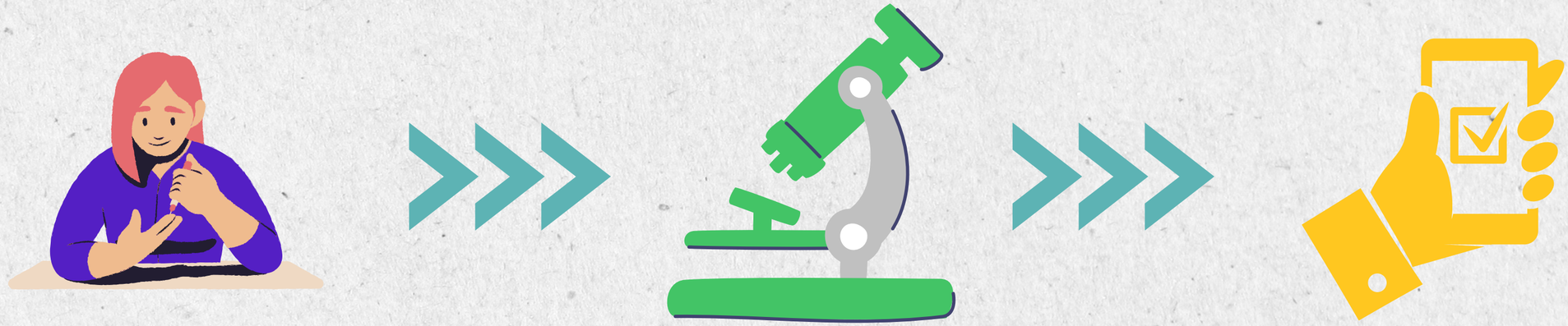
Diabetes Predictive Model



Feras Rafeh - Iranel González - Bruna Santos

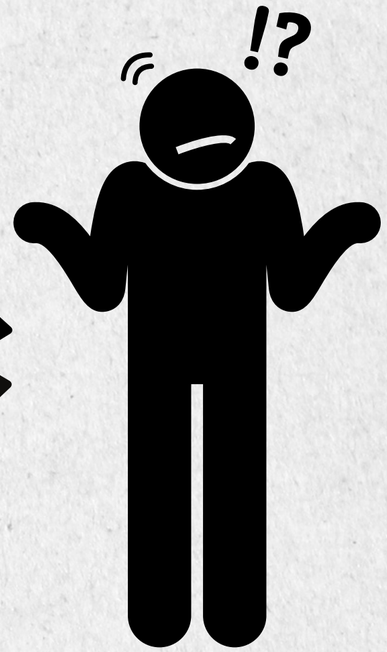
Agenda

1. Problem 
2. Data Introduction 
3. Data Journey 
4. Getting into the insights 
5. Conclusions 

The problem

37.3 million → 1 in 10 → **1 in 5**

96 million → 1 in 3 Prediabetes → **8 in 10** Prediabetes



Reference: https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html#:~:text=37.3%20million%20Americans%E2%80%94about%201,t%20know%20they%20have%20it.

The problem

Diabetes

Reference: https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html#:~:text=37.3%20million%20Americans%E2%80%94about%201,t%20know%20they%20have%20it.

Data Introduction

1 Data

The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative).

Data Shape: (100.000, 9) kaggle

2 Features

Medical Data

- Hypertension
- Heart disease
- Body mass index (BMI)
- Glycosylated hemoglobin
- Blood glucose level

Demographic Data

- Age
- Gender
- Smoking history

3 Objectives

🎯 Developing a predictive model to identify patients at risk of developing Diabetes based on their medical history and demographic information.

🎯 Analyze the impact and importance of the features (independent variables) used to identify the main factors that increase the risk to develop Diabetes.

Data Journey

Cleaning

- Changing Data type in "Age" column
- Droped ages from 0 to 3
- Droped "No info" values in "smoking history" (31964 rows)
- Merge some data from "smoking history"

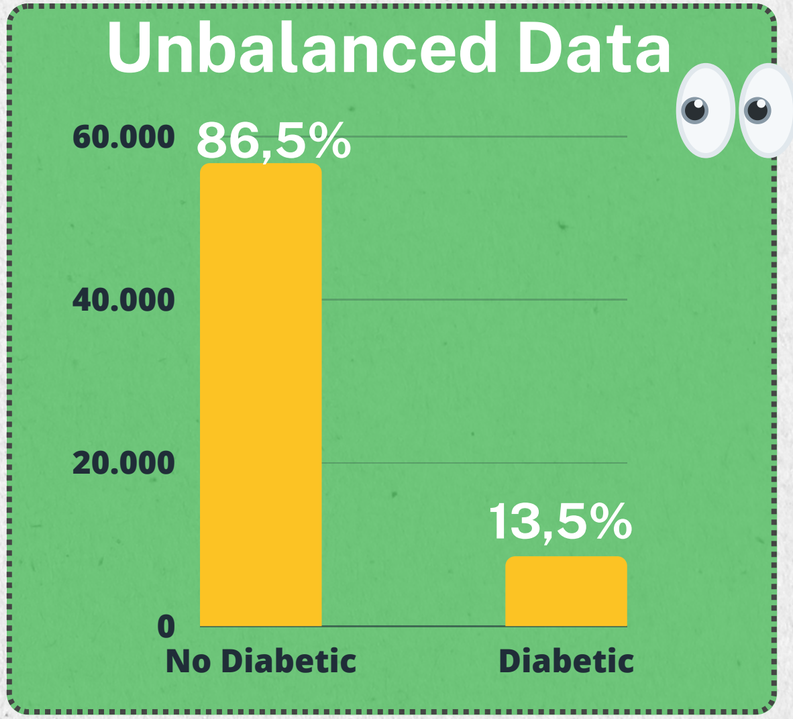
Data Journey

Cleaning

- Changing Data type in "Age" column
- Droped ages from 0 to 3
- Droped "No info" values in "smoking history" (31964 rows)
- Merge some data from "smoking history"

Explore

- The data was unbalanced: 56,665 negative results (no diabetes, 86.5% of our data) and 8,500 positive results, 13.5% of our data



Data Journey

Cleaning

- Changing Data type in "Age" column
- Droped "No info" values in "smoking history"
- Merge some data from "smoking history"
- Droped ages from 0 to 3

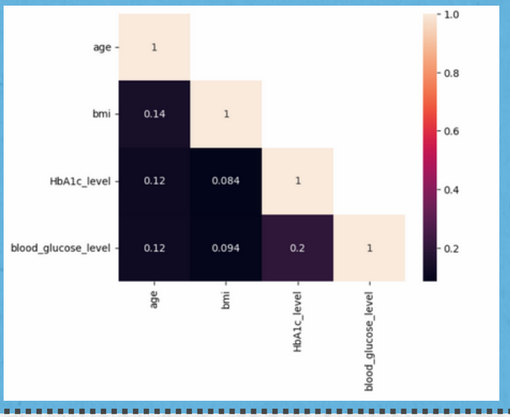
Explore

- The data was unbalanced: 56,665 negative results (no diabetes) and 8,500 positive results)

Process

- Checking Correlation between independent variable
- Data transformation into numerical and categorical
- Split data into train and test data
- Scaler data (using min and max)
- Handle Imbalanced data using (Synthetic Minority Over-sampling Technique (SMOTE))

Heatmap



Data Journey

Cleaning

- Changing Data type in "Age" column
- Droped "No info" values in "smoking history"
- Merge some data from "smoking history"
- Droped ages from 0 to 3

Explore

- The data was unbalanced: 56,665 negative results (no diabetes) and 8,500 positive results)

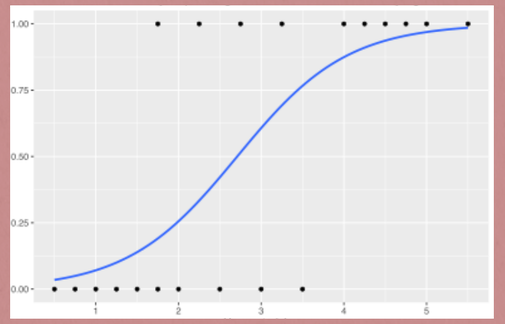
Process

- Checking Correlation between independent variable
- Data transformation into numerical and categorical
- Split data
- Scaler data (using min and max)
- Handle Imbalanced data using (Synthetic Minority Over-sampling Technique (SMOTE))

Apply Model

- Logistic Regression Model
- Model Accuracy: 88.27%

Logistic Regression Model



Data Journey

Cleaning

- Changing Data type in "Age" column
- Droped "No info" values in "smoking history"
- Merge some data from "smoking history"
- Droped ages from 0 to 3

Explore

- The data was unbalanced: 56,665 negative results (no diabetes) and 8,500 positive results)

Process

- Checking Correlation between independent variable
- Data transformation into numerical and categorical
- Split data
- Scaler data (using min and max)
- Handle Imbalanced data using (Synthetic Minority Over-sampling Technique (SMOTE))

Apply Model

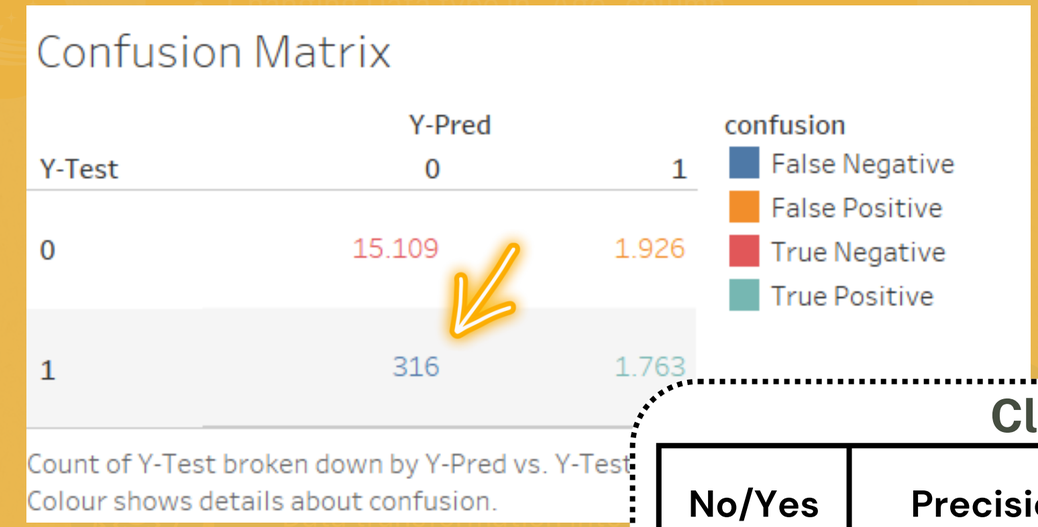
- Logistic Regression Model
- Model Accuracy: 88.27%

Validate

- Classification Report
- Confusion Matrix

Data Journey

Confusion Matrix



Classification Report

No/Yes	Precision	Recall	F1-score	Support
0	0,98	0,89	0,93	17035
1	0,48	0,85	0,61	2079

- Split data
- Scaler data (using min an
- Handle Imbalanced data
  Minority Over-sampling T

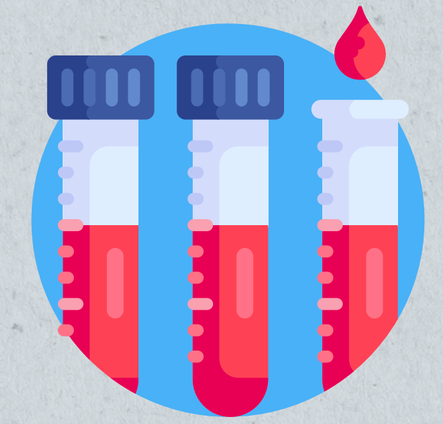
Apply Model

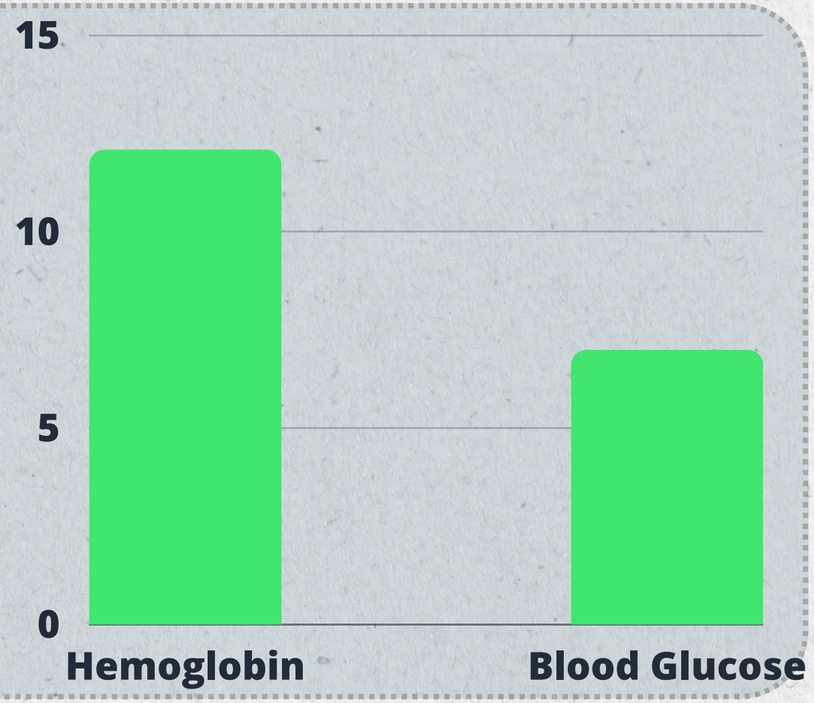
- Logistic Regression Model
- Model Accuracy: 88.27%

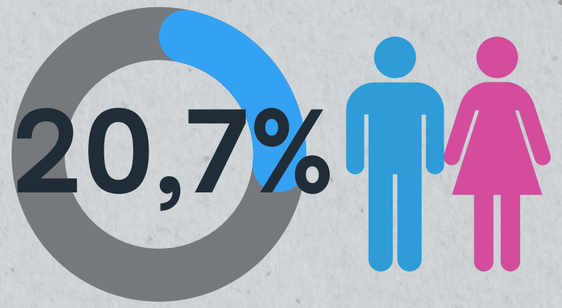
Validate

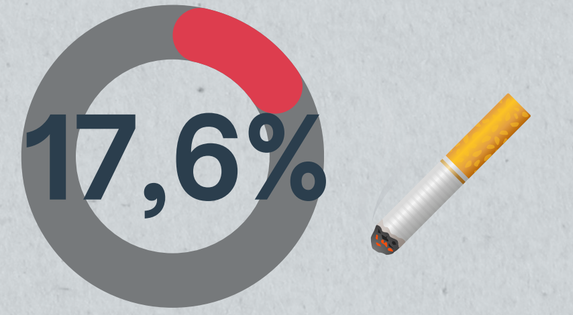
- Classification Report
- Confusion Matrix

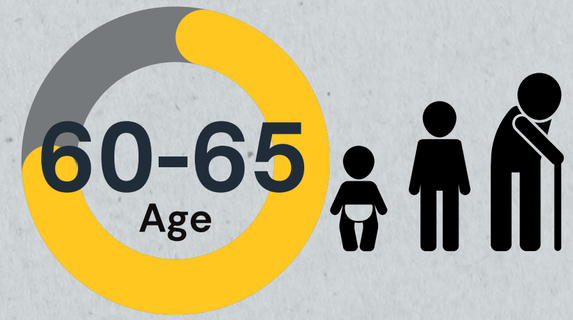
Getting into the insights











- Overweight 25.0 - 29.9
- Obesity 1st Class 30.0 - 34.9
- Obesity 2nd Class 35.0 - 39.9

Coefficient: 8.39

Let's see our model in action

Conclusions

We developed a predictive model that identifies patients at risk of developing diabetes based on their medical history and demographic information.

Model Accuracy: 88.27%

Performance Validation: 316 FN

We analyzed the impact and importance that our independent variables have in detecting patients at risk of developing diabetes.

- Glycosylated hemoglobin
- BMI
- Blood glucose

- Age

60–65

- Gender
- Smoking Story

Thank you for your attention!

Find us on Slack

Ferass Rafeh - Iranel González - Bruna Santos