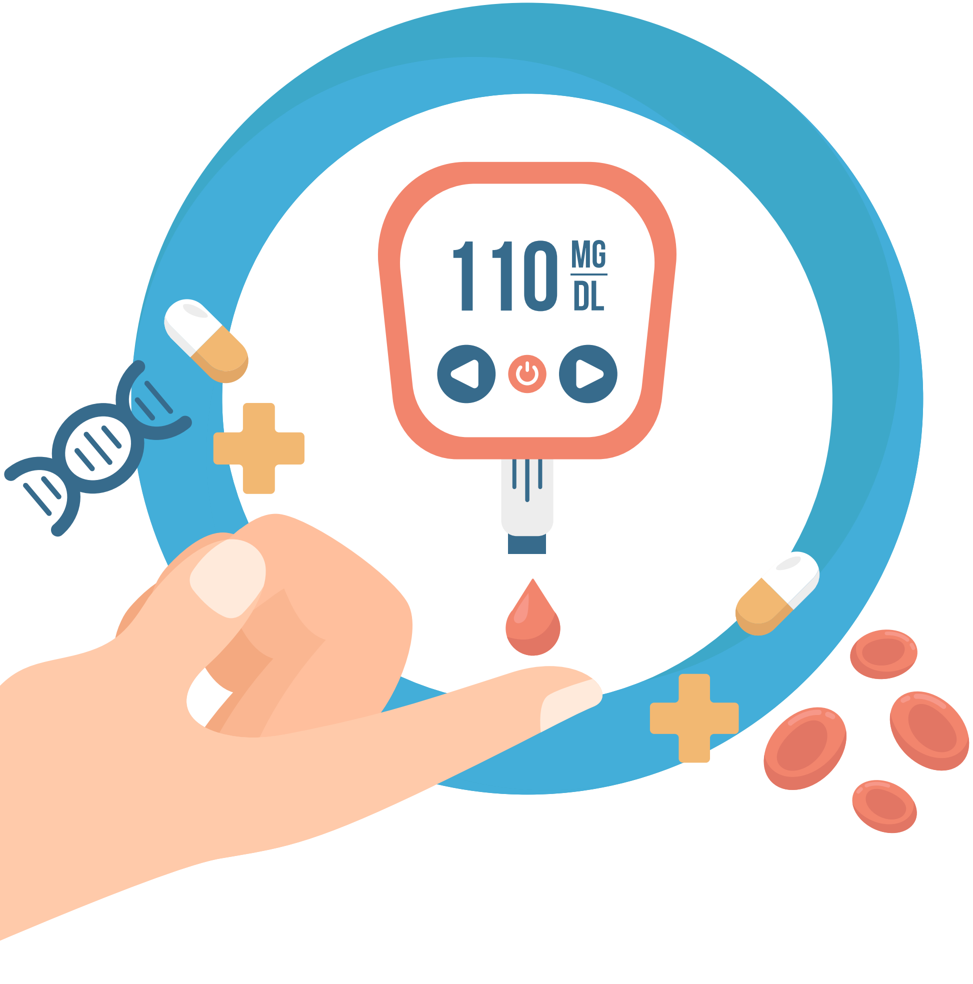
DigitalSkola

# Final Project Data Science

Fauzia Yumna Ayupuspita
Iqbal Muhammad
Muhammad Fahmi
Siti Rabiatul Adwiyah
Yoga Mahardika Sidiq



# Diabetes Prediction

# Outline

1. Overview About Diabetes
2. Problem Definition, Goal, Methodology
3. Result Overview
4. EDA
5. Machine Learning
6. Conclusion

# Overview about Diabetes

## What is Diabetes?

Disease that occurs when your blood glucose is too high

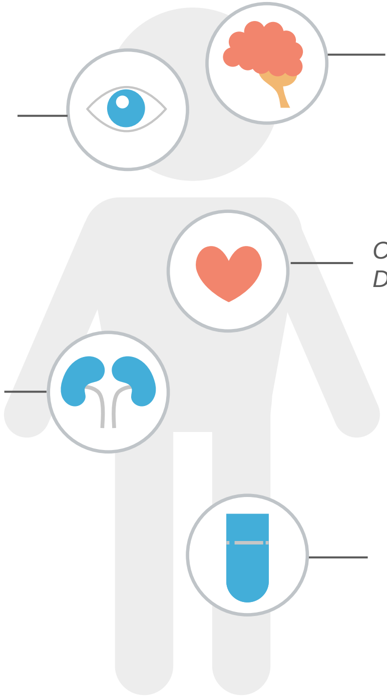
## Symptoms

Thirst

Fatigue

Weight loss

## Organs Affected



Stroke

Damage to the eyes

Cardiovascular Disease

Chronic kidney disease

Foot Ulcers

# Problem, Goals, Methodology

## Problem

- Healthcare sector have large amount databases
- Existing method for diabetes detection is uses lab tests
- Existing Method is time consuming

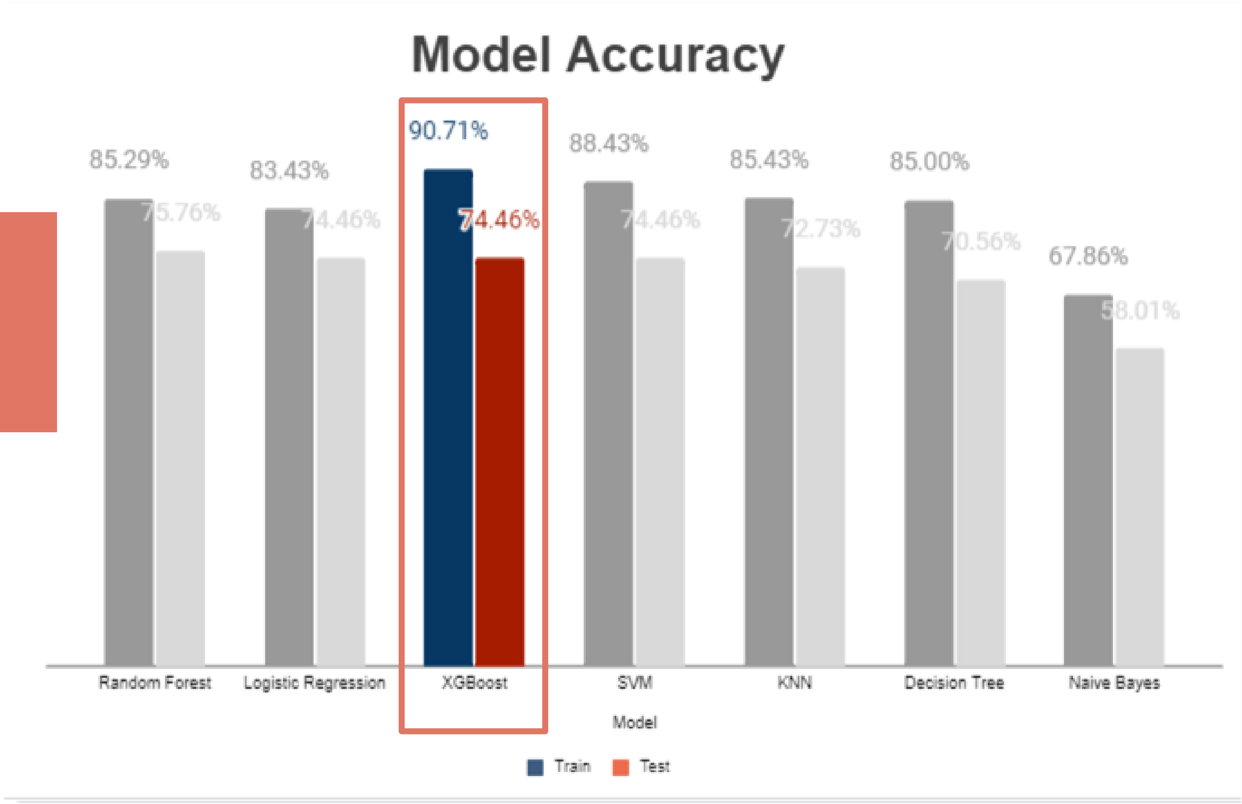
## Goals

Building predictive model for diabetes prediction so that indicated patient can check further as soon as possible

## Methodology

Using Machine Learning - Supervised

# Result Overview



The best model is XGBoost with Accuracy Test is 74.46%

# Methodology

01 Database

Extracting dataset from database *)

02 EDA

Understanding data from univariate, bivariate even multivariate analysis

03 Data Preprocessing

Feature Engineering (Classifying age, Glucose, BMI, Blood Pressure, Insulin and Pregnancies

04 Modelling

Splitting data, model training & model testing, accuracy

05 Result

Diabetes Prediction Model

*) *Source: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database*

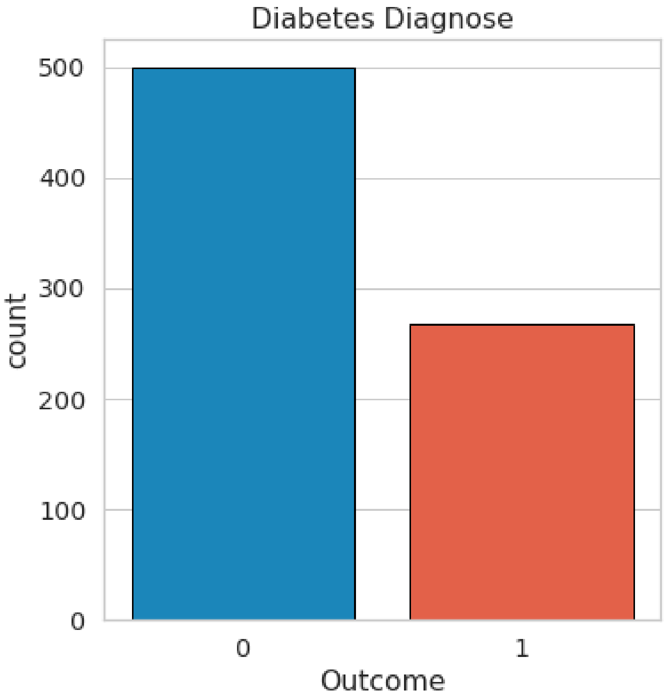
# Exploratory Data Analysis (EDA)

Data Check

| No | Feature | IsNull |
|---|---|---|
| 1 | Pregnancies | 0 |
| 2 | Glucose | 5 |
| 3 | Blood Pressure | 227 |
| 4 | Skin Thickness | 374 |
| 5 | BMI | 11 |
| 6 | Diabetes Pedigree Function | 0 |
| 7 | Age | 0 |
| 8 | Outcome | 0 |

# Exploratory Data Analysis (EDA)

Univariate - Categorical



## Diabetic

**268 or 35%  patients are diagnosed diabetes**

## Non Diabetic

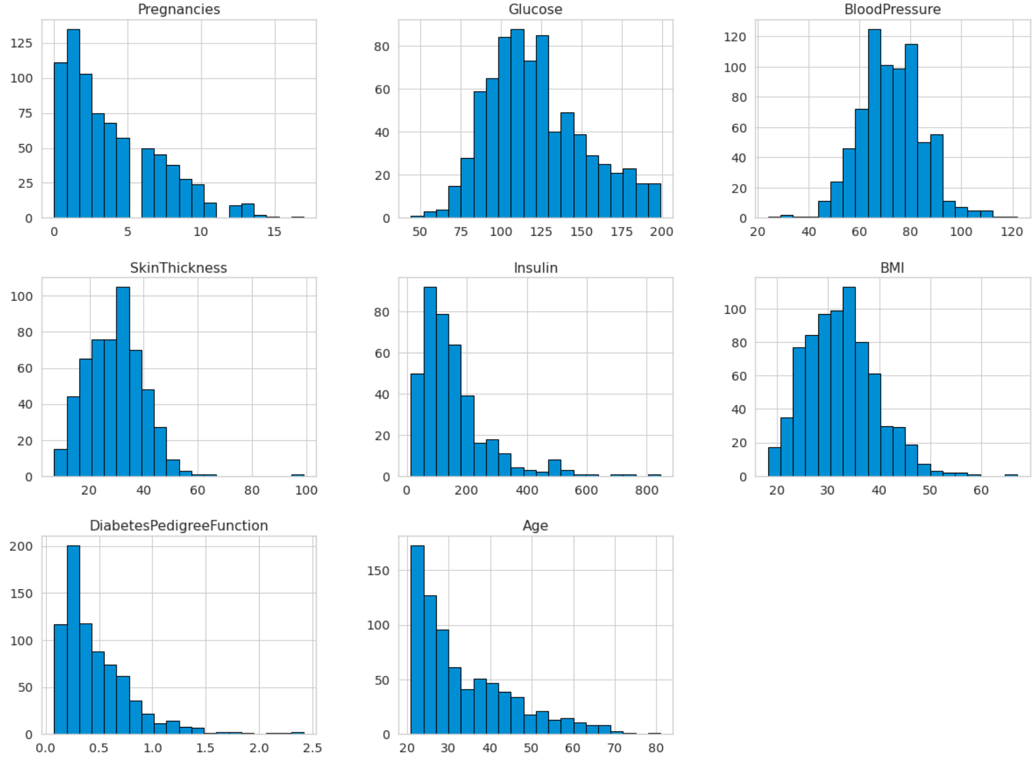
**500 or 65% patients are Non Diabetic**

## Comparison

**The comparison between diabetic and non diabetic is 1:2**

# Exploratory Data Analysis (EDA)

Univariate - Numerical



## Normal Distribution

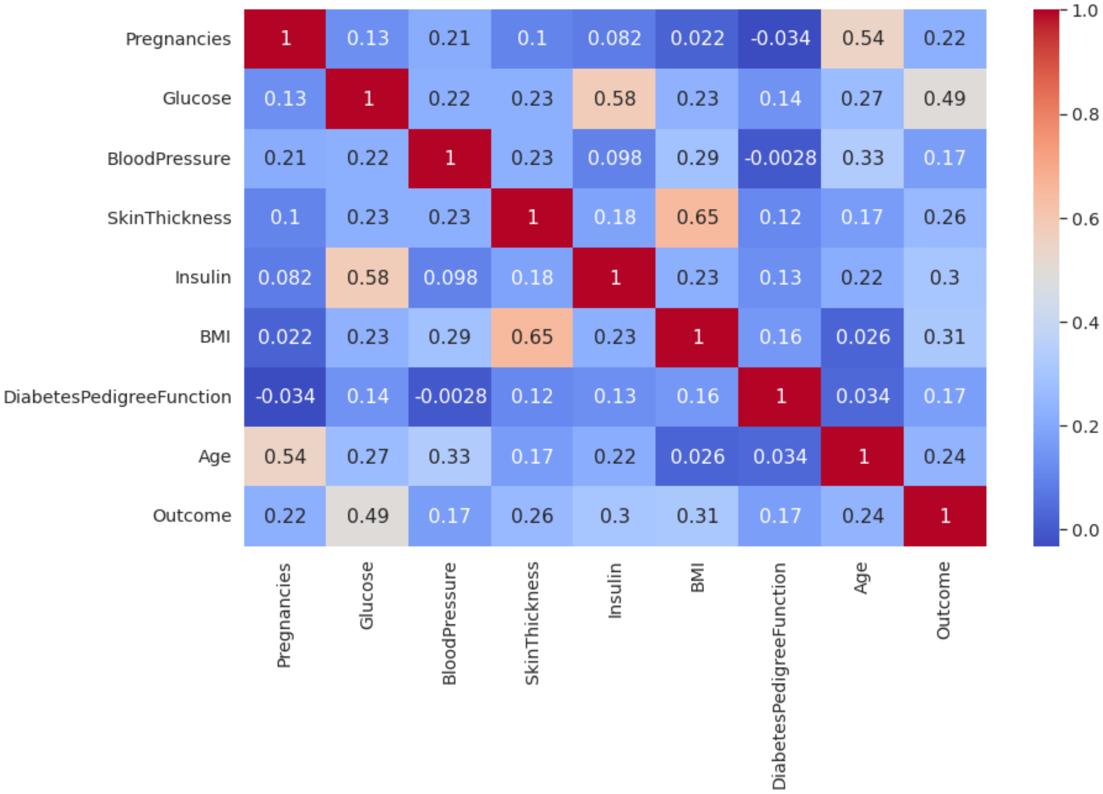
- Glucose and
- Blood Pressure

## Positive Skewed

Most of data are positive skewed except **Glucose and Blood Pressure data**

# Exploratory Data Analysis (EDA)

Correlation Matrix



| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| Pregnancies | 1 | 0.13 | 0.21 | 0.1 | 0.082 | 0.022 | -0.034 | 0.54 | 0.22 |
| Glucose | 0.13 | 1 | 0.22 | 0.23 | 0.58 | 0.23 | 0.14 | 0.27 | 0.49 |
| BloodPressure | 0.21 | 0.22 | 1 | 0.23 | 0.098 | 0.29 | -0.0028 | 0.33 | 0.17 |
| SkinThickness | 0.1 | 0.23 | 0.23 | 1 | 0.18 | 0.65 | 0.12 | 0.17 | 0.26 |
| Insulin | 0.082 | 0.58 | 0.098 | 0.18 | 1 | 0.23 | 0.13 | 0.22 | 0.3 |
| BMI | 0.022 | 0.23 | 0.29 | 0.65 | 0.23 | 1 | 0.16 | 0.026 | 0.31 |
| DiabetesPedigreeFunction | -0.034 | 0.14 | -0.0028 | 0.12 | 0.13 | 0.16 | 1 | 0.034 | 0.17 |
| Age | 0.54 | 0.27 | 0.33 | 0.17 | 0.22 | 0.026 | 0.034 | 1 | 0.24 |
| Outcome | 0.22 | 0.49 | 0.17 | 0.26 | 0.3 | 0.31 | 0.17 | 0.24 | 1 |

Features that have the highest correlation:

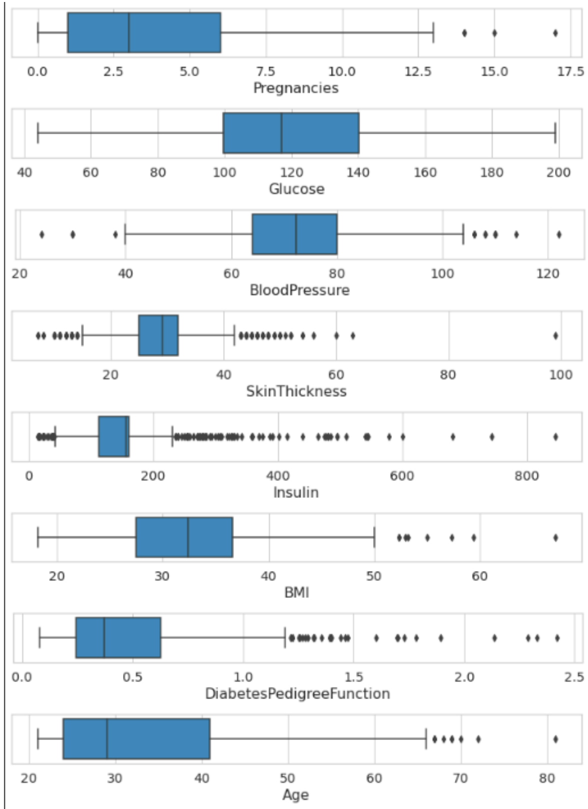
- **Glucose**
- **BMI**
- **Insulin**
- **Skin Thickness**
- **Age**
- **Pregnancies**

All features have correlation < 0.5

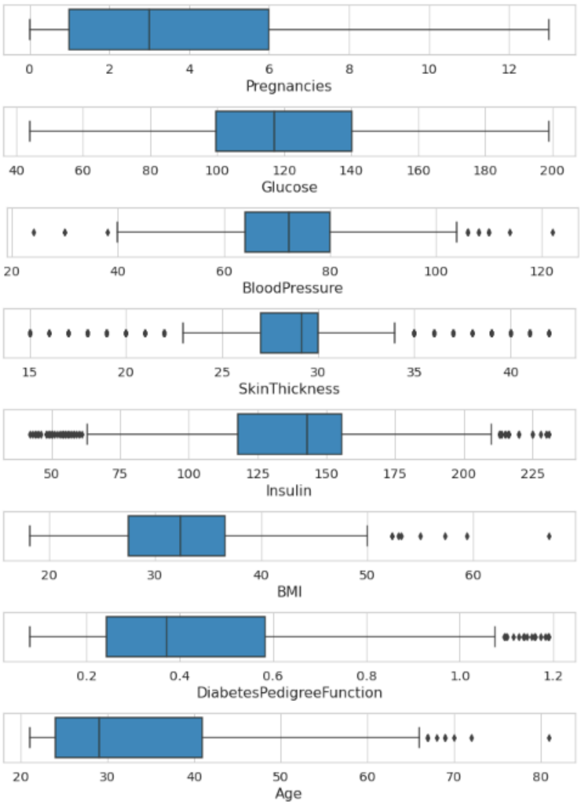
# Exploratory Data Analysis (EDA)

Handling Outlier

## Original Data

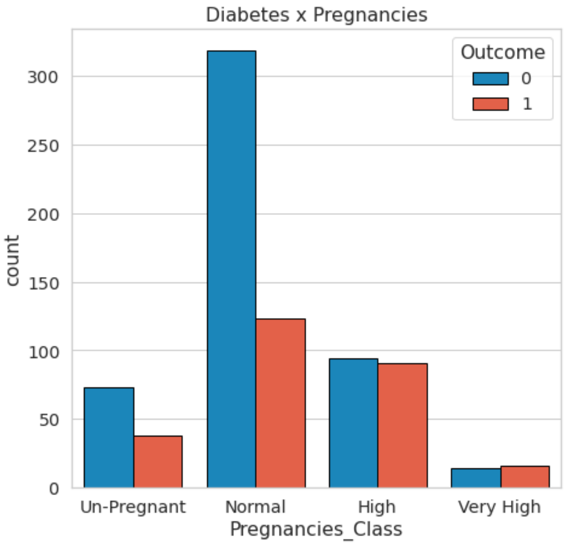


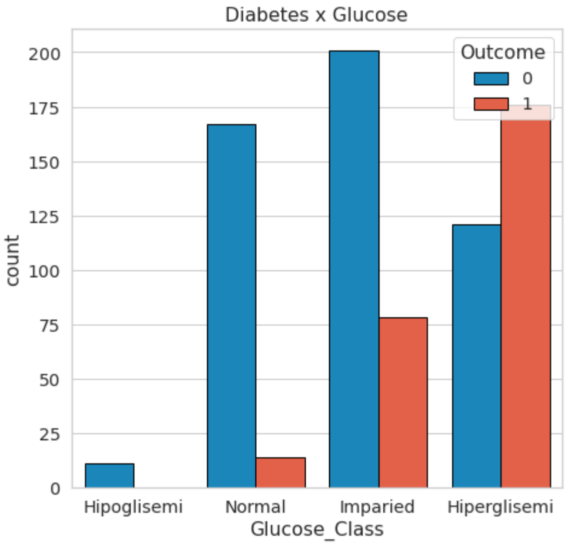
## After Handling Outlier with IQR

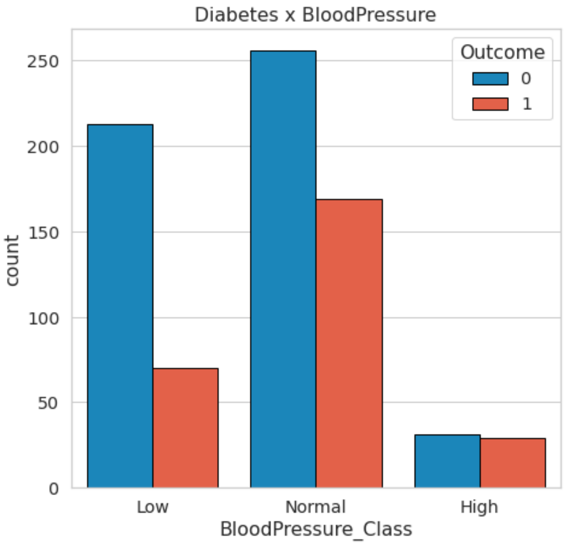


# Exploratory Data Analysis (EDA)

Bivariate – Categorical







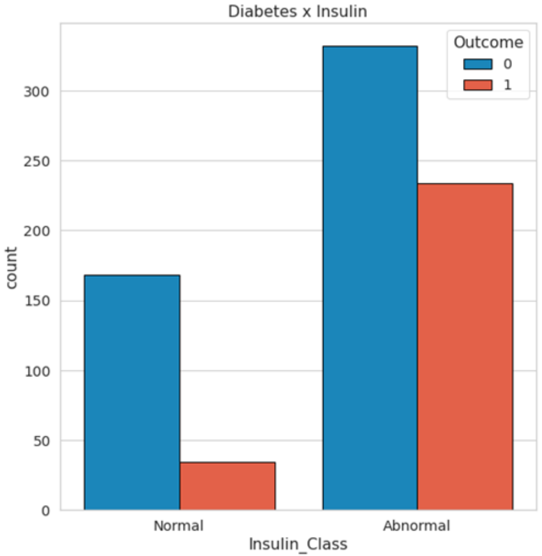
In proportion, **Very High** pregnancies class (>10 times pregnant) is the highest class that have diabetic

- Hipoglisemi Glucose is non diabetic
- Majority Hiperglisemi Class have diabetic

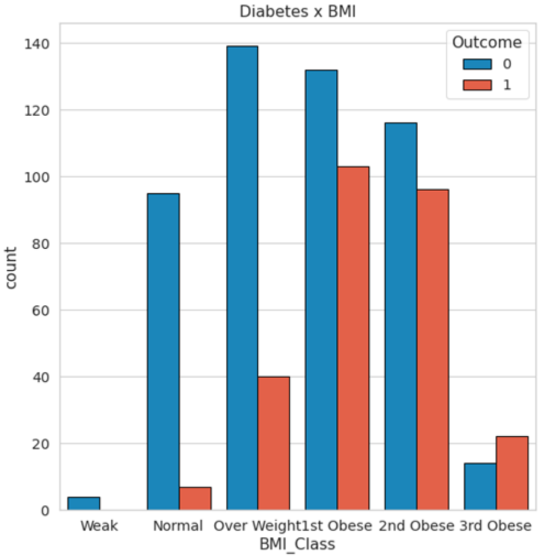
In proportion, **High Blood Pressure Class** (>90) is the highest class that have diabetic

# Exploratory Data Analysis (EDA)

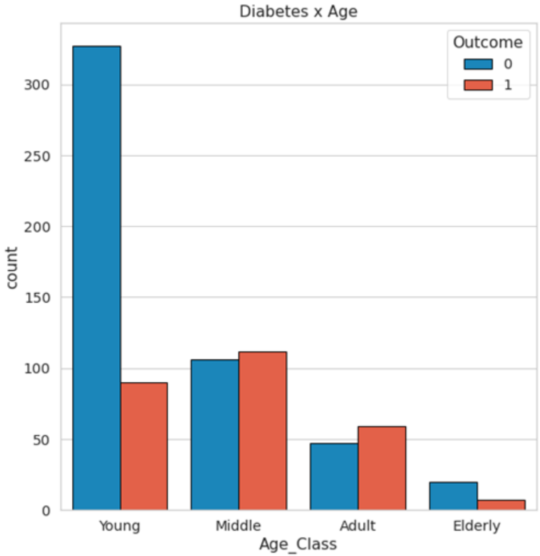
Bivariate – Categorical



In Average, patient with **Abnormal Insulin** class have diabetic

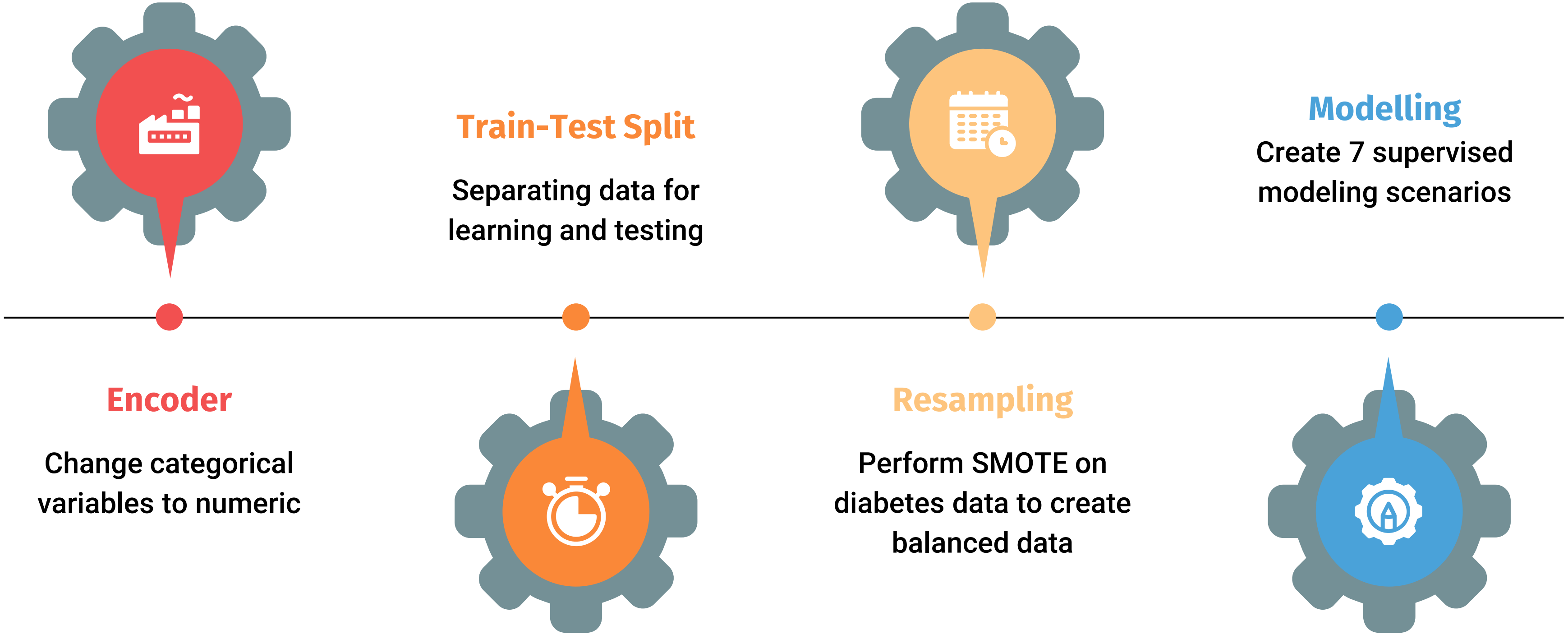


The higher the class of BMI then the more patients have diabetic



Middle Age Class is the highest class have diabetic

# Machine Learning



**Encoder**

Change categorical variables to numeric

**Train-Test Split**

Separating data for learning and testing

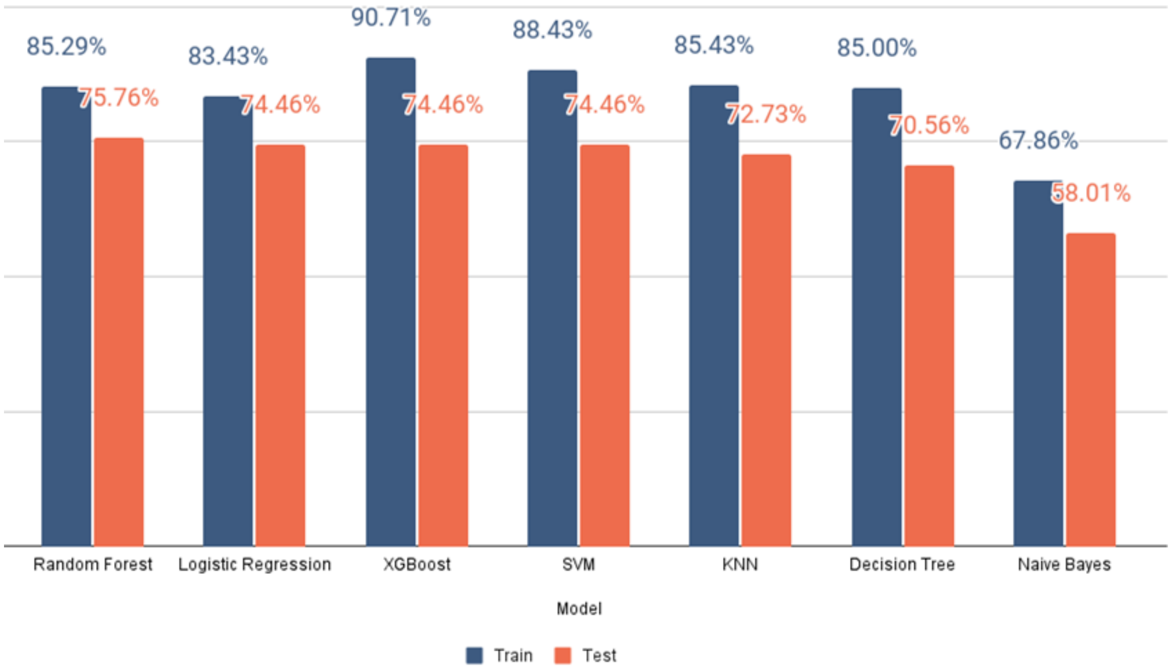
**Resampling**

Perform SMOTE on diabetes data to create balanced data

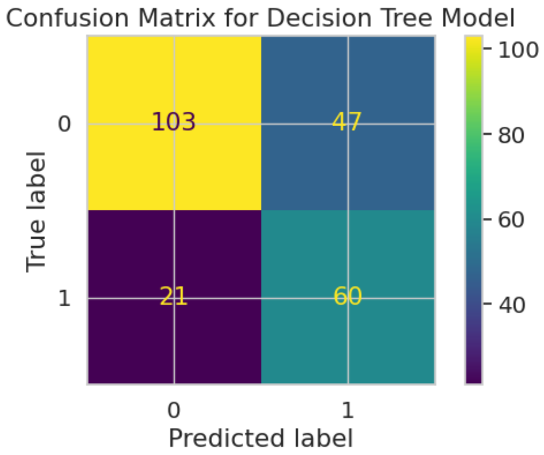
**Modelling**

Create 7 supervised modeling scenarios

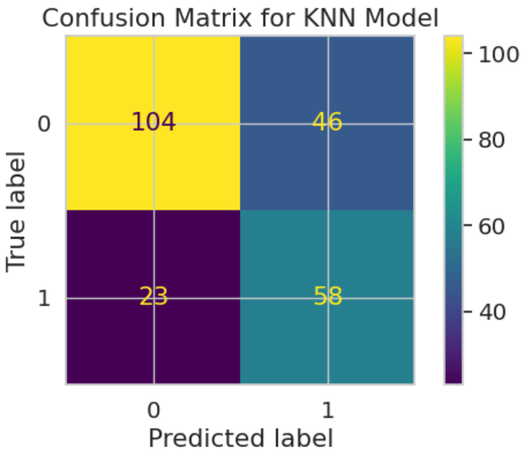
# Result Machine Learning Model



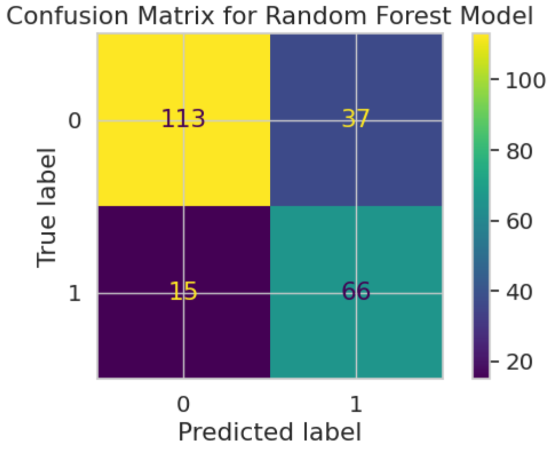
# Confusion Matrix Models



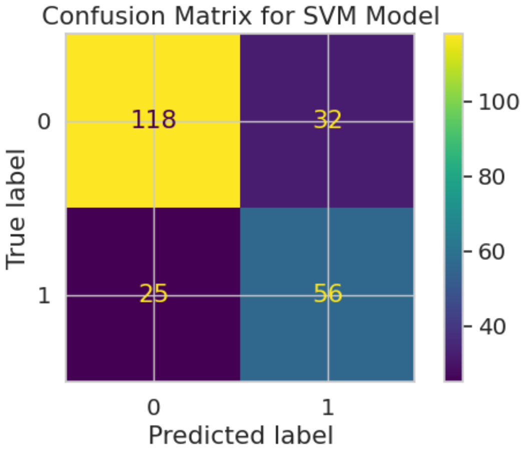
Confusion Matrix for Decision Tree Model



Confusion Matrix for KNN Model

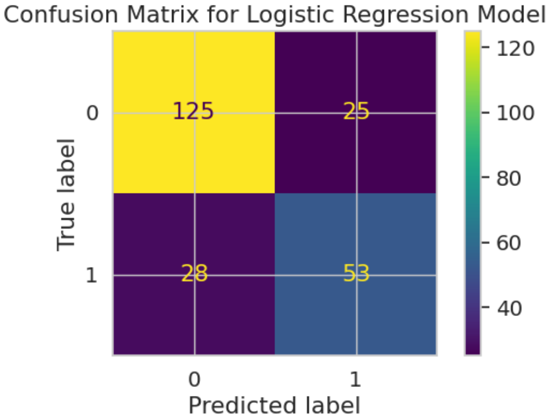


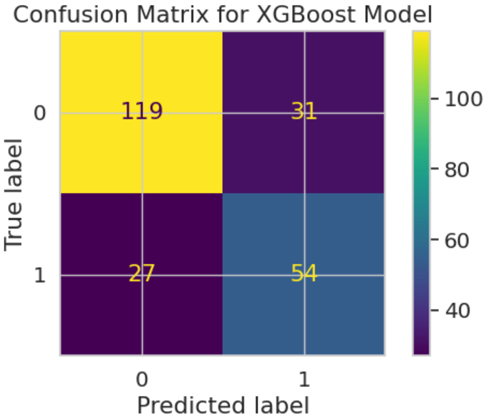
Confusion Matrix for Random Forest Model

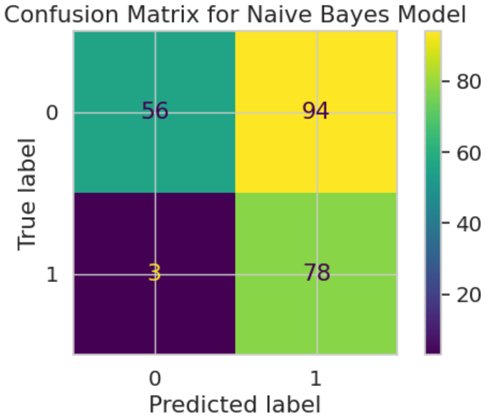


Confusion Matrix for SVM Model

# Confusion Matrix Models


Confusion Matrix for Logistic Regression Model


Confusion Matrix for XGBoost Model


Confusion Matrix for Naive Bayes Model

After looking at the confusion matrix for 7 models, we can draw the conclusion that **the best model to use is XGBoost**, considering the highest proportion of positive true & negative true values

# THANK YOU