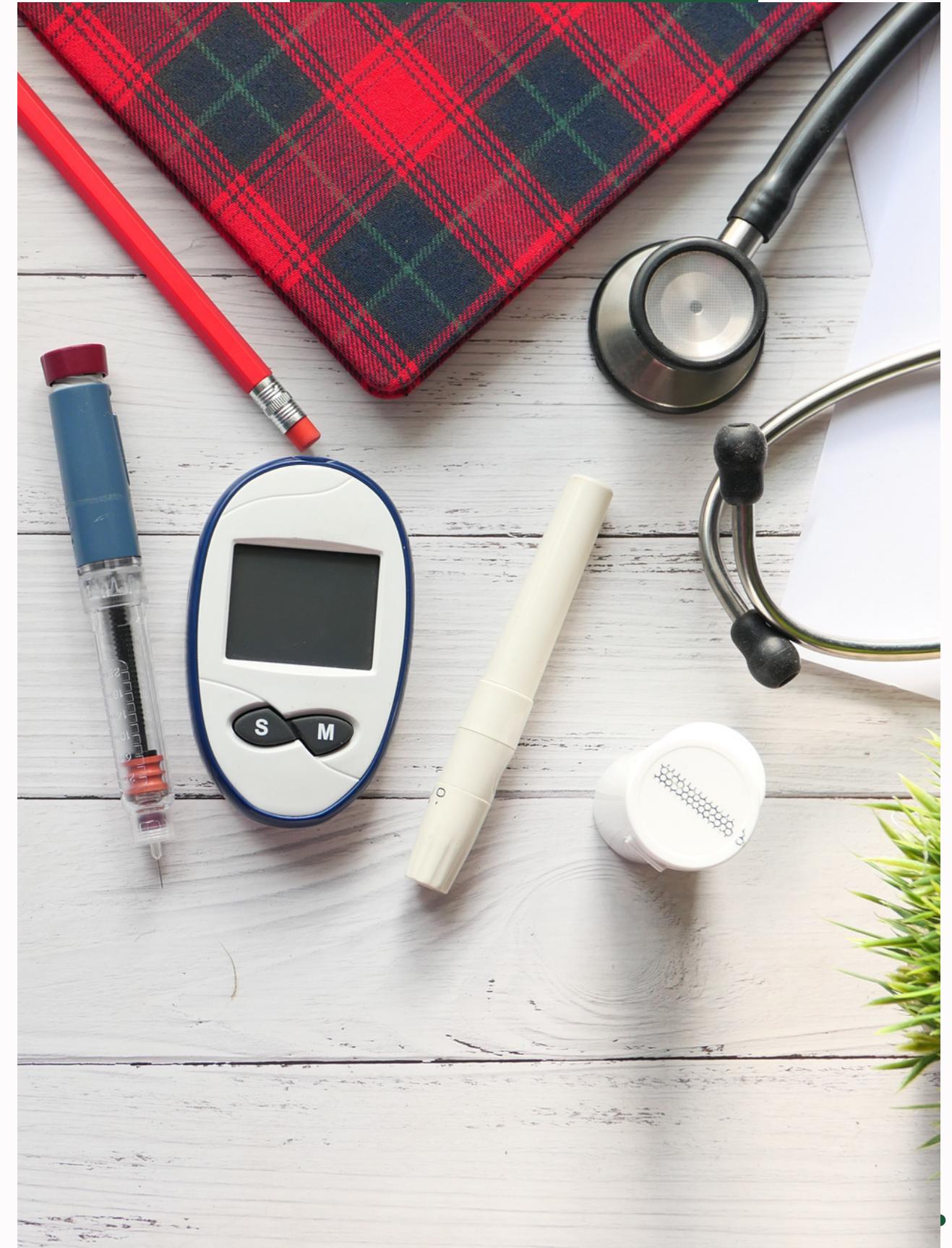


Presentation 2023

PREDIKSI DIABETES

Project Internship MeriSKILL

Asmiyeni Islamiati



Content

01

BACKGROUND

02

EDA & INSIGHT

03

DATA PRE-PROCESSING

04

MODELLING

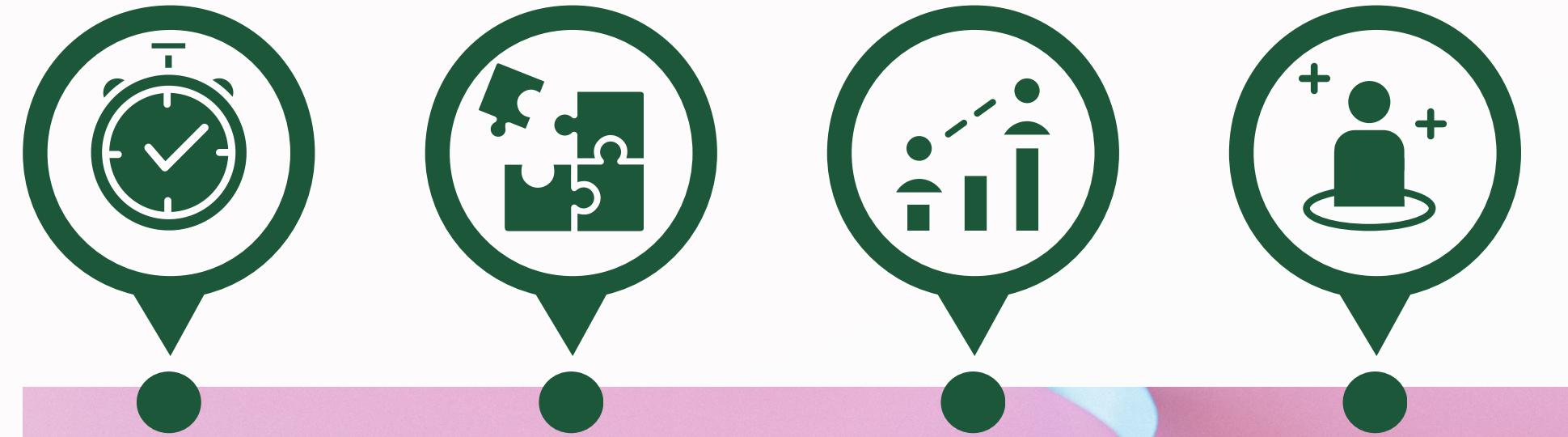


A wide-angle photograph of a modern office common area. The space is filled with lush green plants of various types and sizes, some hanging from the ceiling and others in large planters on the floor. There are several wooden desks arranged in a U-shape, each equipped with a laptop and office chairs. In the background, there are large windows looking out onto a cityscape, and a wooden sofa is visible near one of the windows. The ceiling is made of light-colored wood and features large, circular ductwork. The overall atmosphere is bright and airy.

BACKGROUND

Problem Statement

Masalah yang dihadapi adalah mendeteksi secara diagnostik apakah seorang pasien memiliki diabetes berdasarkan pengukuran diagnostik tertentu yang terdapat dalam dataset. Dataset ini berasal dari National Institute of Diabetes and Digestive and Kidney Diseases dan berfokus pada pasien-pasien perempuan setidaknya berusia 21 tahun dan memiliki keturunan Pima Indian.



Objectives

Mengidentifikasi variabel-variabel medis yang independen dan menjelaskan hubungannya dengan hasil diabetes.

01

Mengembangkan model prediktif menggunakan machine learning untuk memprediksi diabetes berdasarkan variabel-variabel medis yang diberikan.

02

Mengevaluasi kinerja model prediktif untuk memastikan akurasi dan keandalan prediksi.

03

Business Metrics

Diabetes Detection Rate (DDR):

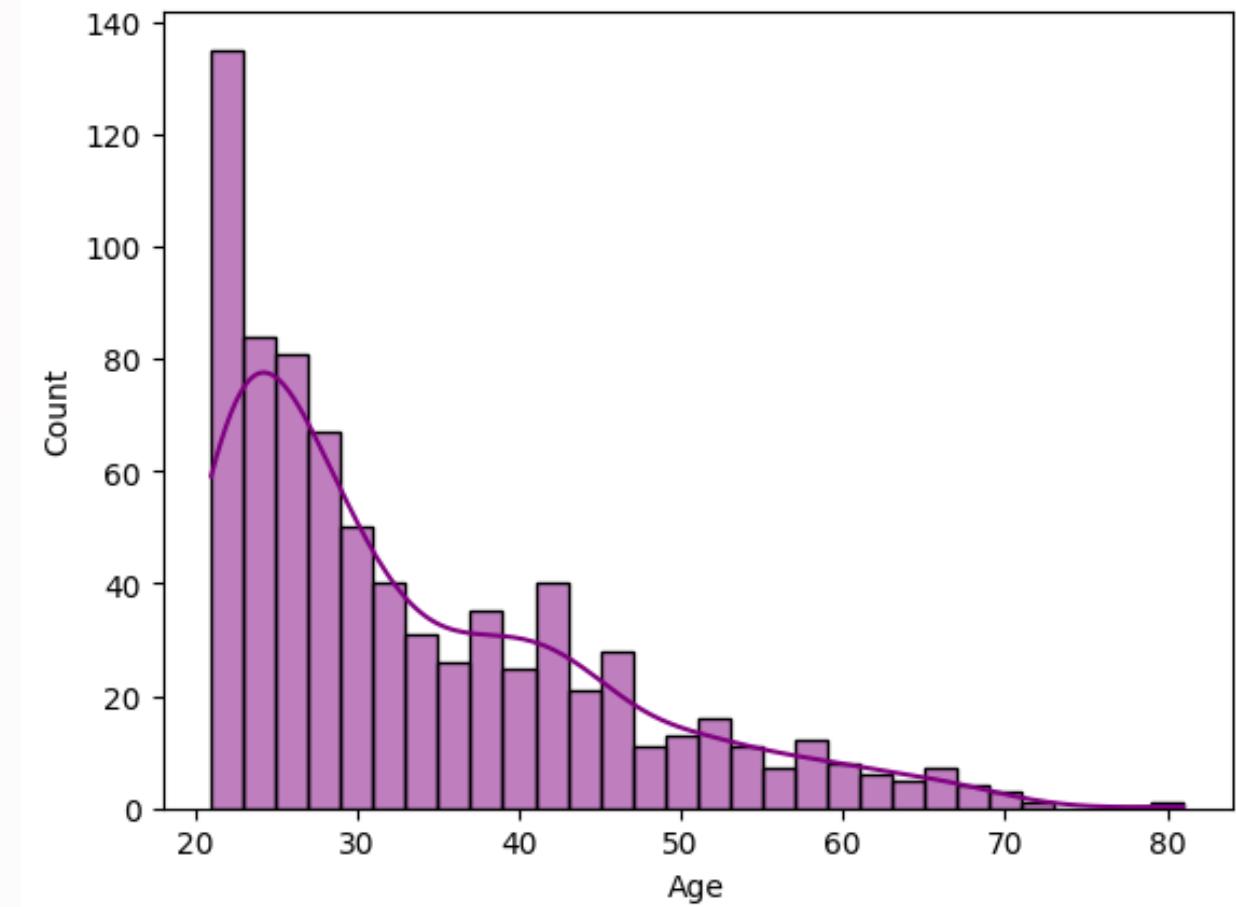
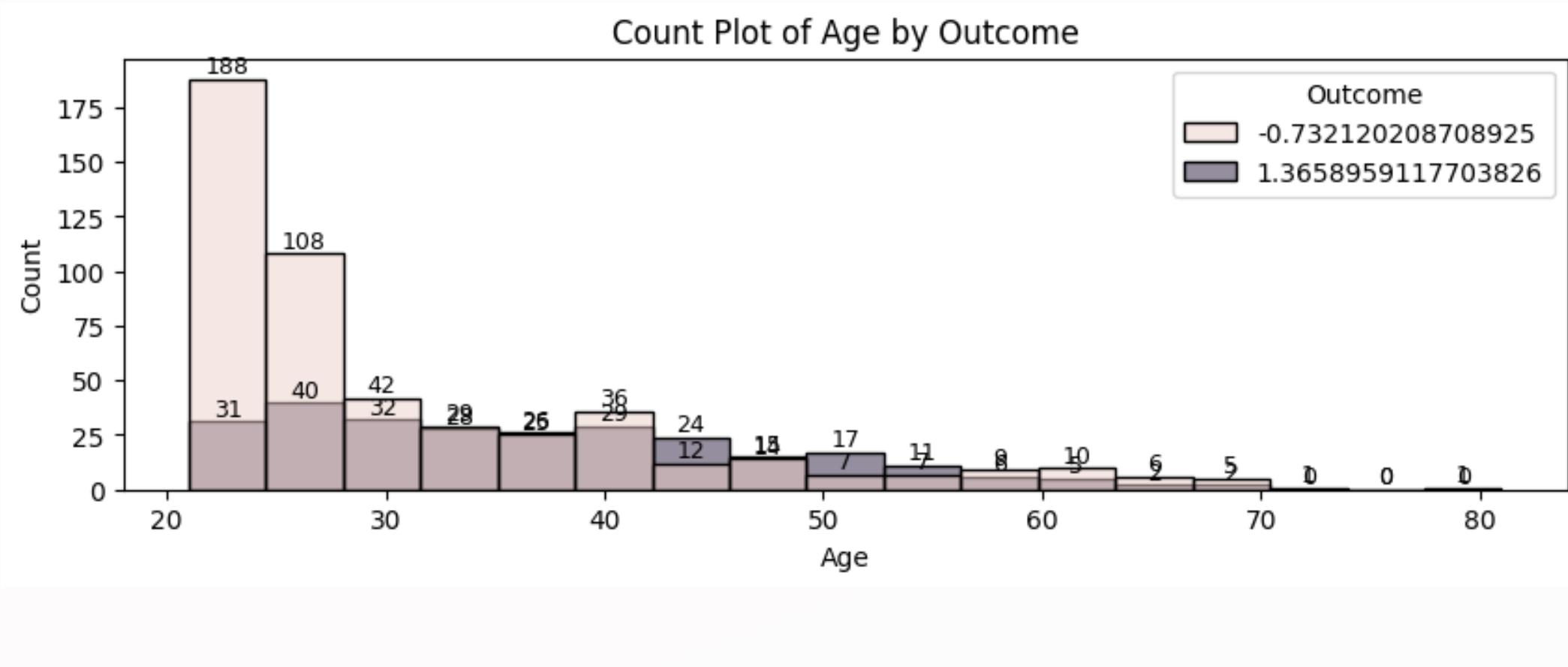
Mengukur persentase pasien yang benar-benar memiliki diabetes dibandingkan dengan keseluruhan pasien yang diidentifikasi sebagai berpotensi memiliki diabetes oleh model.

Goal

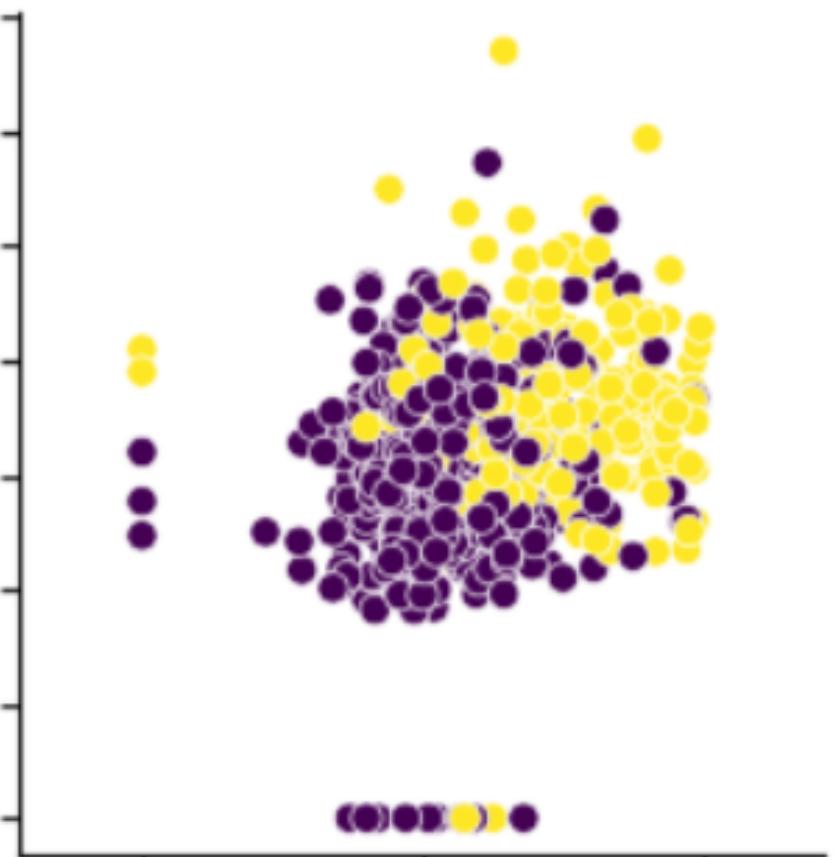
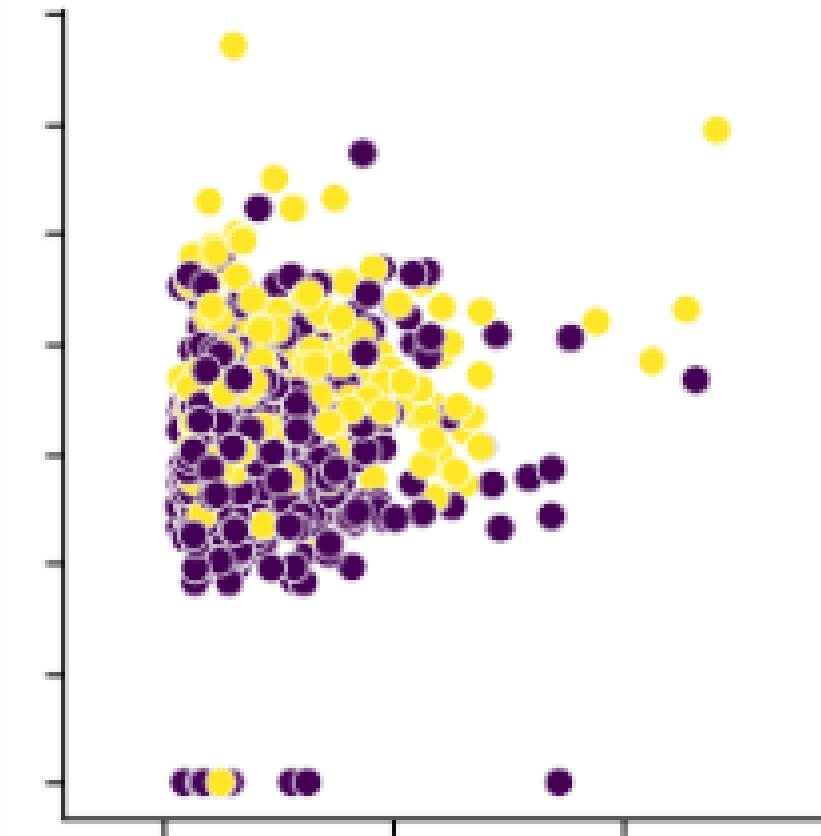
Memprediksi apakah pasien memiliki diabetes atau tidak berdasarkan hubungan antara variabel-variabel medis yang independen dengan variabel target.



EXPLORATORY DATA ANALYSIS & INSIGHT



- Berdasarkan analisis umur, pasien yang berusia di bawah 30 tahun cenderung memiliki risiko outcome yang lebih rendah.
- Perlu diperhatikan pula bahwa pasien dengan outcome terbanyak berada di rentang umur <30 tahun



- Berdasarkan visualisasi pada pairplot, fitur BMI dan Glukosa mengindikasikan korelasi yang baik dengan fitur-fitur lain.
- Diantara seluruh korelasi BMI, yang paling mengindikasikan kecocokan adalah BMI & Diabetes Predigree Function
- Diantara seluruh korelasi Glukose, yang paling mengindikasikan kecocokan adalah Glukose & BMI
- Indikasi korelasi yang baik apabila pada scatter plot terdapat kecenderungan berpisah antara plot fitur dan label



DATA PRE-PROCESSING

Pre Processing Steps

Handling Missing & Duplicated Data

Tidak ditemukan data hilang serta yang terduplicat dalam dataset diabetes

SCALING DATA

Semua data berbentuk integer, sehingga dilakukan standarisasi untuk membuat data berskala sama

Handling Outliers with IQR

- Jumlah baris sebelum memfilter outlier: 768
- Jumlah baris setelah memfilter outlier: 768

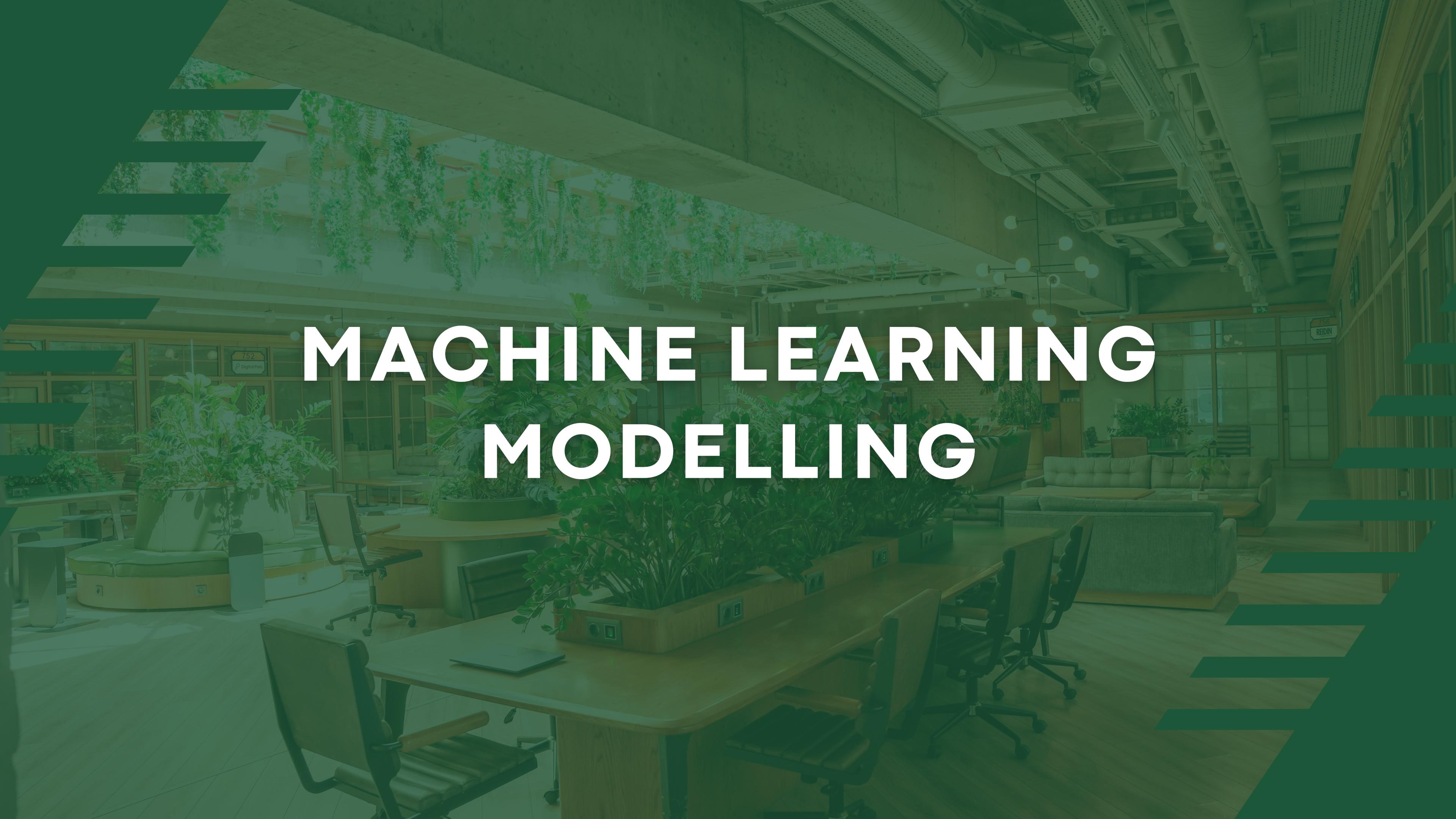
Tidak ada indikasi outliers dalam dataset diabetes

SPLIT TRAIN TEST DATA

- Number of Train Data: 614
- Number of Test Data: 154

HANDLE CLASS IMBALANCE WITH OVERSAMPLING

Diketahui bahwa data train mencakup 80% dari seluruh dataset, oleh karena itu, dilakukan oversampling untuk menangani ketidakseimbangan kelas.



MACHINE LEARNING MODELLING

MODEL

Logistic Regression

ACCURACY	0.75
PRECISION	0.64
RECALL	0.67
F1 SCORE	0.65
DDR	67.27%

	Classification Report:			
	precision	recall	f1-score	support
0	0.81	0.79	0.80	99
1	0.64	0.67	0.65	55
accuracy			0.75	154
macro avg	0.73	0.73	0.73	154
weighted avg	0.75	0.75	0.75	154

- Model memiliki tingkat keakuratan yang tinggi dalam memprediksi pasien yang sebenarnya tidak memiliki diabetes di antara semua prediksinya. Precision mengukur seberapa banyak dari kasus yang diprediksi sebagai negatif (tidak ada diabetes) oleh model yang benar.



THANK YOU

