

DIABETES - INTERMEDIATE PREDICTION

Using Logistic Regression



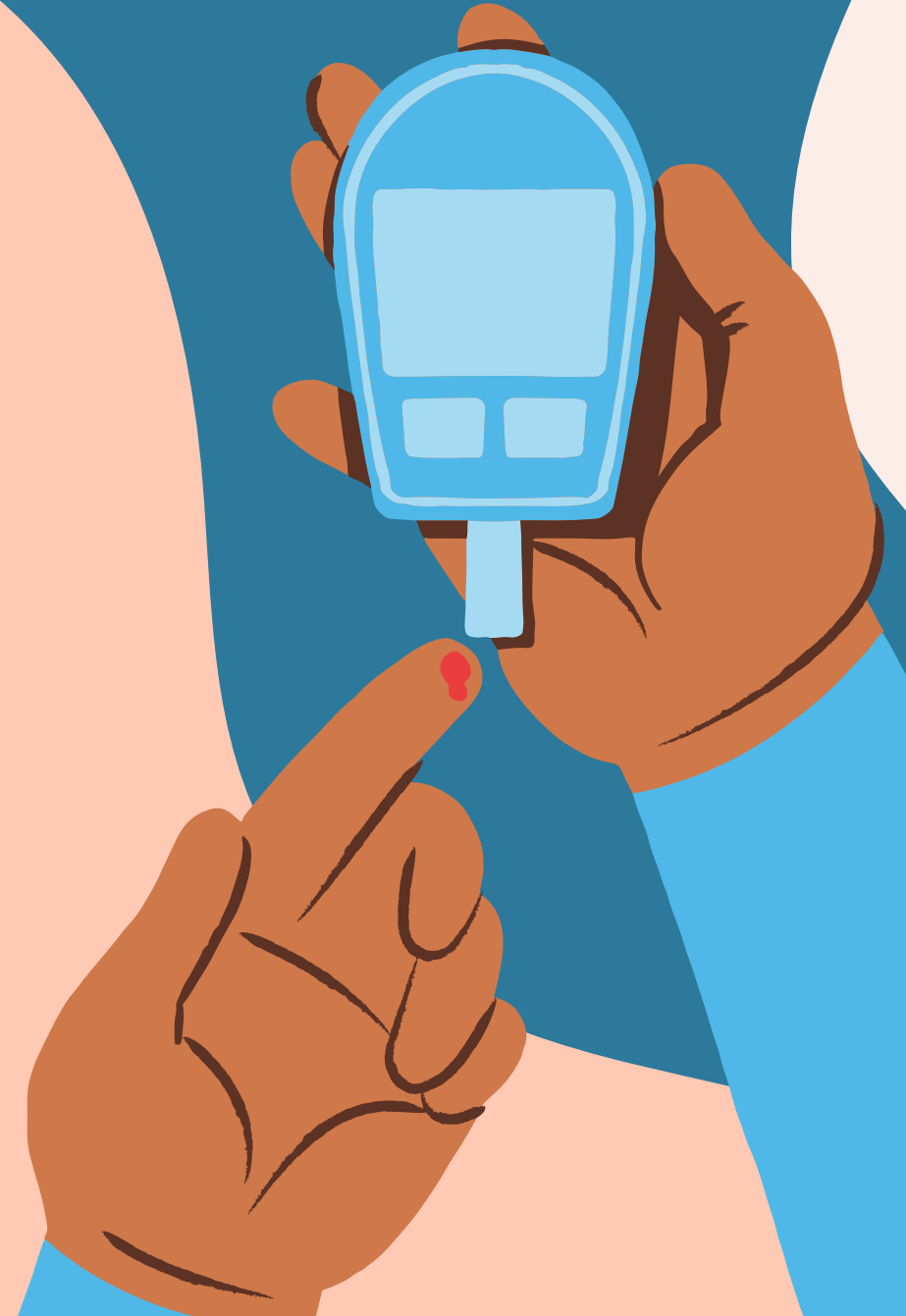
INTRODUCTION

ในปัจจุบันโรคเบาหวานเป็นหนึ่งในปัญหาสุขภาพที่รุนแรงและเป็นที่พึ่งพาของมหาชนทั่วโลก โรคนี้มีการระบาดอย่างรวดเร็วและมีผลกระทบต่อคุณภาพชีวิตของผู้ป่วยทุกคน

โรคเบาหวาน (Diabetes) คือโรคที่เกิดจากความผิดปกติของการทำงานของฮอร์โมนที่ชื่อว่า อินซูลิน (Insulin) ทำให้ร่างกายไม่สามารถนำน้ำตาลที่อยู่ในกระแสเลือดไปใช้ได้อย่างเต็มประสิทธิภาพ ทำให้มีปริมาณน้ำตาลคงเหลือในกระแสเลือดมากกว่าปกติ

DATASET

ชุดข้อมูลนี้จะเกี่ยวกับโรคเบาหวานระดับกลาง ซึ่งชุดข้อมูลนี้จะ
เป็นข้อมูลเกี่ยวกับด้านสุขภาพและผลลัพธ์ของการเป็นโรคเบาหวาน
โดยจะประกอบไปด้วย 8 columns และ 768 rows



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

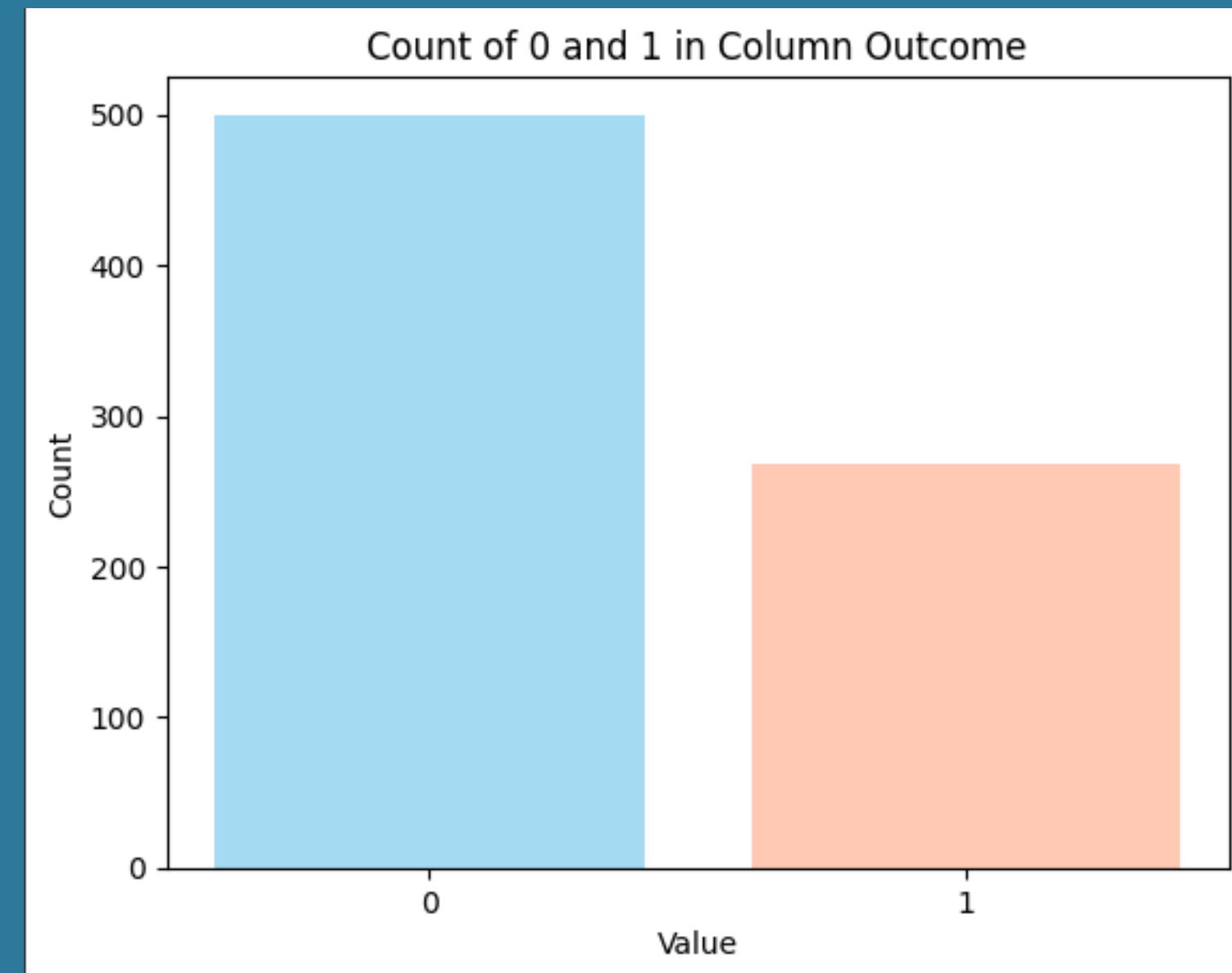
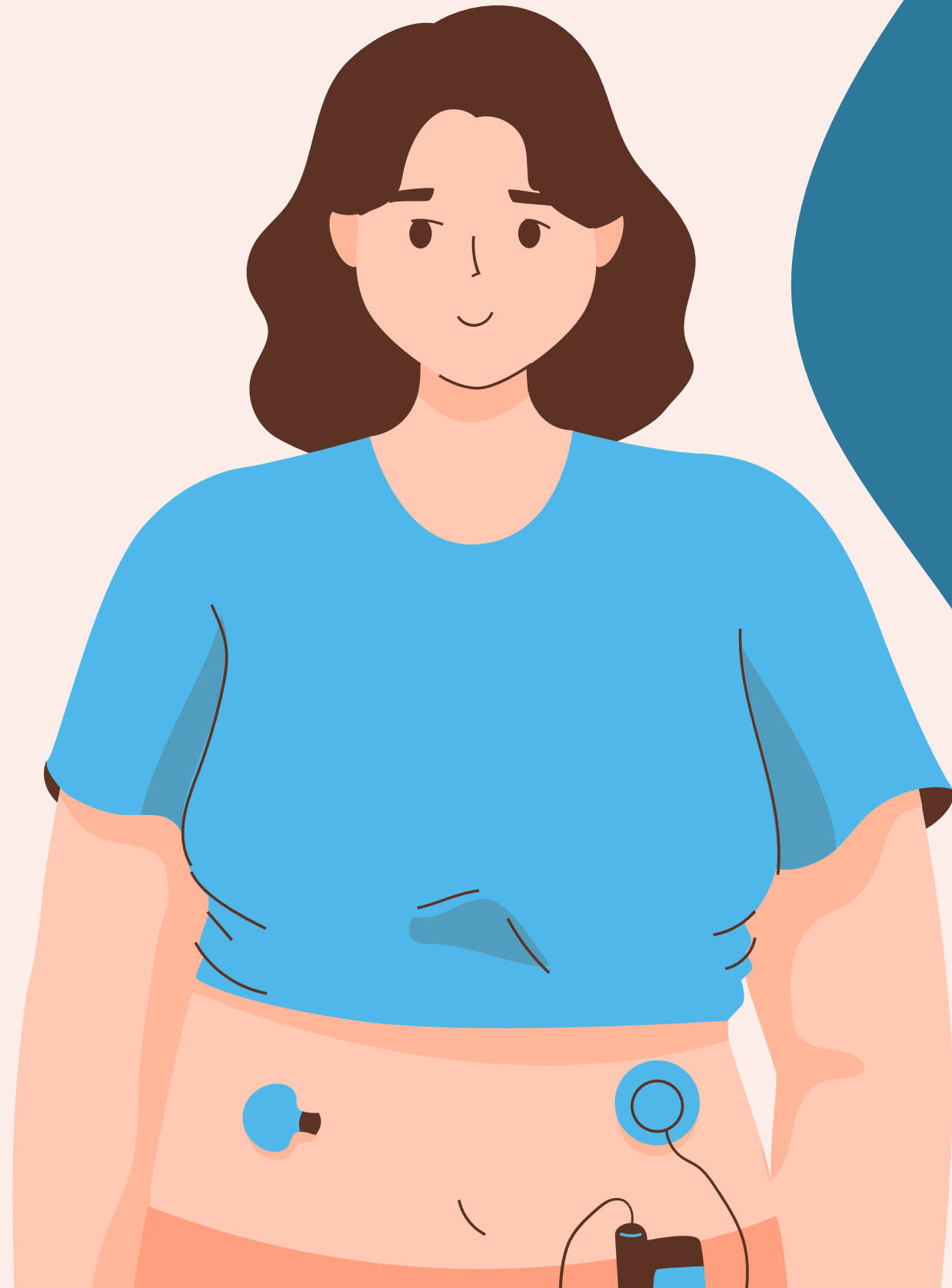
DATASET

Input Values

- Pregnanciessort - จำนวนครั้งที่ผู้ป่วยตั้งครรภ์
- Glucose - ความเข้มข้นของกลูโคสในพลาสมาที่ 2 ชั่วโมงในการทดสอบความทนทานต่อกลูโคสในช่องปาก
- BloodPressure - ความดันโลหิต (mmHg)
- SkinThickness - ความหนาของ Triceps skinfold (มม.)
- Insulin - ระดับอินซูลินในเลือดในชั่วเวลา 2 ชั่วโมง (mu U/ml)
- BMI - ดัชนีมวลกาย (BMI)
- DiabetesPedigreeFunction - ฟังก์ชัน Pedigree ของโรคเบาหวาน
- Age - อายุ (ปี)

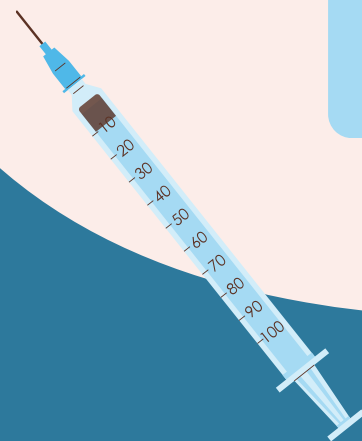


OUTCOME - ข้อมูลคือเป้าหมาย
โดยที่ 0 หมายถึง 'ไม่' (ดีต่อสุขภาพ)
และ 1 แสดงถึง 'ใช่' (เบาหวาน)



LIBRAIRES

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5 from sklearn.metrics import confusion_matrix
6 from sklearn.preprocessing import StandardScaler
7 from sklearn.metrics import precision_recall_curve
```

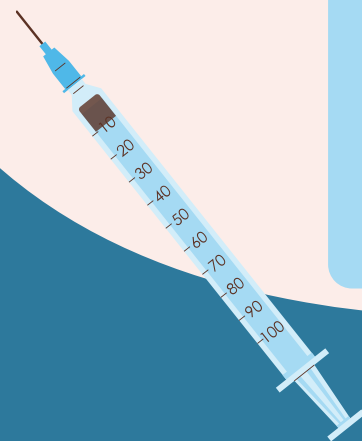
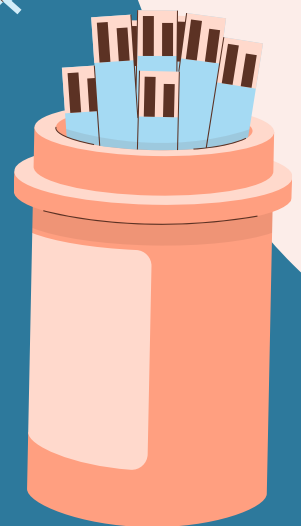


SHUFFLE THE DATAFRAME

```
1 shuffled_df = diabetes_df.sample(frac=1.0, random_state=42)
2 shuffled_df.reset_index(drop=True, inplace=True)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	98	58	33	190	34.0	0.430	43
1	2	112	75	32	0	35.7	0.148	21
2	2	108	64	0	0	30.8	0.158	21
3	8	107	80	0	0	24.6	0.856	34
4	7	136	90	0	0	29.9	0.210	50
...
763	5	139	64	35	140	28.6	0.411	26
764	1	96	122	0	0	22.4	0.207	27
765	10	101	86	37	0	45.6	1.136	38
766	0	141	0	0	0	42.4	0.205	29
767	0	125	96	0	0	22.5	0.262	21

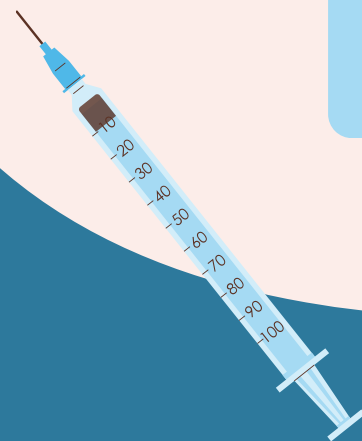
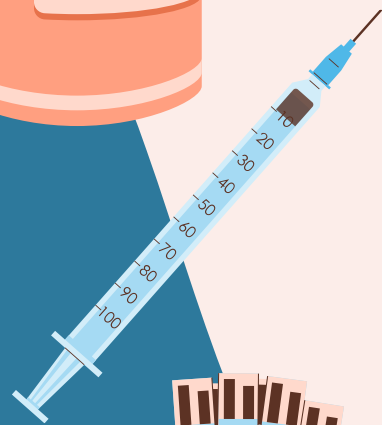
768 rows x 8 columns



TEST TRAIN SPILT DATA

```
1 np.random.seed(42)
2
3 train_ratio = 0.75 # 75% train, 25% test
4
5 train_size = int(train_ratio * len(diabetes_df))
6 test_size = len(diabetes_df) - train_size
7
8 train_data = diabetes_df.iloc[:train_size]
9 test_data = diabetes_df.iloc[train_size:]
10
11 X_train = train_data.drop(columns=['Outcome'])
12 y_train = train_data['Outcome']
13 X_test = test_data.drop(columns=['Outcome'])
14 y_test = test_data['Outcome']
15
```

```
Number of training set: 576
Number of testing set: 192
```



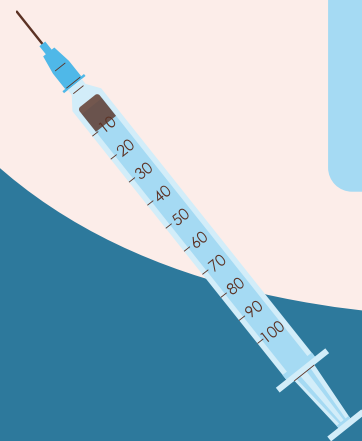
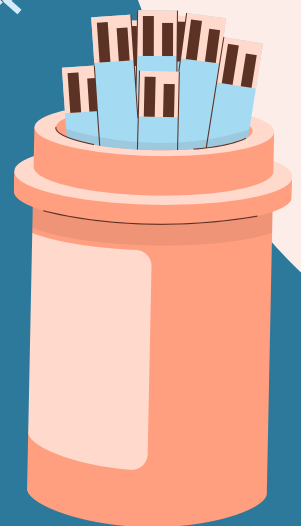
TRAINING MODEL

กำหนด

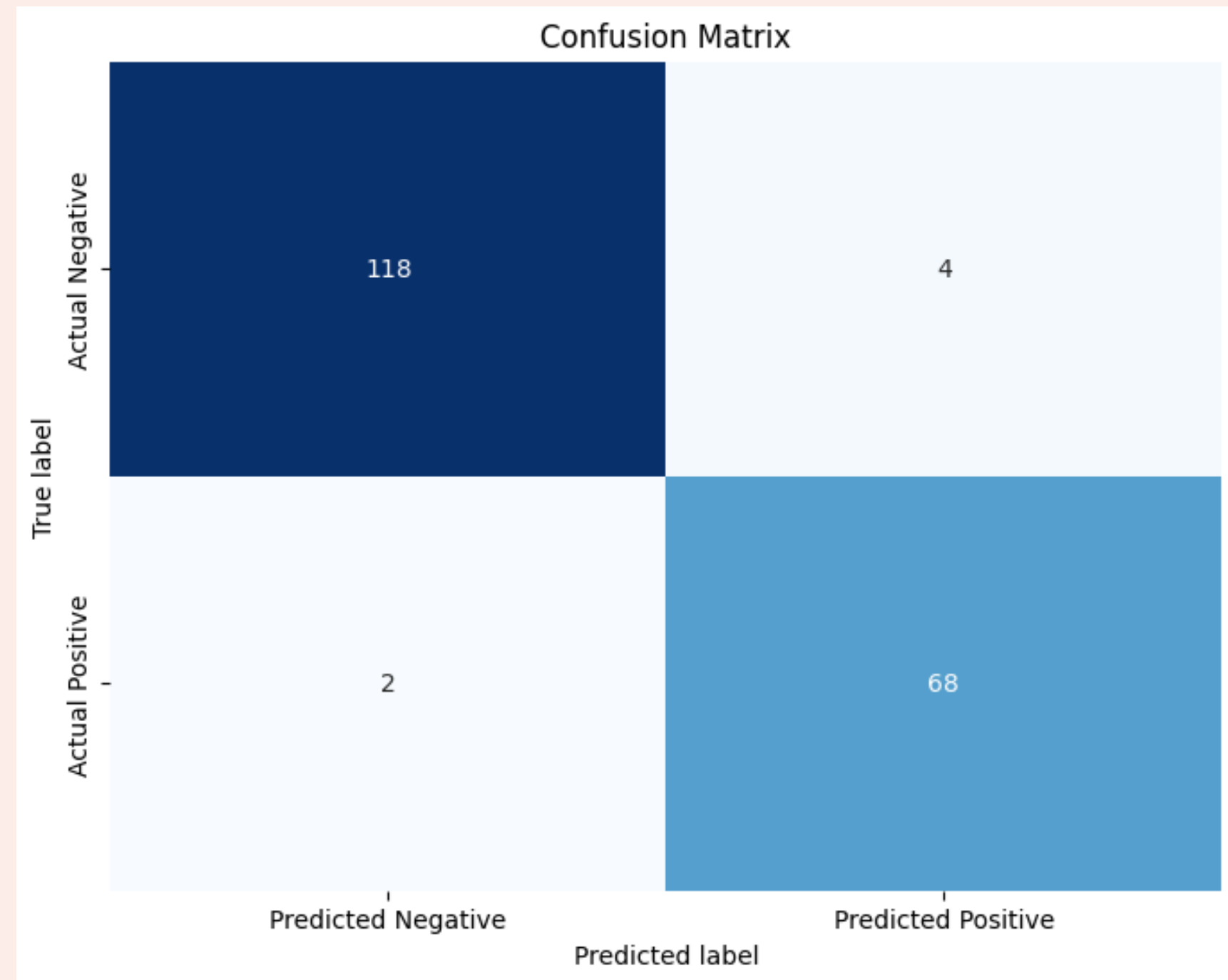
- learning rate = 0.00001
- num_iterations = 1000
- self.bias = 0
- loss function = log loss

Scores

- Precision: 94.44%
- Recall: 97.14%
- Accuracy: 96.87%
- Loss: 68.95%



CONFUSION MATRIX



SUMMARY

จากค่า **accuracy** กับค่า **loss** มีค่าที่สูงเหมือนกัน แสดงให้เห็นว่าโมเดลอาจจะมีความสามารถในการจำแนกข้อมูลที่ถูกต้อง (**accuracy**) แต่ก็มีความไม่แม่นยำในการคาดการณ์ค่าความน่าจะเป็น (**probability**) ของคลาสในบางกรณี ซึ่งค่า **loss** ที่สูงแสดงให้เห็นถึงความไม่แม่นยำในการคาดการณ์นี้

The background features abstract, flowing shapes in a muted blue and a soft peach or light orange color. These shapes are positioned in the corners and along the sides, creating a modern, organic frame for the central text.

THANK YOU!