Machine Learning

Diabetes Classification Model

Presented By: Tina Pham

Table of Contents

01. **Project Goal**

02. **About the Data**

03. **Results**

04. **Future Goals**

01. Project Goal

Create a supervised machine learning model

“To diagnose patient positive or negative to diabetes based on health assessment variables”

02.

About the data

Elaborate on what you want to discuss.

Predictor

"Health Assessment Variables"

Pregnancies

Glucose

Blood pressure

Skin thickness

Insulin

BMI

DiabetesPedigreeFunction

Age

Outcome

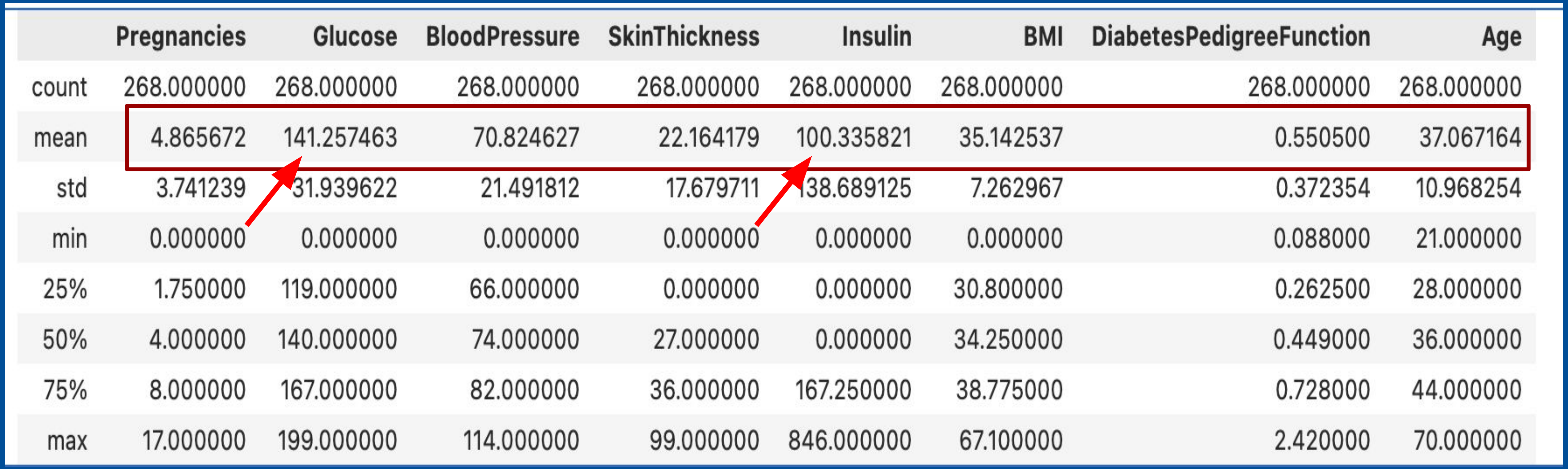
Outcome

negative=0

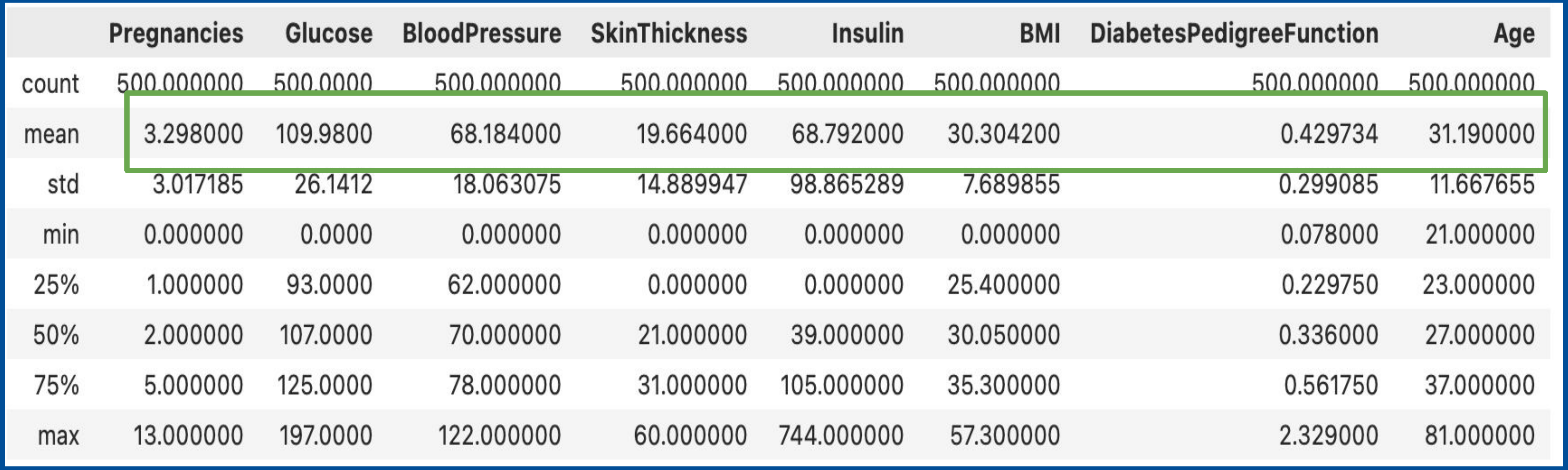
positive=1

Exploratory Data Analysis

Outcome= 1 ⟶ positive



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000
mean	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	0.550500	37.067164
std	3.741239	31.939622	21.491812	17.679711	138.689125	7.262967	0.372354	10.968254
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.088000	21.000000
25%	1.750000	119.000000	66.000000	0.000000	0.000000	30.800000	0.262500	28.000000
50%	4.000000	140.000000	74.000000	27.000000	0.000000	34.250000	0.449000	36.000000
75%	8.000000	167.000000	82.000000	36.000000	167.250000	38.775000	0.728000	44.000000
max	17.000000	199.000000	114.000000	99.000000	846.000000	67.100000	2.420000	70.000000



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	500.000000	500.0000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000
mean	3.298000	109.9800	68.184000	19.664000	68.792000	30.304200	0.429734	31.190000
std	3.017185	26.1412	18.063075	14.889947	98.865289	7.689855	0.299085	11.667655
min	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	93.0000	62.000000	0.000000	0.000000	25.400000	0.229750	23.000000
50%	2.000000	107.0000	70.000000	21.000000	39.000000	30.050000	0.336000	27.000000
75%	5.000000	125.0000	78.000000	31.000000	105.000000	35.300000	0.561750	37.000000
max	13.000000	197.0000	122.000000	60.000000	744.000000	57.300000	2.329000	81.000000

Outcome= 0 ⟶ negative

1. **No Null** data in dataset

2. **Invalid** data with **Blood Pressure, Insulin, BMI, Skin Thickness, Glucose =0**

3. Mean for all **predictors** are **higher** in **positive outcome**

4. Significant higher in measurement for **Insulin and Glucose** for positive outcome

Data Cleaning

Pregnancies **As is**	**Glucose** drop [Glucose]=0	**Blood Pressure** drop [BloodPressure]= 0	**IBM** drop [IBM]=0
Skin Thickness Assume change in skin thickness. **As is**	**Insulin** Assume change in Insulin. **As is**	**PedigreeFunction** **As is**	**Age** **As is**

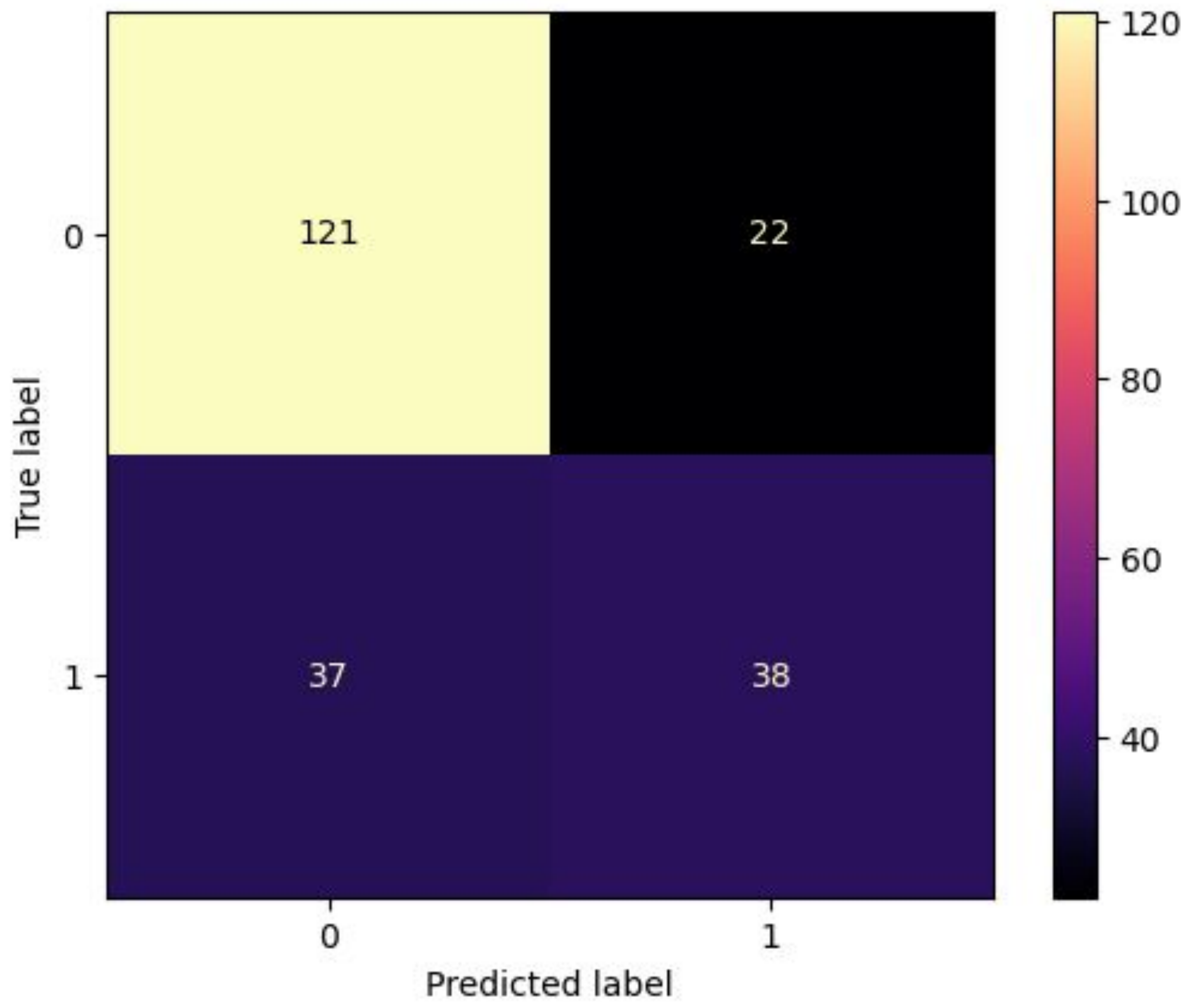
03.

Results

Elaborate on what you want to discuss.

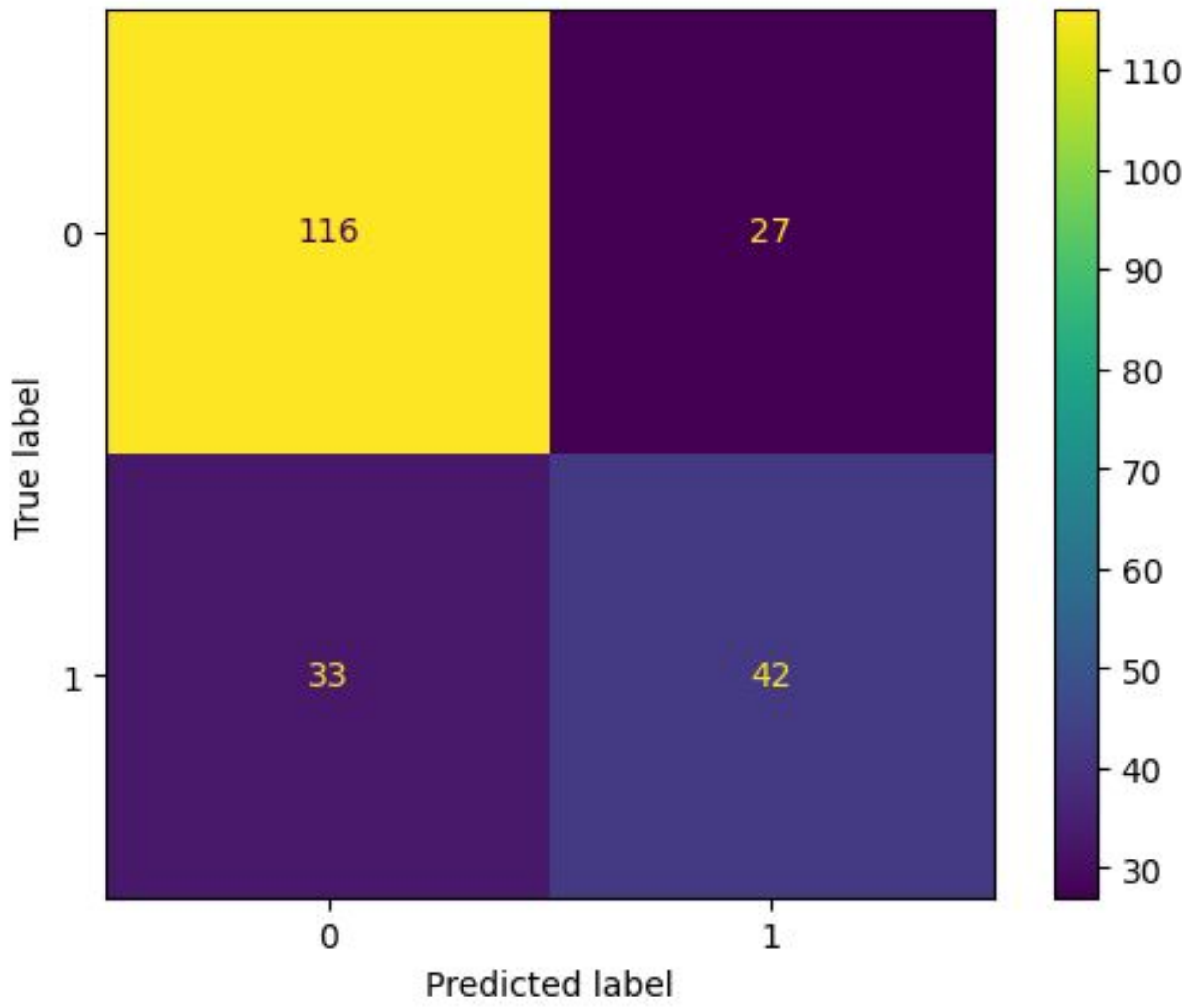
Confusion Matrix

Logistic Model



Accuracy = 0.72936
Precision = 0.63333
Recall = 0.50667
F1 score = 0.56296

Random forest



Accuracy = 0.72477
Precision = 0.60870
Recall = 0.56000
F1 score = 0.58333

“Logistic Regression is **slightly better model**. However it has a **higher false negative** which is a **drawback** of the model for medical prediction”

04.

Future Goals

1. **Data Processing**

- More cleaning of data to better fit the model

2. More Model Fitting

- Due to time constraints, these are just base models
- More model fitting and assessment to enhance the prediction of test value

3. More model

- Creates more supervised learning model and determine best model.
- Compare with real life data to find actual vs prediction to test the model.

THANK YOU