

An analysis and machine learning project by B.Nedyalkov

Diabetes

Statistics and insights of diabetic patients (Pima
native American women)

With dataset supplied by Akshay D.
(<https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset>)
National Institute of Diabetes and Digestive and Kidney Diseases

Overview

Diabetes dataset

The data consists of:

- The of **clinical parameters** (Blood glucose, age, BMI, etc.)
- Of **768 women**
 - 268 of whom with type II diabetes
 - 500 without
- From **Pima Indian heritage**

Diabetes type II is a disease that stops the body from using insulin properly

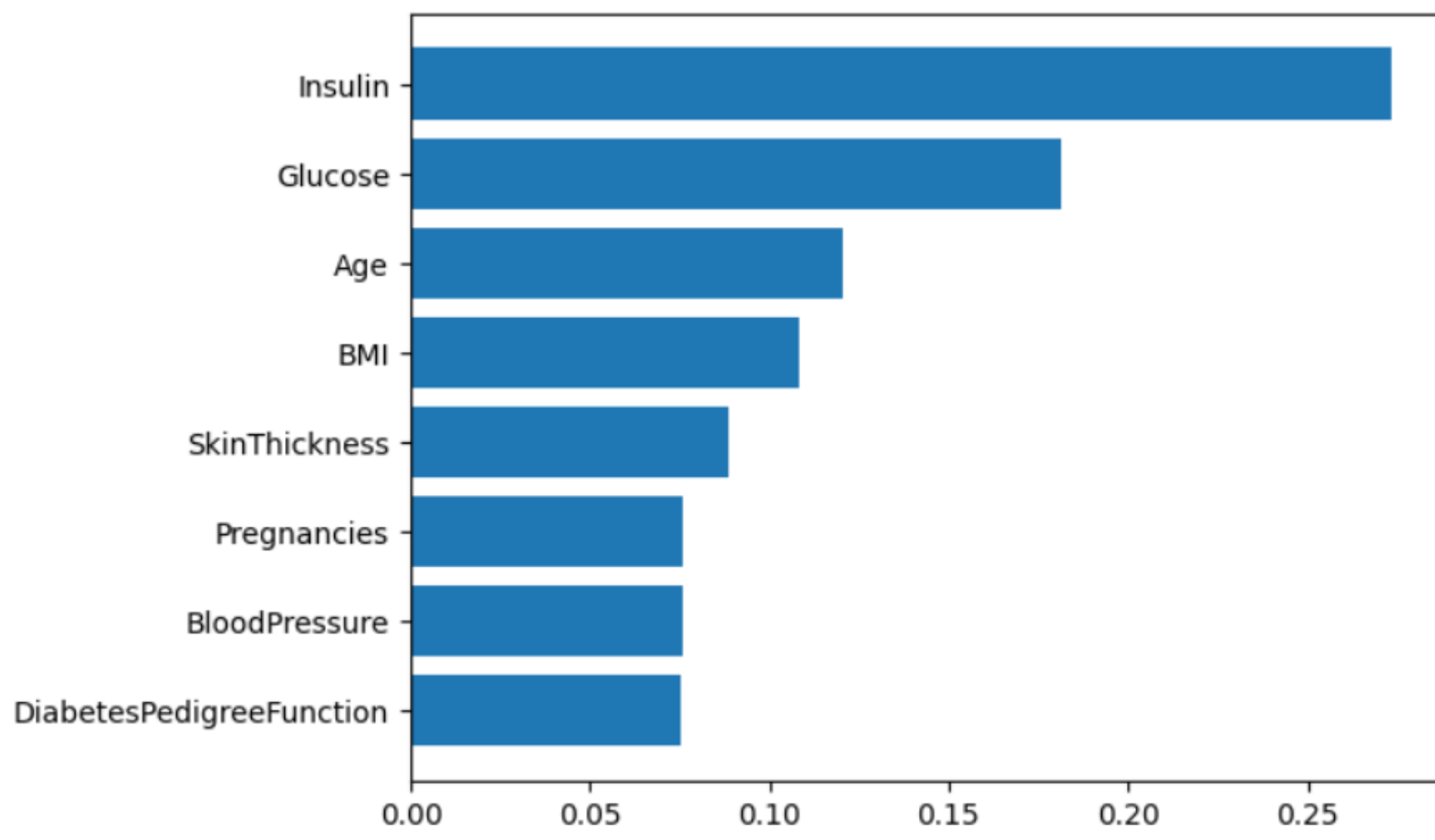


Overview

Diabetes facts

- Over time, type 2 diabetes can cause serious damage to the body, especially nerves and blood vessels
- Type 2 diabetes is often preventable
- Diabetes is a risk factor for many diseases and health complications
- It is the leading cause of blindness and amputation in adults
- Diabetics have at least 2 times the medical costs of someone without diabetes



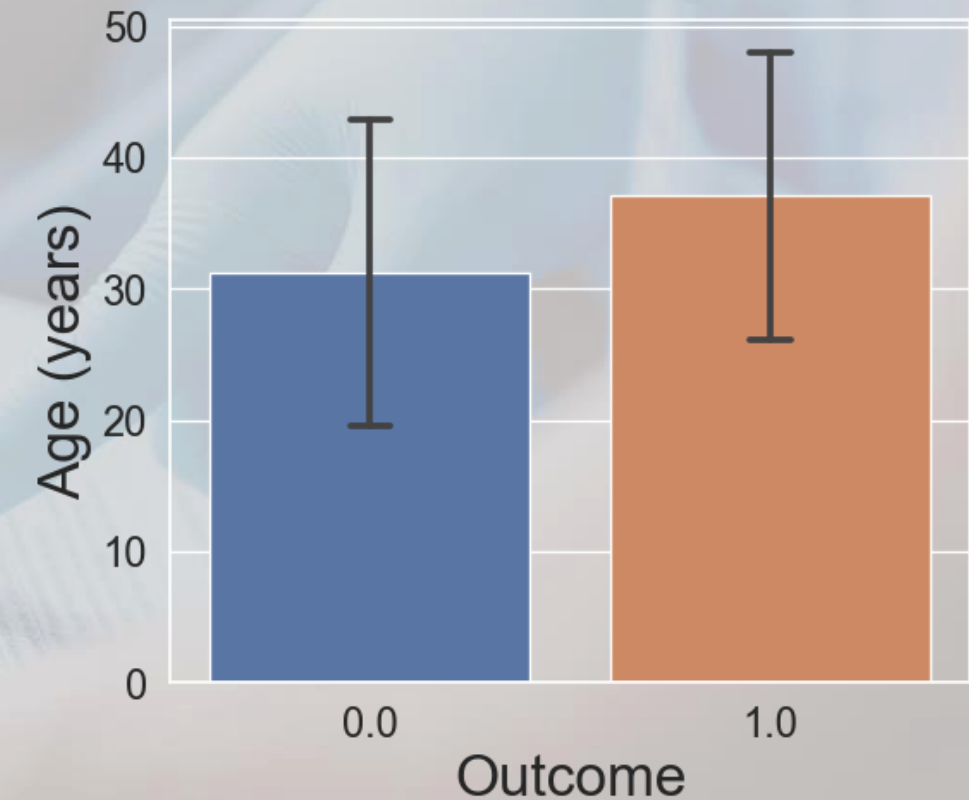
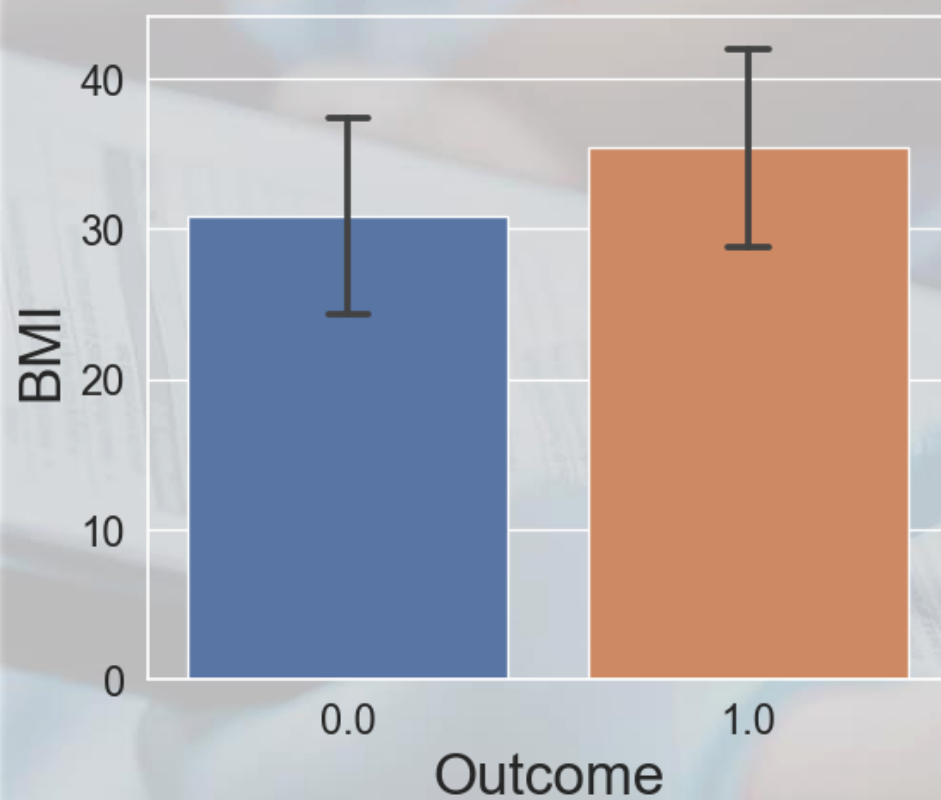
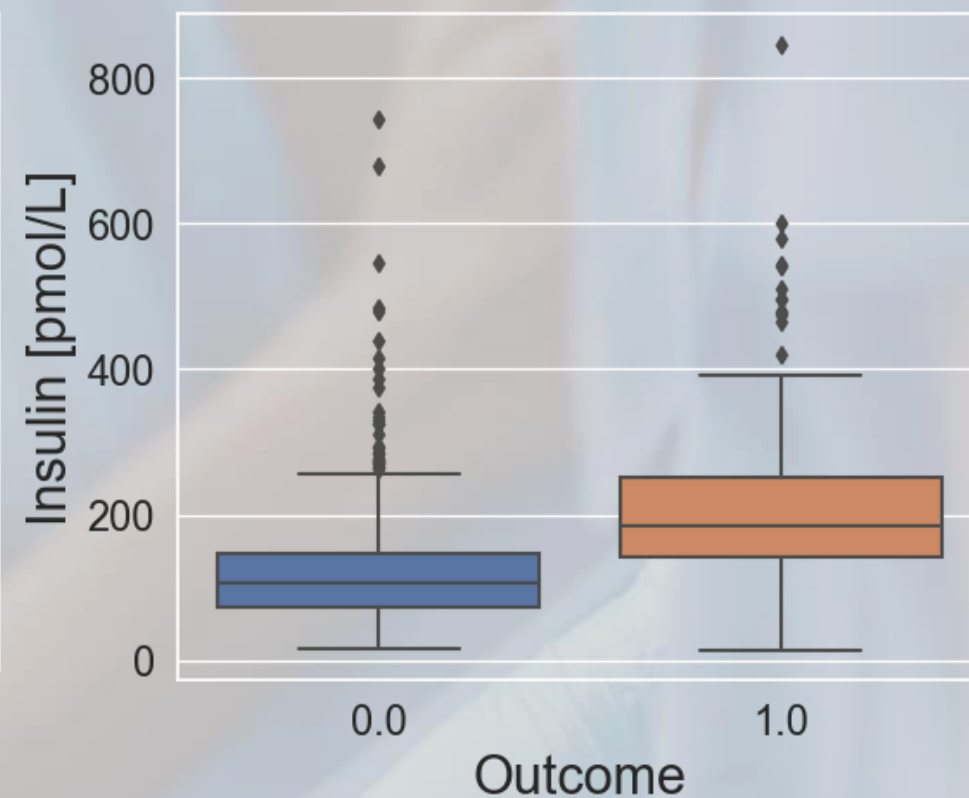
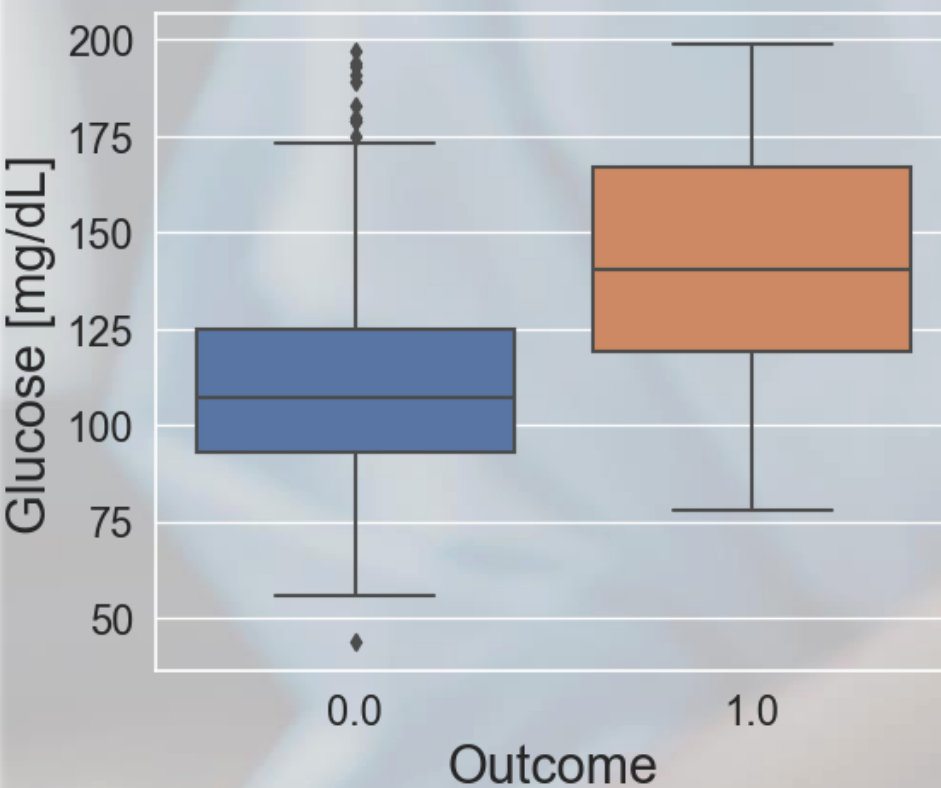


Analysis of the Data

Most important factors

- BMI
- Insulin levels
- Glucose levels (Blood sugar)
- Age



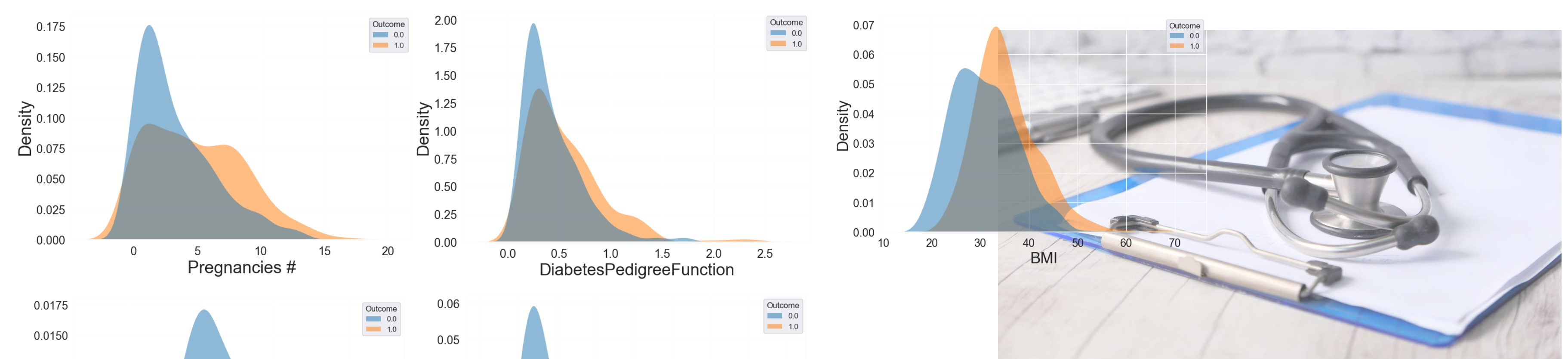


Analysis of the data

Most important factors

- BMI
- Insulin levels
- Glucose levels (Blood sugar)
- Age

0 = No Diabetes
1 = Has Diabetes



Insights from the data



- Women with more than 5 pregnancies are a risk group
- The diabetes pedigree has very limited impact below value = 1
- Glucose level is a major sign of diabetes
- Diabetes develops predominantly in women in their 30s to 50s
- Overweight individuals (BMI > 25) are much more likely to have diabetes

Distributions of the data

Analysis of the data

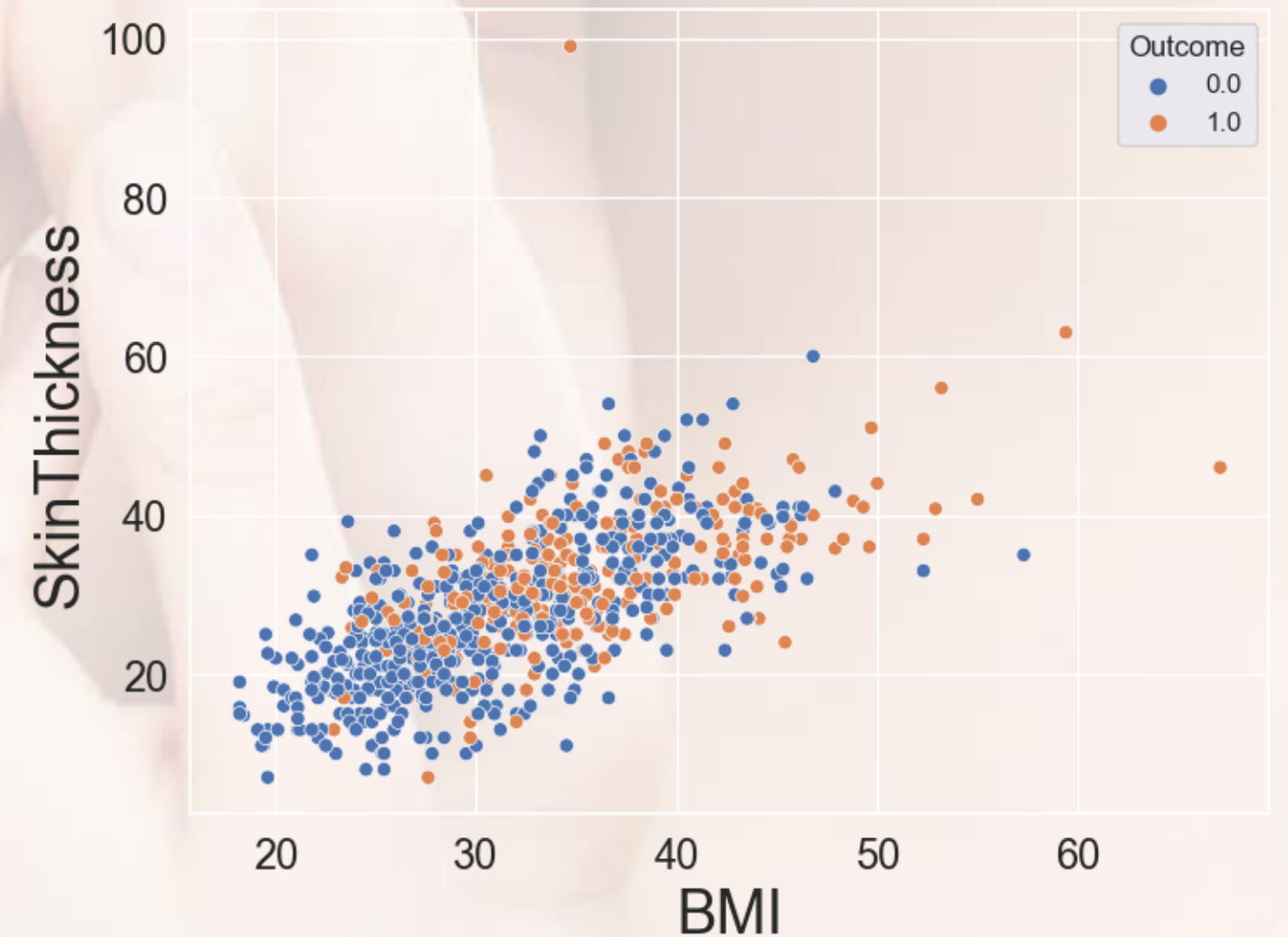
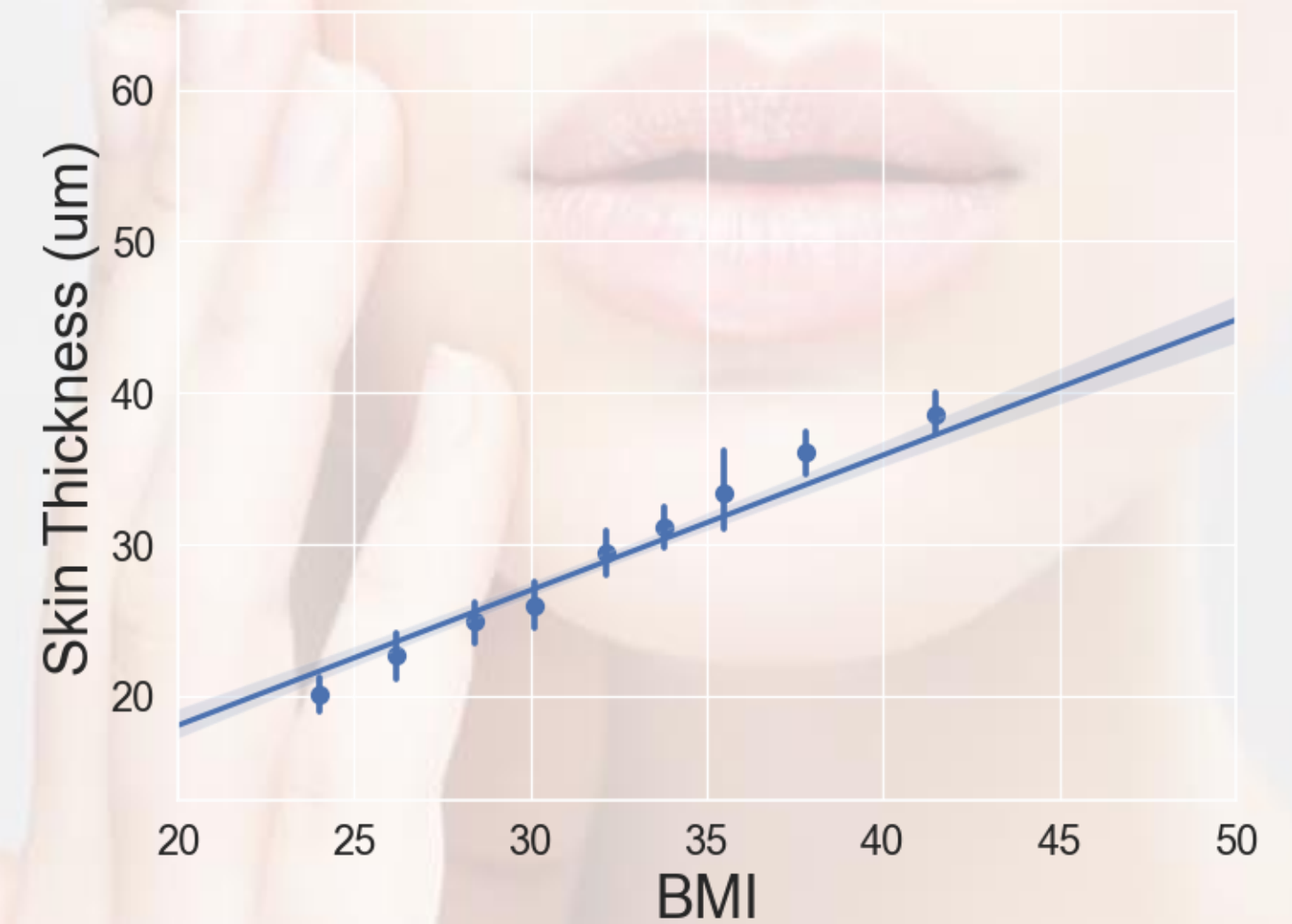
Curios findings

Skin Thickness vs BMI



Skin Thickness $\approx 1.1 \times \text{BMI} + 4.4$

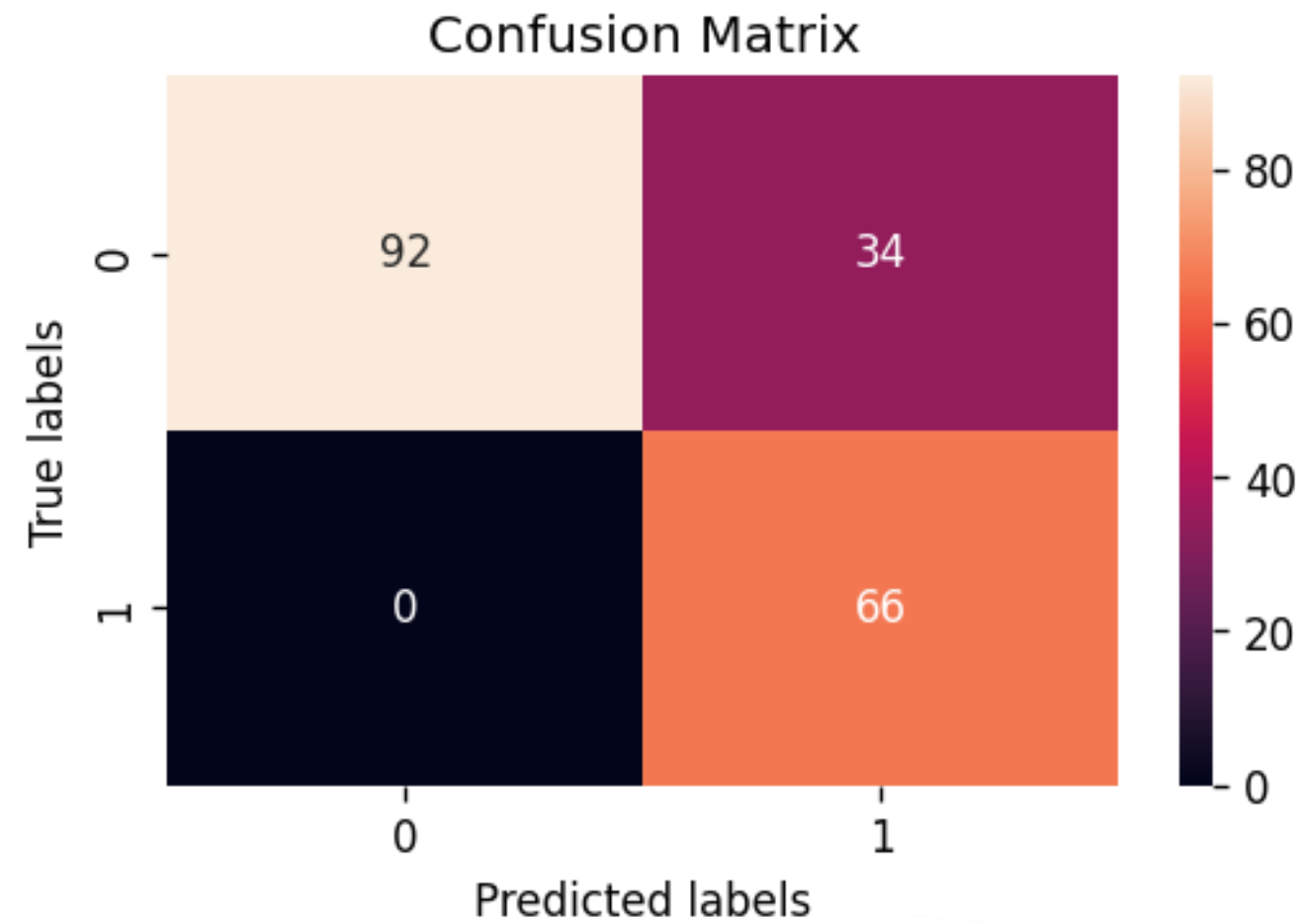
The thickness of the skin in is approximately linearly proportional to the Body mass index (aka. How much muscle and fat are there in one's body)



Machine Learning

The Model

By having 100% recall rate we can guarantee that 0% of the patients from the test set will come up as False Negatives



The models

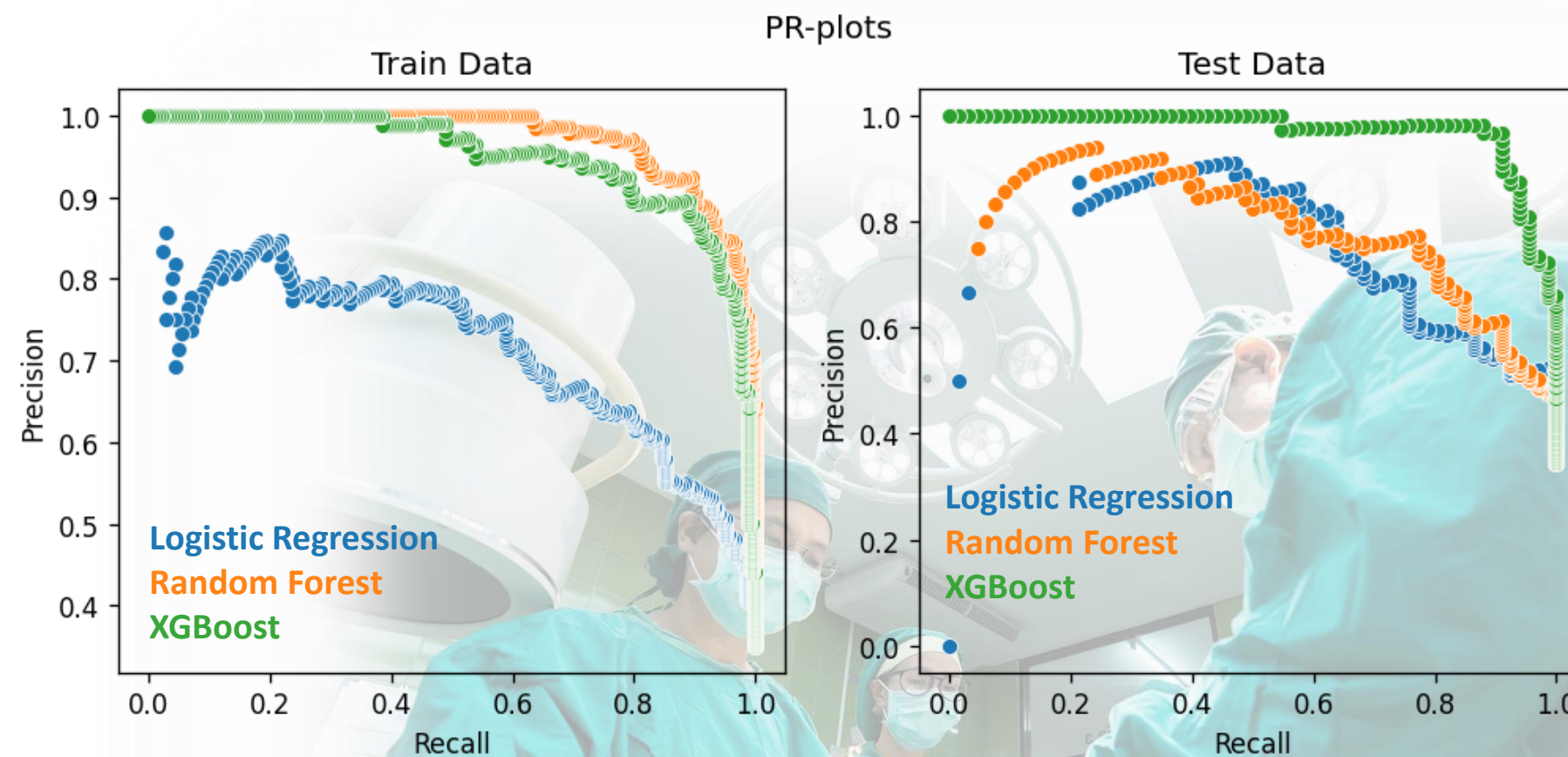
3 ML models were tried

- Logistic regression
- Random Forest
- XGBoosted Random Forests



The tuning

Random and grid search were implemented



Best model

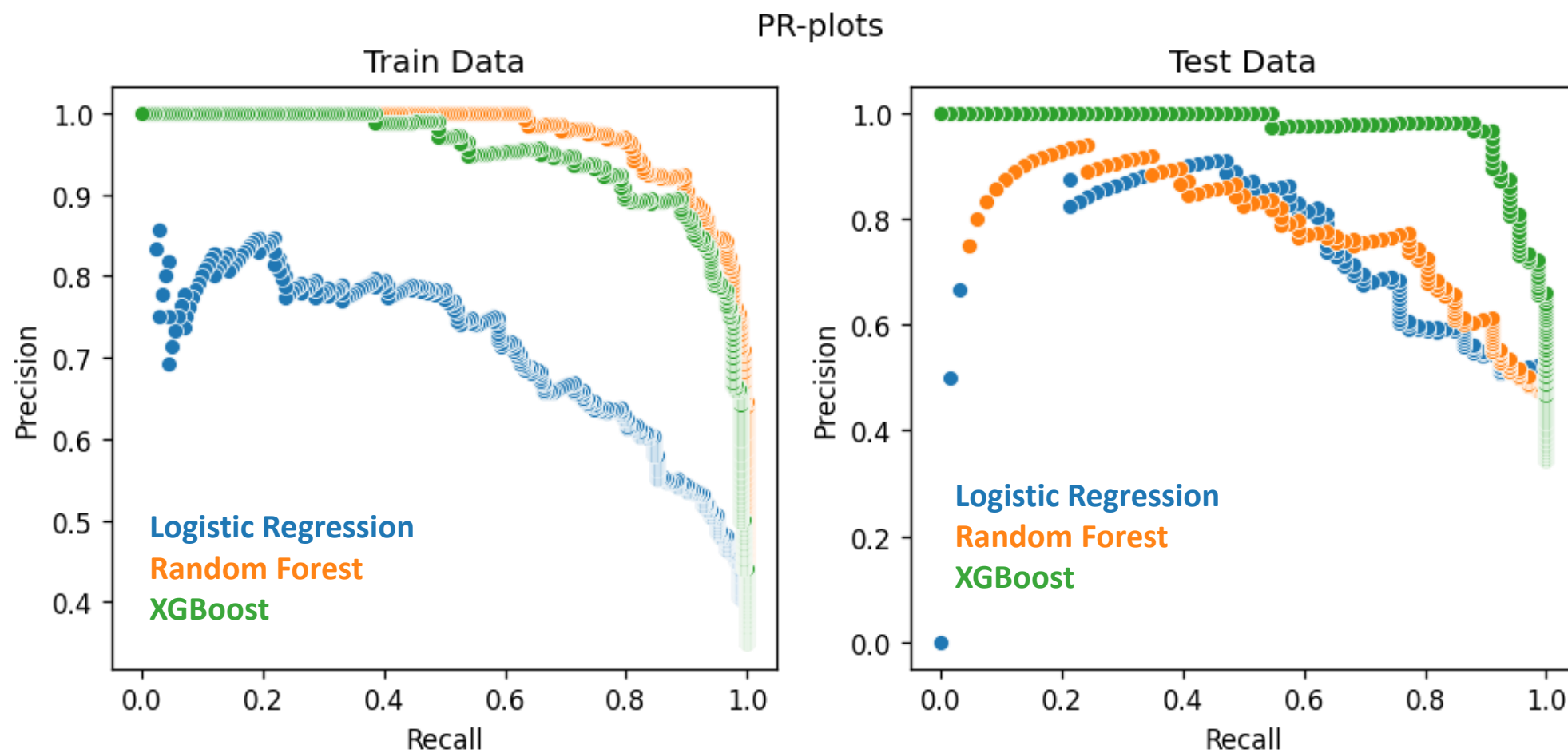
Hyper-tuned XGBoost with:

- Recall = 100%
- Precision = 63%

Machine Learning

Notes on the Models

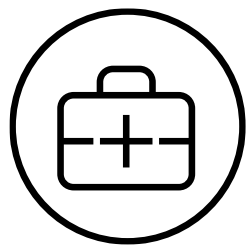
The accuracy of the models was chosen based on the area under the Precision-Recall curve, which shows a score **between 0 and 1** (correspondingly no fit and perfect fit), this is called the “Average accuracy score”



- The Linear regression model seems to function well, but struggles with its selectivity (Average accuracy score $\approx 0.8 \pm 0.15$)
- The Random Forest seems to overfit (Average accuracy score $\approx 0.77 \pm 0.15$)
- XGBoosted Trees seems to perform the best, giving a nice fit and an Average accuracy score of $\approx 0.78 \pm 0.08$

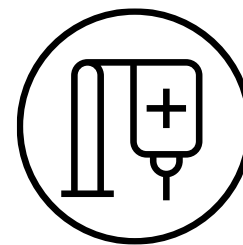
Message to take home

Diabetes type II is a mostly predictable and preventable disease



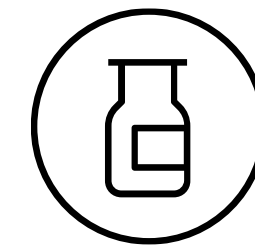
Weight is a big factor

Having Body Mass Index over 25 correlated with high probability of having diabetes



Blood sugar levels

Balanced diet is could be a key for avoiding/treating diabetes



The Model

A model is available that can with high accuracy identify diabetics from their clinical parameters



Thank You

This project was delivered to you by Boris Y.
Nedyalkov

