

**Diabetes
Disease
Prediction**

PIMA INDIANS DIABETES

**CV 2 - WEEKLY
ASSIGNMENT II**

KEMAJUAN INDUSTRI KESEHATAN DI INDONESIA

Industri kesehatan di Indonesia telah mengalami kemajuan yang pesat dalam beberapa tahun terakhir. Hal ini didorong oleh berbagai faktor, antara lain:

- *Pertumbuhan ekonomi yang berkelanjutan. Pertumbuhan ekonomi yang berkelanjutan telah meningkatkan daya beli masyarakat Indonesia, sehingga permintaan terhadap layanan kesehatan juga meningkat.*
- *Peningkatan kesadaran masyarakat tentang pentingnya kesehatan. Masyarakat Indonesia semakin menyadari pentingnya kesehatan, sehingga mereka lebih rajin memeriksakan kesehatan dan melakukan tindakan preventif.*
- *Teknologi kesehatan yang semakin maju. Kemajuan teknologi kesehatan telah membuka peluang baru untuk meningkatkan kualitas layanan kesehatan.*



Background problem

Seberapa penting sistem cerdas pada industri kesehatan di Indonesia dibandingkan pendekatan diagnostik yang sudah ada ?

- Diabetes adalah penyakit serius yang mempengaruhi kesehatan masyarakat.
- Metode diagnosa tradisional diabetes memerlukan waktu dan biaya yang signifikan.
- ANN adalah model komputasi yang efektif dalam prediksi penyakit.
- ANN dapat memproses data pasien, termasuk faktor risiko, untuk prediksi akurat.
- ANN memiliki potensi untuk meningkatkan pencegahan dan manajemen diabetes.

EDA

All Numerical Features

#	Column	Non-Null Count	Dtype
0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	BloodPressure	768 non-null	int64
3	SkinThickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	DiabetesPedigreeFunction	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

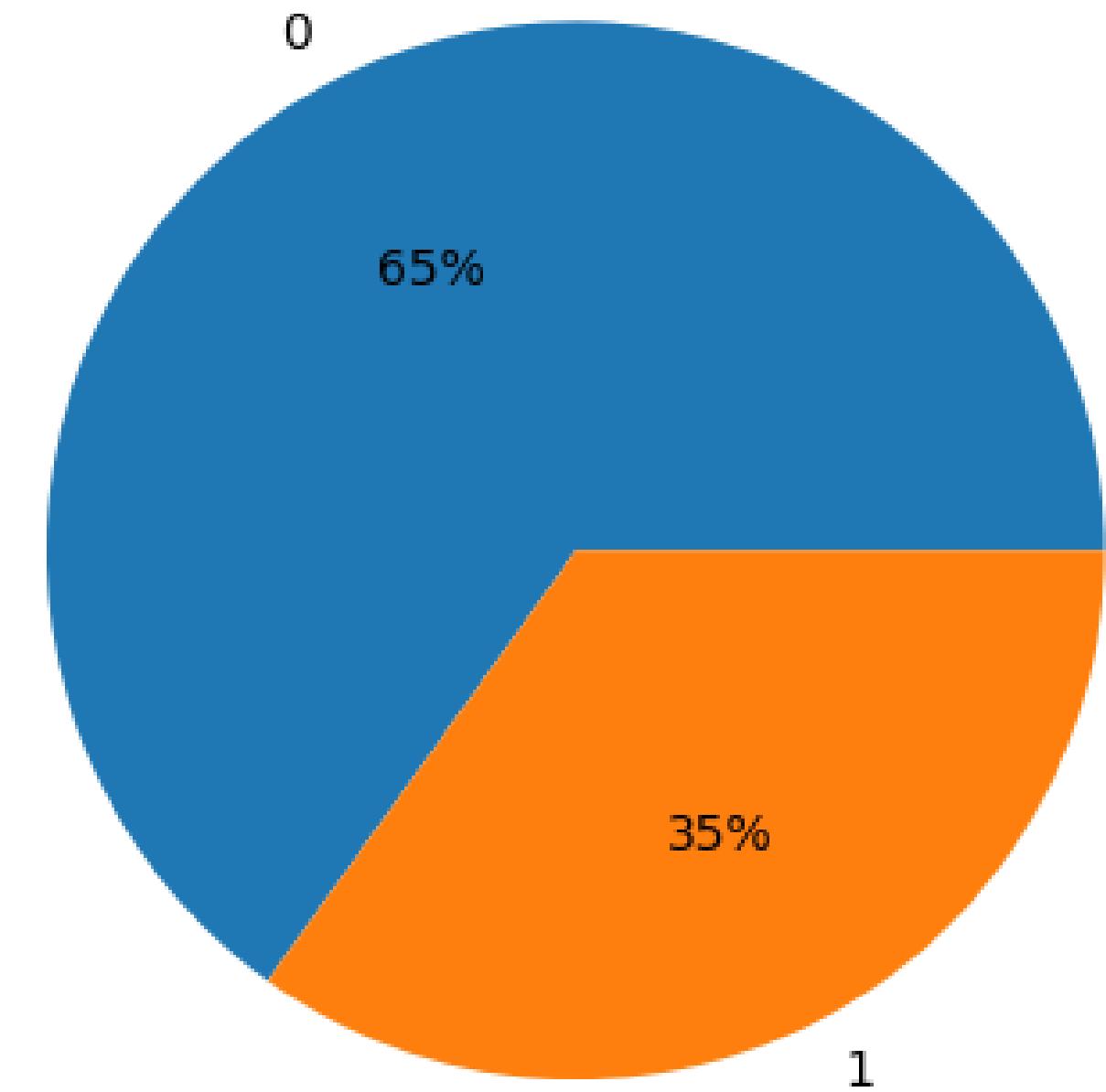
dtypes: float64(2), int64(7)
memory usage: 54.1 KB

- Target (Outcome)
- Binary Classification

Distribusi Target

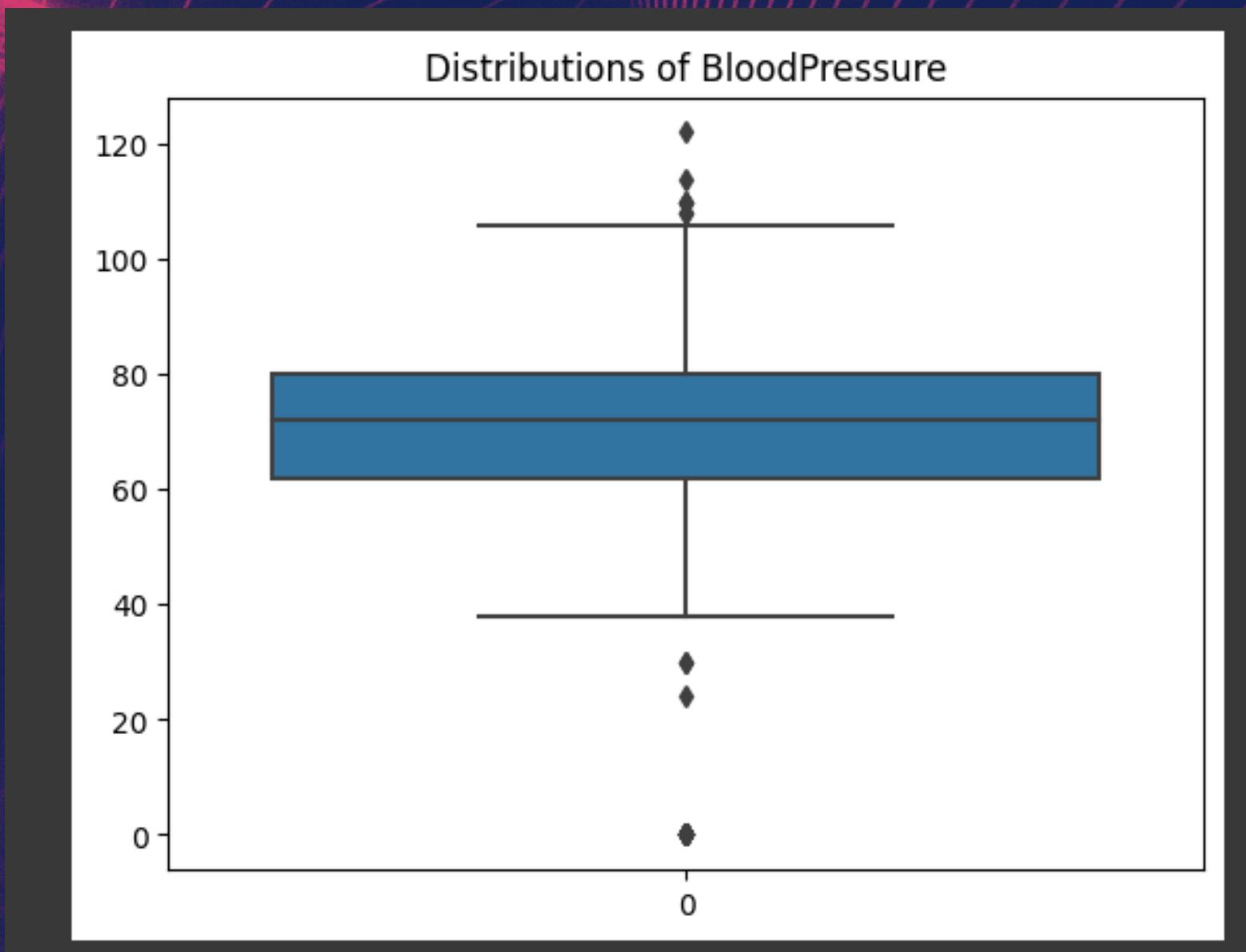
Imbalanced Class

Distributions of Outcome



Distribusi Features

Outlier Exist



Abnormal Data

Jumlah nilai 0 di dalam features

Pregnancies	111	Normal
Glucose	5	
BloodPressure	35	
SkinThickness	227	
Insulin	374	Missing Values
BMI	11	

PREPROCESSING

- Handling Missing Values
- Handling Outlier
- Train Test Split
- Normalisasi

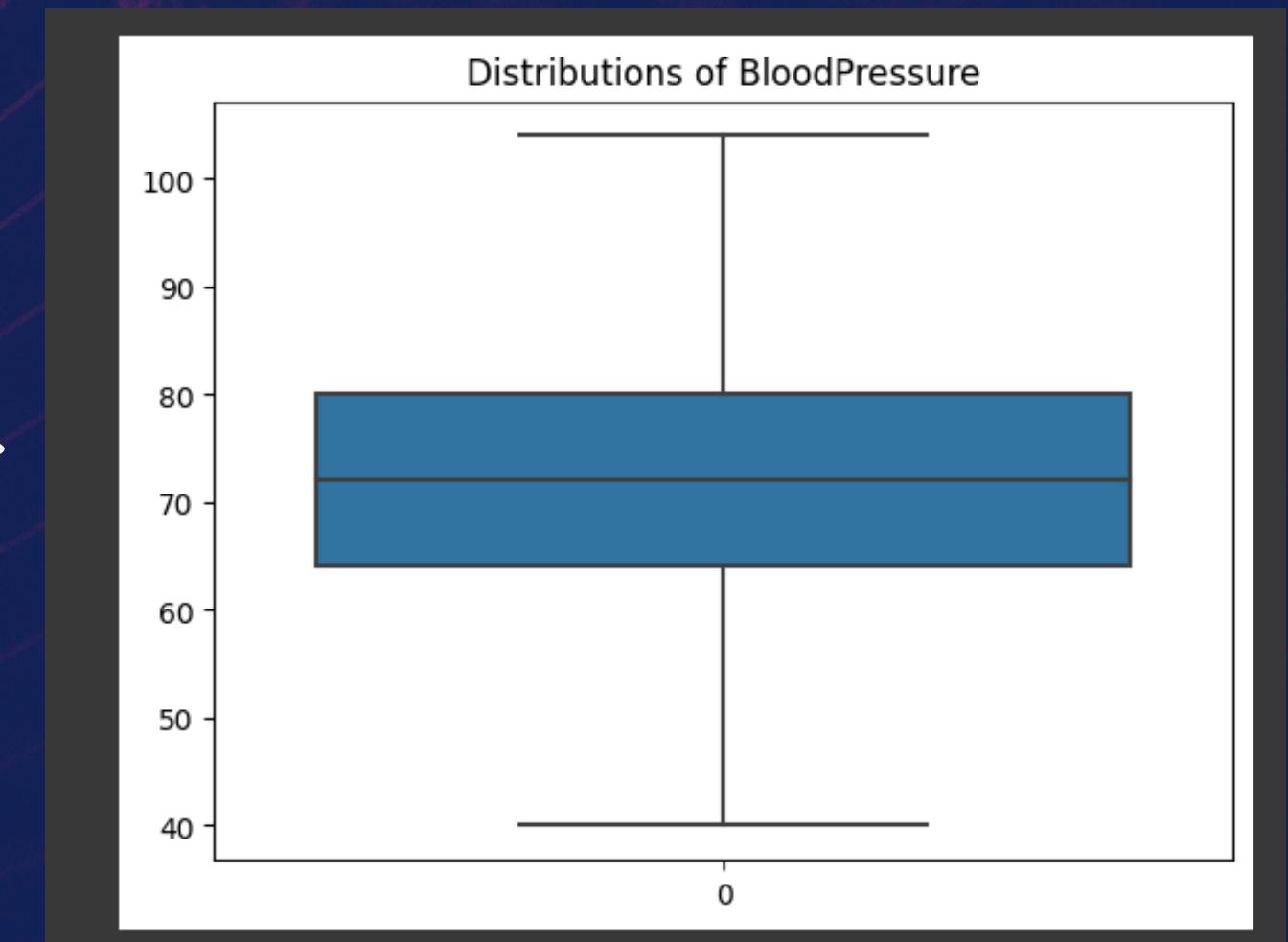
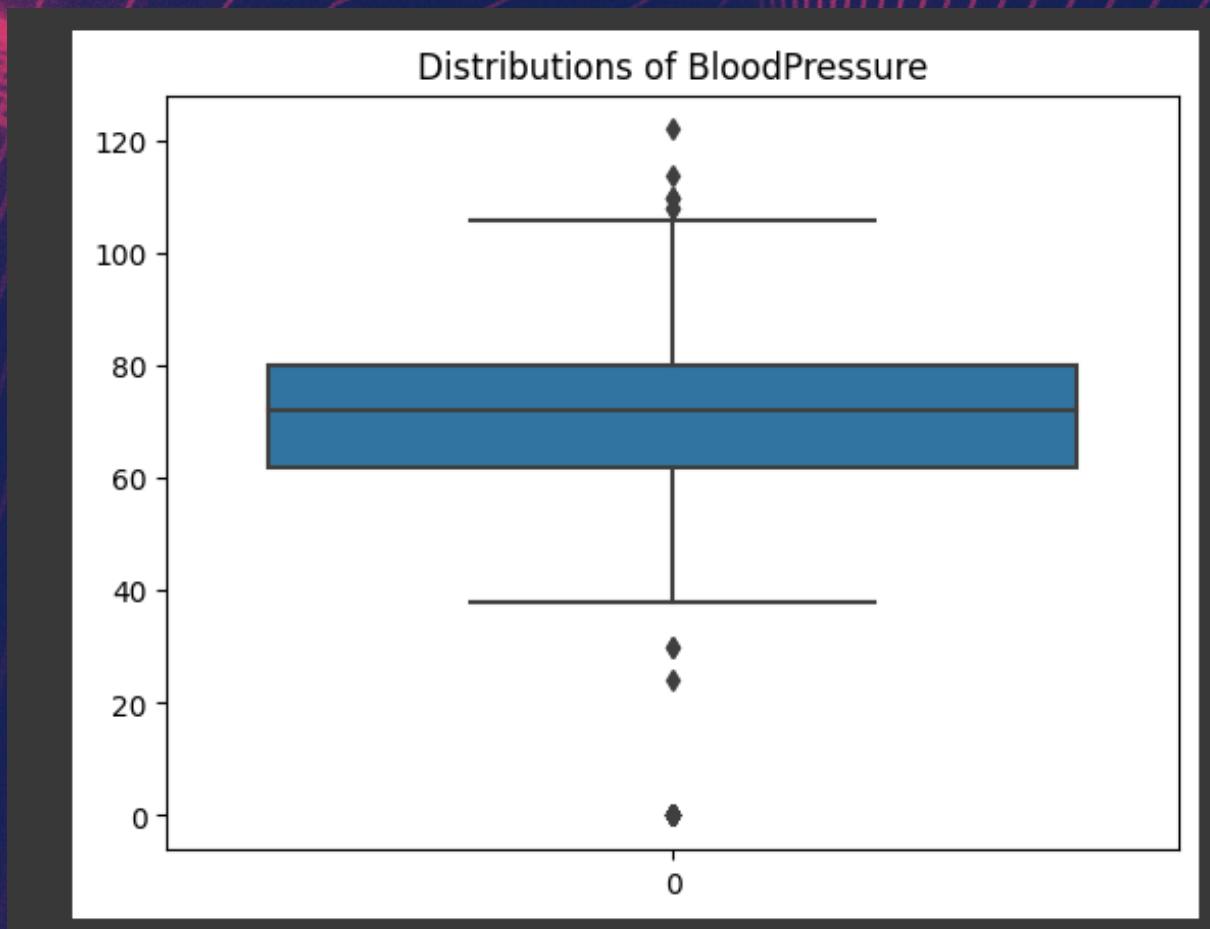
Handling Missing Values

- Drop rows > 70% missing values → 768 -> 733 rows
- Mean Imputer

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Count	0.0	5.00	35.00	227.00	374.0	11.00	0.0	0.0	0.0
Percentage	0.0	0.65	4.56	29.56	48.7	1.43	0.0	0.0	0.0

Handling Outlier

- Capping



Train Test Split & Normalisasi

Test size 20% stratify



Train 586 rows, Test 147 rows

Normalisasi -> StandardScaler

MODELING

- Neural Network Bayesian Optimizer
 - learning rate (0.0001, 0.1)
 - num hidden layers (1, 5)
 - num neurons (5, 50)
 - epochs (5, 50)
- Metric
 - Recall (Utama)
 - Precision
 - F1 Score

iter	target	epochs	learni...	num_hi...	num_ne...
1	0.8235	16.83	0.08164	3.305	18.44
2	0.5882	19.9	0.05582	2.058	40.44
3	0.8235	33.93	0.08268	4.314	8.461
4	0.7451	48.66	0.03527	2.355	41.72
5	0.6471	17.32	0.03288	2.762	26.95
6	0.7451	16.71	0.06267	3.289	18.58
7	0.8824	17.13	0.08929	3.292	18.07
8	0.7451	15.36	0.04863	4.681	44.96
9	0.902	42.81	0.08616	4.191	14.54
10	0.8431	7.691	0.01301	4.209	28.44
11	0.7059	36.64	0.01375	2.439	45.69
12	0.8235	17.38	0.003143	3.204	18.29
13	0.7451	14.62	0.03996	1.861	16.61
14	0.6863	47.34	0.07655	4.573	35.31
15	0.8235	38.56	0.0664	2.353	11.14
16	0.9216	42.95	0.05251	4.679	14.27
17	0.8627	45.56	0.04346	4.053	15.73
18	0.8235	43.18	0.04089	4.365	14.48
19	0.7255	37.44	0.04702	1.163	10.83
20	0.7255	16.58	0.05416	1.657	27.72
21	0.6275	30.36	0.09864	1.816	19.67
22	0.6863	16.13	0.02634	2.16	25.08
23	0.7255	5.532	0.04251	1.553	43.67
24	0.7255	27.91	0.06774	2.723	34.11
25	0.8431	27.3	0.06777	4.611	8.243

Based on optimizer
process range
hyperparameters

- Deep
- Shallow

MODELING

- learning rate = 0.052506
- num hidden layers = 4
- num neurons = 14
- epochs = 42



Recall: 92.16%

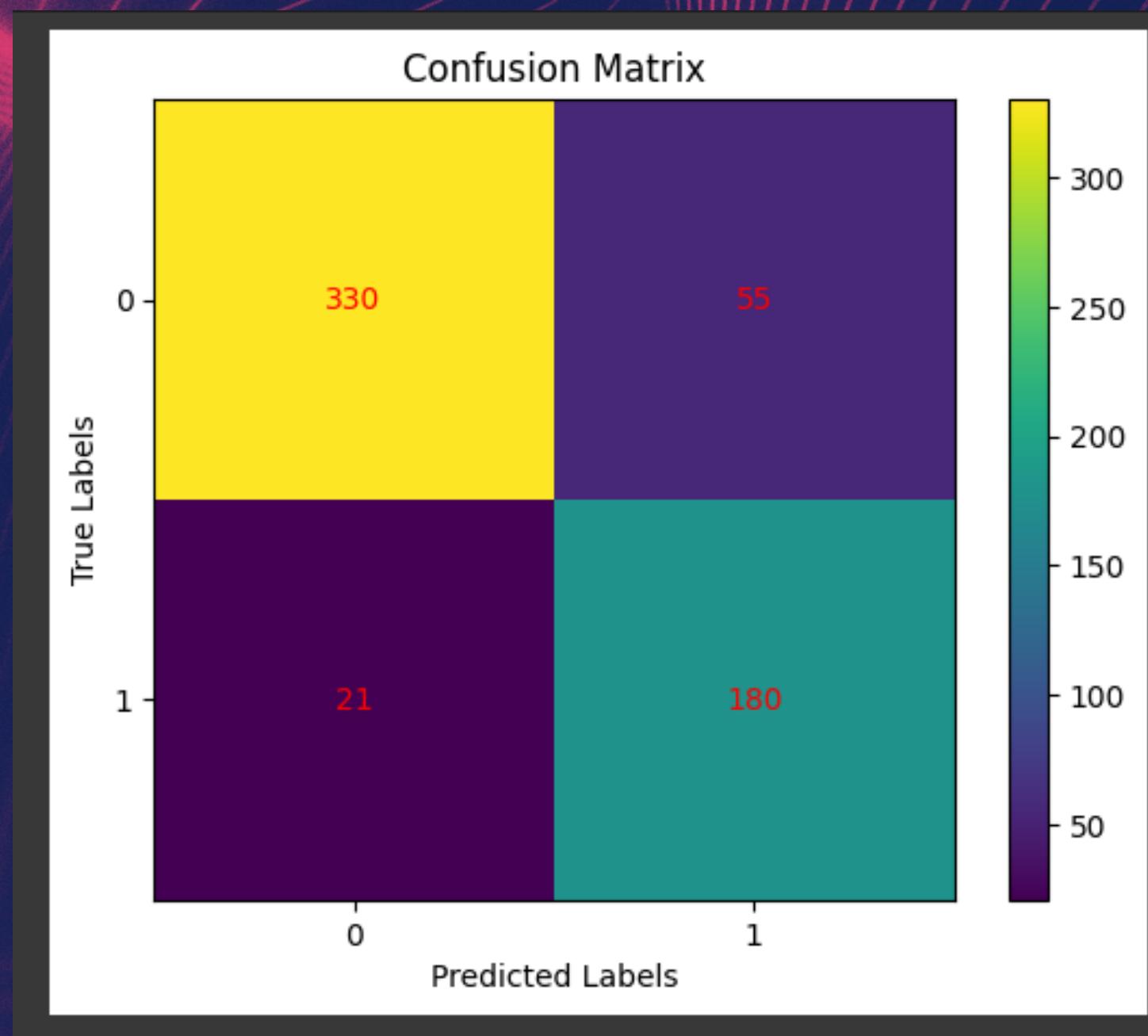
Retrain new model based on optimized hyperparameters

EVALUASI

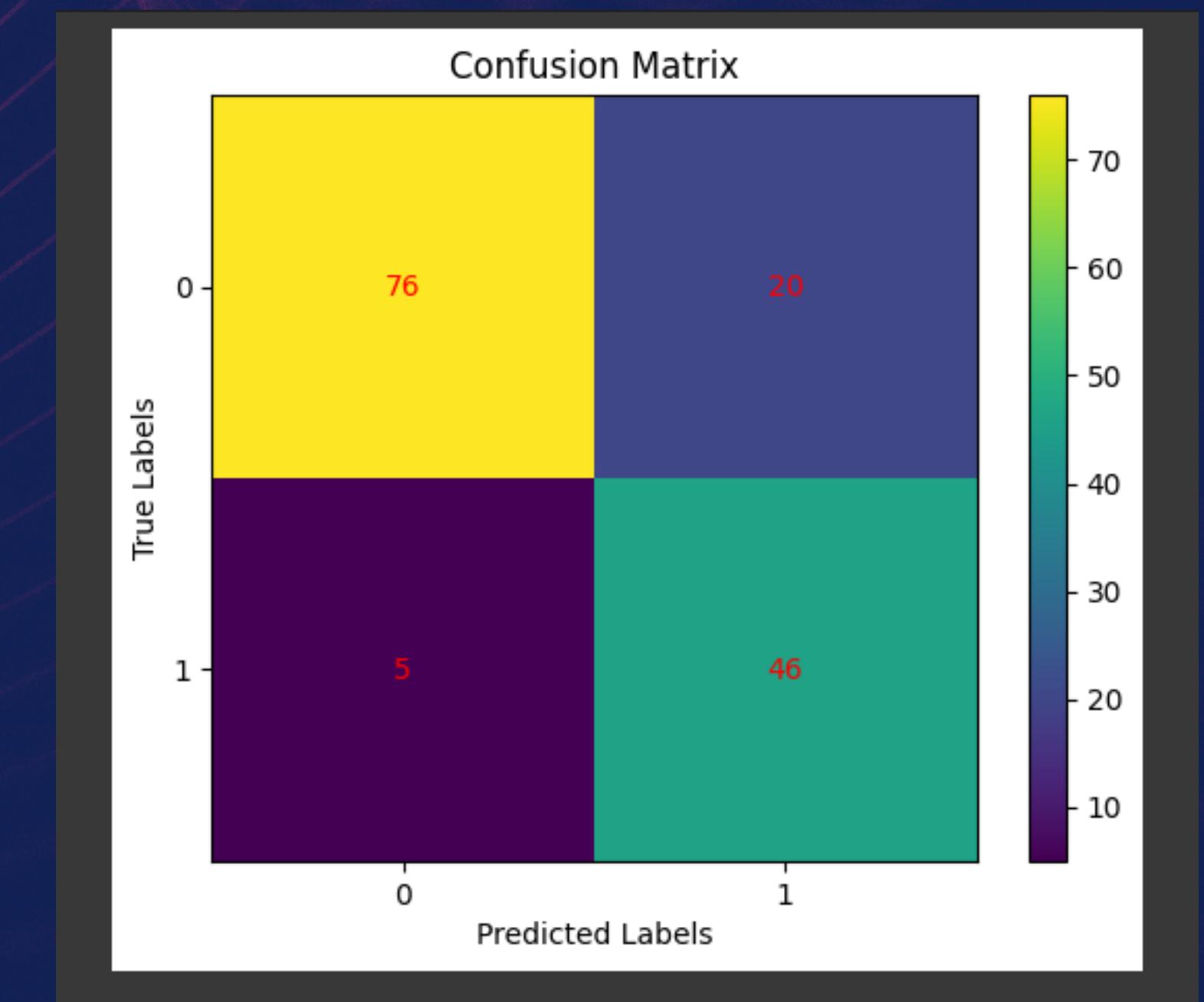
	Recall	Precision	F1 Score
Train	84.08 %	87.11 %	80.32 %
Test	90.20 %	69.70 %	78.66 %

EVALUASI

Train

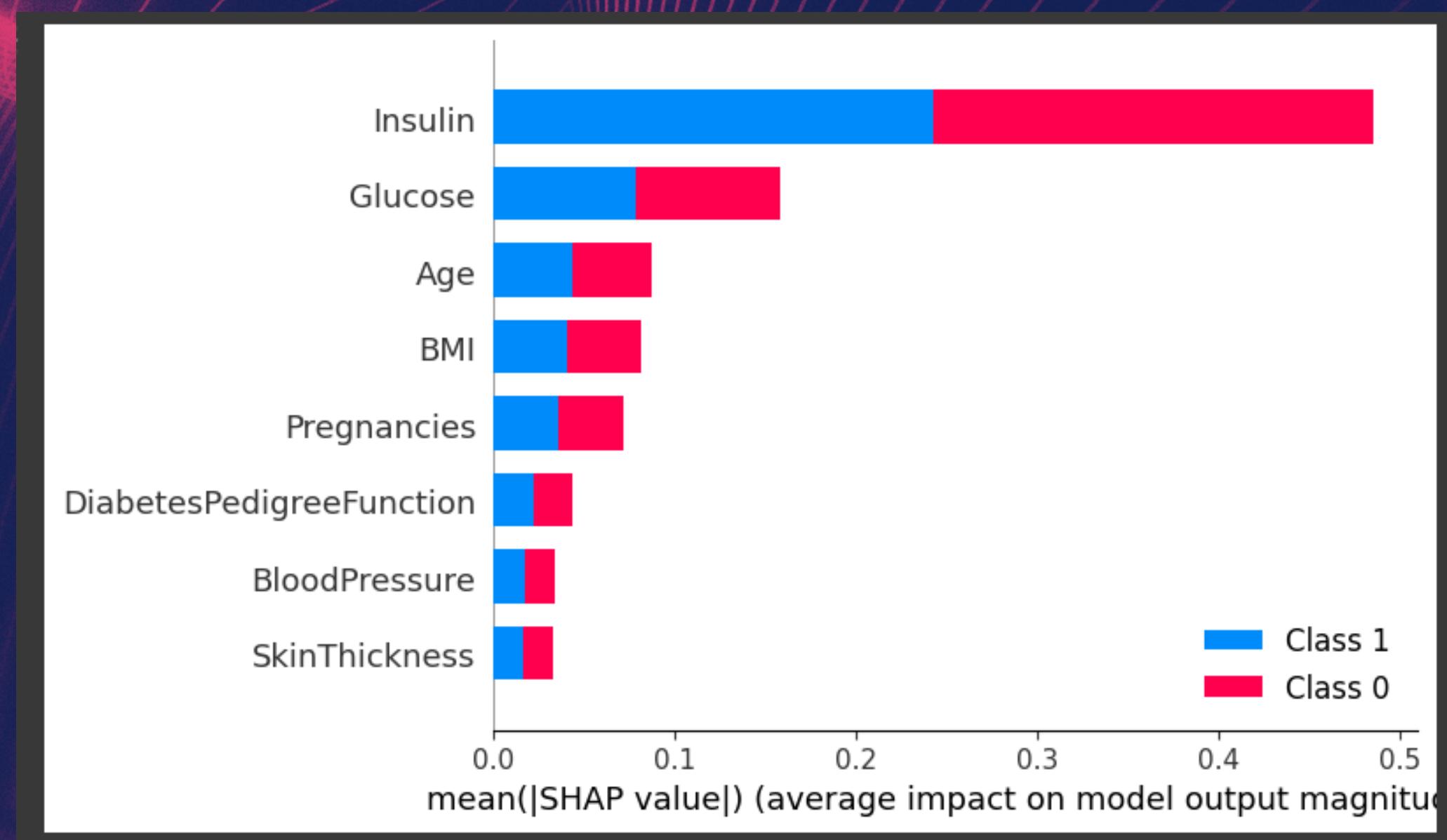


Test



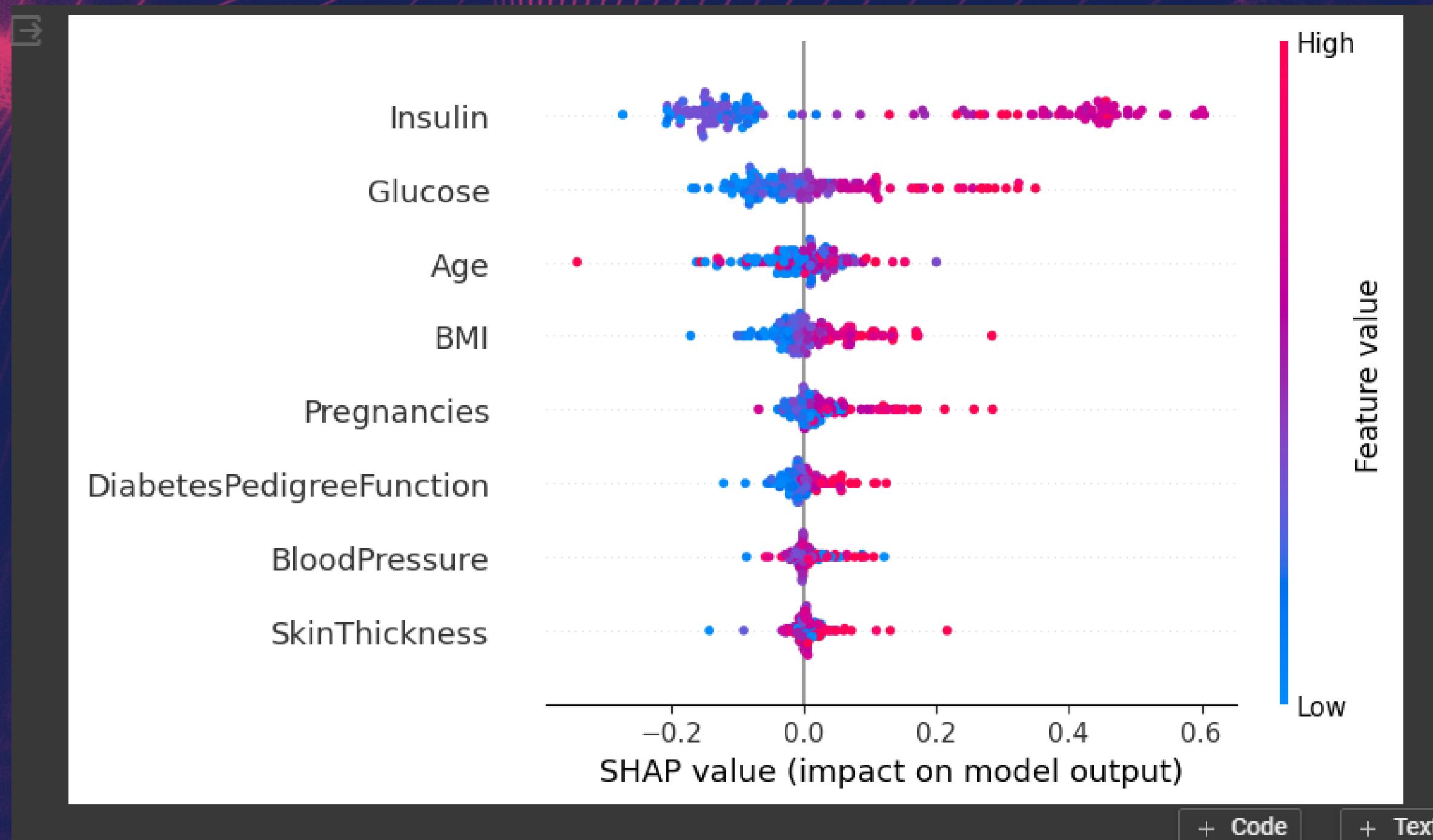
EVALUASI

Feature Importance (Shap)



EVALUASI

Impact on Class 1



KESIMPULAN

1. Insulin dan Glukosa merupakan 2 fitur terpenting
2. Dapat disimpulkan bahwa semakin tinggi nilai Insulin, Glukosa, Usia, BMI dan kehamilan cenderung didiagnosis sebagai diabetes
3. Model ini memiliki recall tinggi (0.9020) dalam mencegah kesalahan diagnosis pasien diabetes sebagai non diabetes, tetapi presisinya (0.6970) lebih rendah dalam mencegah kesalahan diagnosis non diabetes sebagai pasien diabetes. Model cocok digunakan sebagai langkah awal pemeriksaan calon pasien diabetes, namun perlu konfirmasi lebih lanjut untuk diagnosis yang akurat.
4. Untuk meningkatkan kinerja, karena kumpulan data hanya berisi 768 baris, lebih baik mengumpulkan lebih banyak data dan data yang lebih bersih (misalnya Insulin memiliki banyak nilai yang hilang)



THANK YOU

CV2_OPPENHEIMER

