabetes

December 13, 2021

Bianca Finol Hernandez



# Table of Contents

[Table of figures 1](#_Toc90255720)

[Seneca statement 2](#_Toc90255721)

[Project Overview 3](#_Toc90255722)

[Overall information about the data 3](#_Toc90255723)

[Metadata: 3](#_Toc90255724)

[Data Preparation and Data Cleaning: 4](#_Toc90255725)

[The model 6](#_Toc90255726)

[Regression Model 6](#_Toc90255727)

[Decision tree 9](#_Toc90255728)

[Comparison models 10](#_Toc90255729)

[Conclusion: 11](#_Toc90255730)

# Table of figures

[Figure 1 Sample Properties 3](#_Toc90254932)

[Figure 2 Metada 3](#_Toc90254933)

[Figure 3 Variables and Attributes 4](#_Toc90254934)

[Figure 4 Data\_row to illustrate missing values 4](#_Toc90254935)

[Figure 5 Missing value for insulin and outcome 4](#_Toc90254936)

[Figure 6 Missing value for Age, BMI, BloodPressure, DiabetesPedigreeFunction, Glucose, Pregnancies and SkinThickness 5](#_Toc90254937)

[Figure 7 Validation of cleaning missing values 5](#_Toc90254938)

[Figure 8 Changing insulin character 5](#_Toc90254939)

[Figure 9 Number of replacements 6](#_Toc90254940)

[Figure 10 Graph after run regression model stepwise method 7](#_Toc90254941)

[Figure 11 Statistical output 8](#_Toc90254942)

[Figure 12 Analysis of Maximum Likelihood Estimates 8](#_Toc90254943)

[Figure 13 Odds Ratio Estimates 9](#_Toc90254944)

[Figure 14 Decision tree- Variable importance 9](#_Toc90254945)

[Figure 15 leaf node 9](#_Toc90254946)

[Figure 16 Percentage of 0 and 1 10](#_Toc90254947)

[Figure 17 Statistical information 10](#_Toc90254948)

[Figure 18 Comparison models 10](#_Toc90254949)

[Figure 19 Confusion matrix 11](#_Toc90254950)

# Seneca statement

“*I, Bianca Finol Hernandez, declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students*.”

# Project Overview

The purpose of this dataset is to predict a person does or does not have diabetes. Validate the predictions with appropriate metrics. The target column is ‘outcome’. Try to find out what factors could increase the risk of diabetes.

Variables:

 Pregnancies: Number of time pregnant

 Glucose : concentration a 2 hours in an oral glucose tolerance test

 Blood Pressure: (mm Hg)

 Skin Thickness: (mm)

 Insulin: (mu U/ml)

 BMI : (weight in kg/(height in m)^2)

 DiabetesPedigreeFunction: provides information about history diabetes in relative and the genetic relationship of those relatives to the patient.

 Age (years)

 Outcome: Binary variable 1 =yes | 0=No

# Overall information about the data

Information about the dataset:

There are 9 variables.

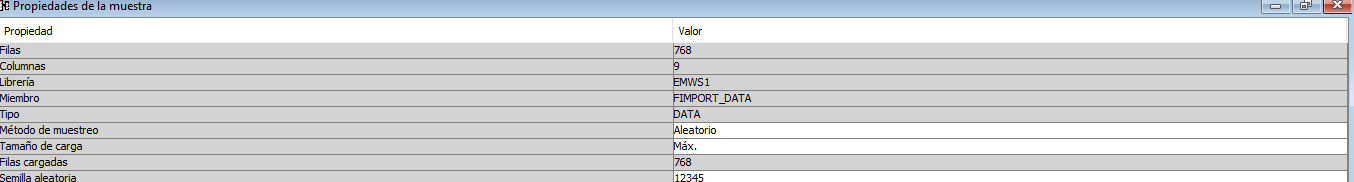


Figure Sample Properties

## Metadata:

The variables are: Age, BMI, BlooPressure, Diabetes Pedigree, Glucose, Insulin, Pregnancies, Skinthickness and Outcome as the target variable. Outcome is a binary variable where 1 means has diabetes and 0 no diabetes.

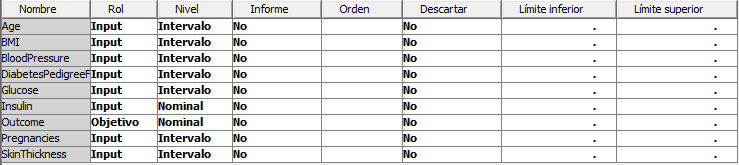


Figure Metada

The following table shows the description of the variables

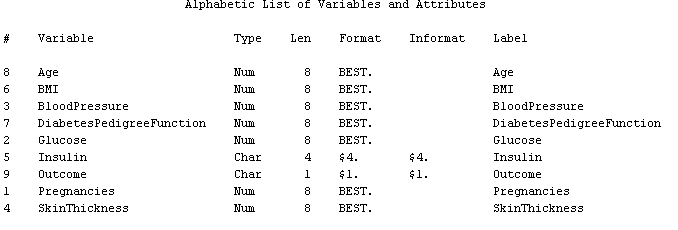


Figure Variables and Attributes

## Data Preparation and Data Cleaning:

Before to do the analysis, it is important to review the data and process with the data cleansing.

1. At first sight, there are missing values on the data. Therefore it is requested to clean data.

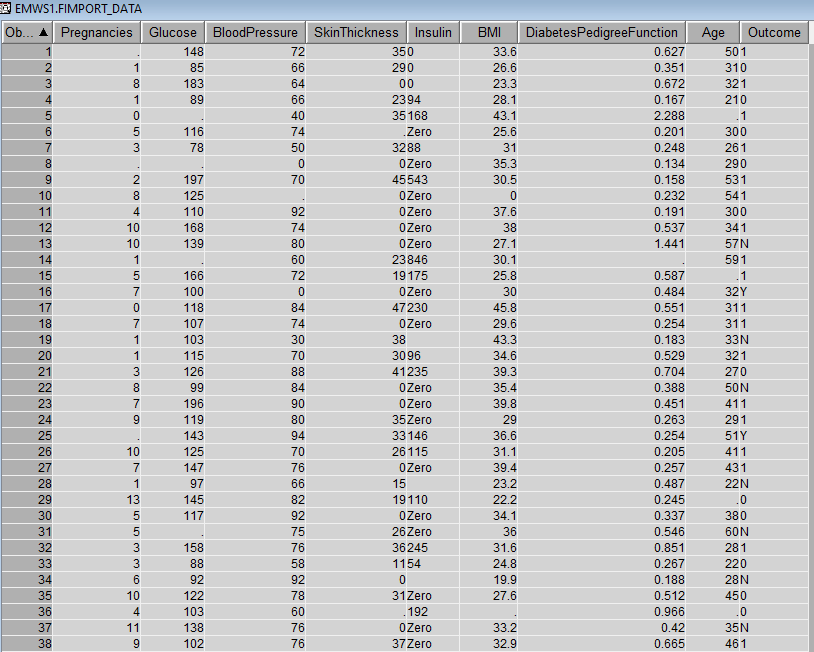


Figure Data\_row to illustrate missing values

Run some statistic to know if there are missing values. The following tables illustrate the missing values for the 9 variables.

The figure 5 point up there are 51 missing values for insulin.

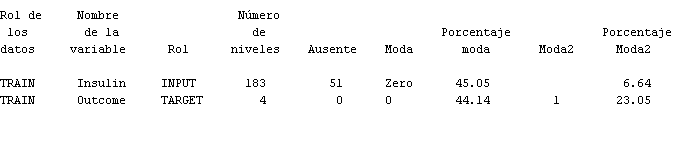


Figure Missing value for insulin and outcome

From table below is possible to know the missing values for the Age (51 missing), BMI (35 missing), Bloodpressure (34 missing), diabetes pedigree function (40 missing), glucose (38 missing), pregnancies (37 missing), Skin thickness (34 missing).

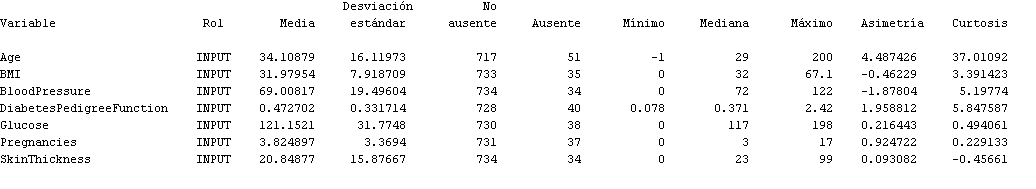


Figure Missing value for Age, BMI, BloodPressure, DiabetesPedigreeFunction, Glucose, Pregnancies and SkinThickness

It was run the data cleaning and to validate if the change was made, following table shows no missing value for the variables. However, there is possible to see that insulin has the value of “Zero” (character), therefore it is required to change for numerical.

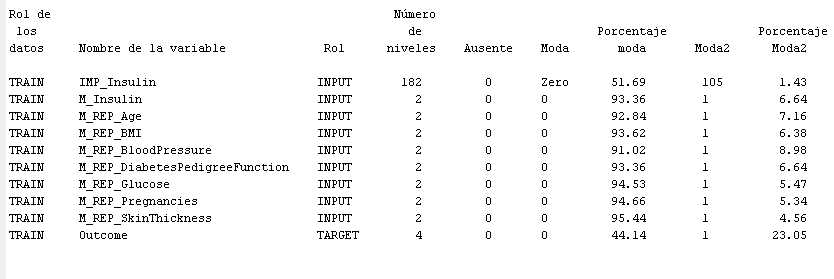


Figure Validation of cleaning missing values

The following table illustrates the transformation from character to numerical related to insulin variable.

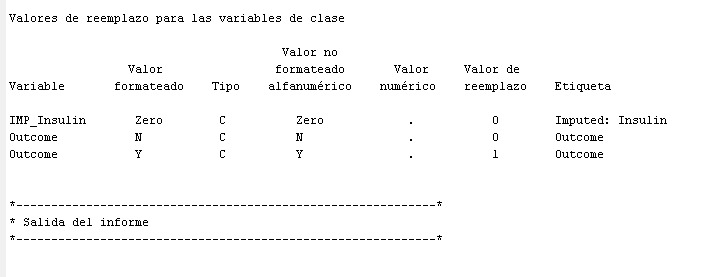


Figure Changing insulin character

The figure 9 is a summary of the replacement made for the 9 variables, in which it is impossible to see that insulin has the largest number of replacements. Additionally, it is important to highlight for the variable outcome, the replacement was Y for 1 and N for 0.

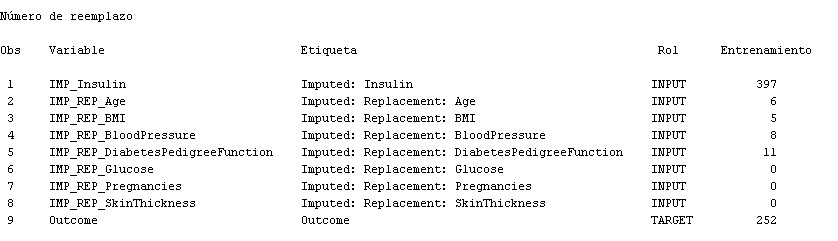
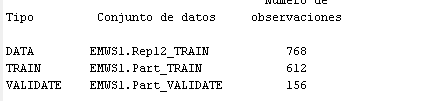


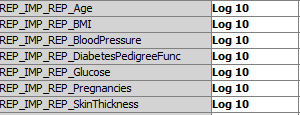
Figure Number of replacements

# The model

After the data cleaning, the data is splitting in training with 612 observation (representing 80%) and validation with 156 observation (20% of the data). The purpose of split the data is avoid overfitting or underfitting.



The following variables were transformed to log10: Age, BMI, BloodPressure, DiabetesPedigree Function, Glucose, Pregnancies, Skin Thickness. The transformation was made to decrease the skewness of the variables.



# Regression Model

The multiple regression model is based on the following equation:

𝑦 = 𝛽0 + 𝛽1x1 + 𝛽2x2 + … + 𝛽pxp

For this model, the variables are the following:

* Dependent variable is Outcome (Y).
* Independent variables (in their LG10 version): Age, BMI, Bloodpressure, DiabetesPedigreeFunction, Glucose, Pregnancies, SkinThickness, and insulin.

For the regression model is was choosed **stepwise** selection method. The stepwise selection method ends when either of these condictions is met:

* No variables outside the model has a significant F statistic, and every variable in the model is significant at the significance level that is specified in the model.
* The variables to be added to the model is the variable that was just deleted from it.

For this model the significance level is α=0.05.

After running the regression model, the stepwise method ends at 6th step. The result consist in the following effects:

Intercept LG10\_REP\_IMP\_REP\_BMI LG10\_REP\_IMP\_REP\_DiabetesPedigre LG10\_REP\_IMP\_REP\_Glucose LG10\_REP\_IMP\_REP\_Pregnancies M\_Insulin M\_REP\_Age

The equation based on the result of the stepwise selection is the following:

Outcome = 𝛽0 + 𝛽1 LG10\_REP\_IMP\_REP\_BMI + 𝛽2 LG10\_REP\_IMP\_REP\_DiabetesPedigre + 𝛽3 LG10\_REP\_IMP\_REP\_Glucose + 𝛽4 LG10\_REP\_IMP\_REP\_Pregnancies + 𝛽5 M\_Insulin + 𝛽6 M\_REP\_Age

The following graph shows the lines for training and validation data. The lines follow same tendency, and they are closed.

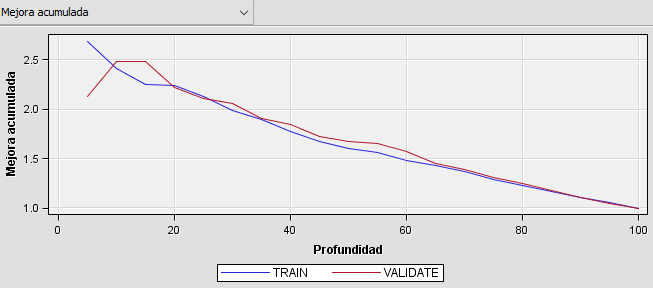


Figure Graph after run regression model stepwise method

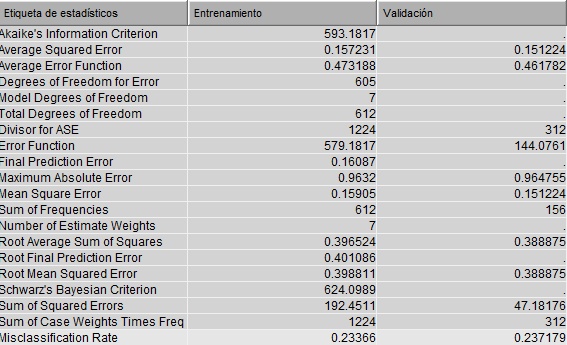


Figure Statistical output

Information to be highligh about the table above:

* The smaller MSE closer to finding the line of best fit.
* The Sum of Squared Error is the difference between the observed value and the predicted value

The Null hypothesis is the following:

Null hypothesis H0: 𝛽1 = 𝛽2= 𝛽3= 𝛽4= 𝛽5= 𝛽6=0

The Null hypothesis is rejected, and the coefficients are the following:

* The intercept is -36.1802
* The regression coefficient for Log10\_Rep\_Imp\_rep\_BMI is 7.2348
* The regression coefficient for Log10\_rep\_imp\_rep\_diabetespedigre is 5.5630
* The regression coefficient for Log10\_rep\_imp\_rep\_glucose is 10.8427
* The regression coefficient for Log10\_Rep\_Imp\_rep\_pregnancies is 1.3148
* The regression coefficient for M\_insulin is 0.5492
* The regression coefficient for M\_rep\_age is -0.4257.

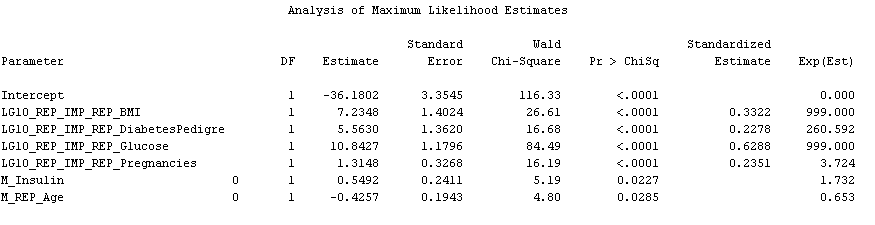


Figure Analysis of Maximum Likelihood Estimates

Only 6 regression variables from 9 seen to be statiscaly significant at an alpha of 0.05. Those variables are BMI, DiabetesPedigree, Glucose, Pregnancies, Insulin and Age.

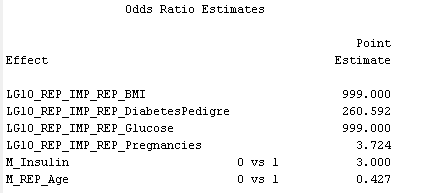


Figure Odds Ratio Estimates

Higher the Point estimate, higher the likelihood to have an outcome of 1, meaning the person are more likely to have diabetes.

# Decision tree

Accordingly with decision tree model, the three most significat variables for the outcome are Glucose, Age and BMI.

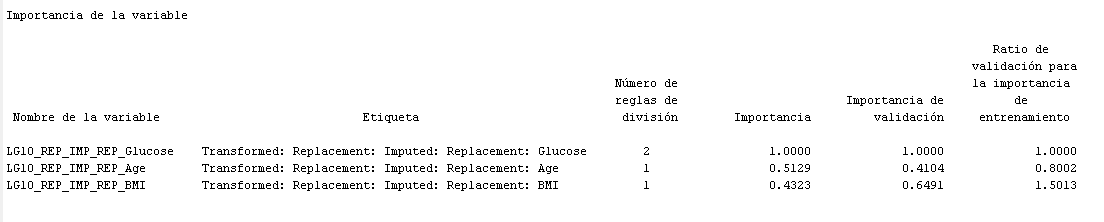


Figure Decision tree- Variable importance

There are 5 leaf node in the tree. This tree is automatically pruned to an optimal size. Therefore, the node numbers that appear in the final tree are not sequential. They reflect the positions of the nodes in the full tree, before pruning.

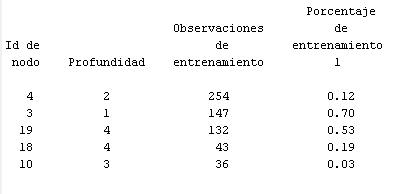


Figure leaf node

The following table states that 34.80% of the training data is for 1 (diabetes). And, 35.26% of the validation is for 1.

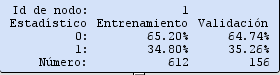


Figure Percentage of 0 and 1

Related to the statistical information, the misclassification rate for validation is lower than training (0.224 vs 0.24 respectively). SSE for validation is lower than training (53.287 vs. 196.82 respectively).

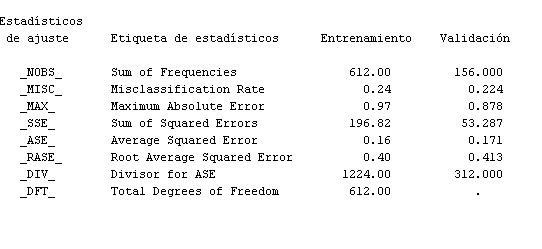


Figure Statistical information

# Comparison models

The KNN model has been included to compare the regression and decision tree explained in this project.

Comparing the values for the three models (regression, decision tree and KNN), it is possible to conclude that they are very similar predictors. The regression model has the lower MSE for training (0.15723) and validation (0.15122), followed by decision tree and KNN. The misclassification rate is lower for the regression model in training data (0.23366) but higher in validation data (0.23718).

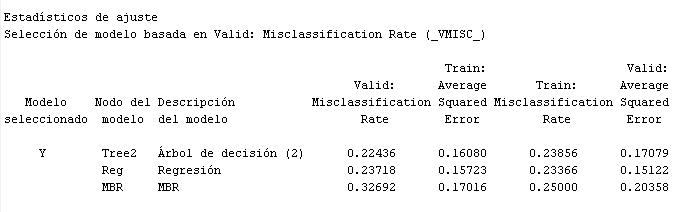


Figure Comparison models

The confusion matrix shows that regression model has more TN for training and validation data. Decision tree has more TP for training and validation data. Regarding FN, decision tree shows lower counts FN than regression and KNN.

FN

TN

TP

FP



Figure Confusion matrix

# Conclusion:

For regression model patients with diabetes pedigree, more number of pregnancies, high level of glucose, high BMI, and high level of insulin are more likely to have diabetes. Regarding the variable Age is also important for the model but in less proportion than the mentioned before.

For the decision tree, the most relevant variables are Age, Glucose and BMI.

Comparing the models, it is possible to conclude that the regression model has more TN for training and validation data. The decision tree has more TP for training and validation data. Regarding FN, the decision tree shows lower counts FN than regression and KNN.