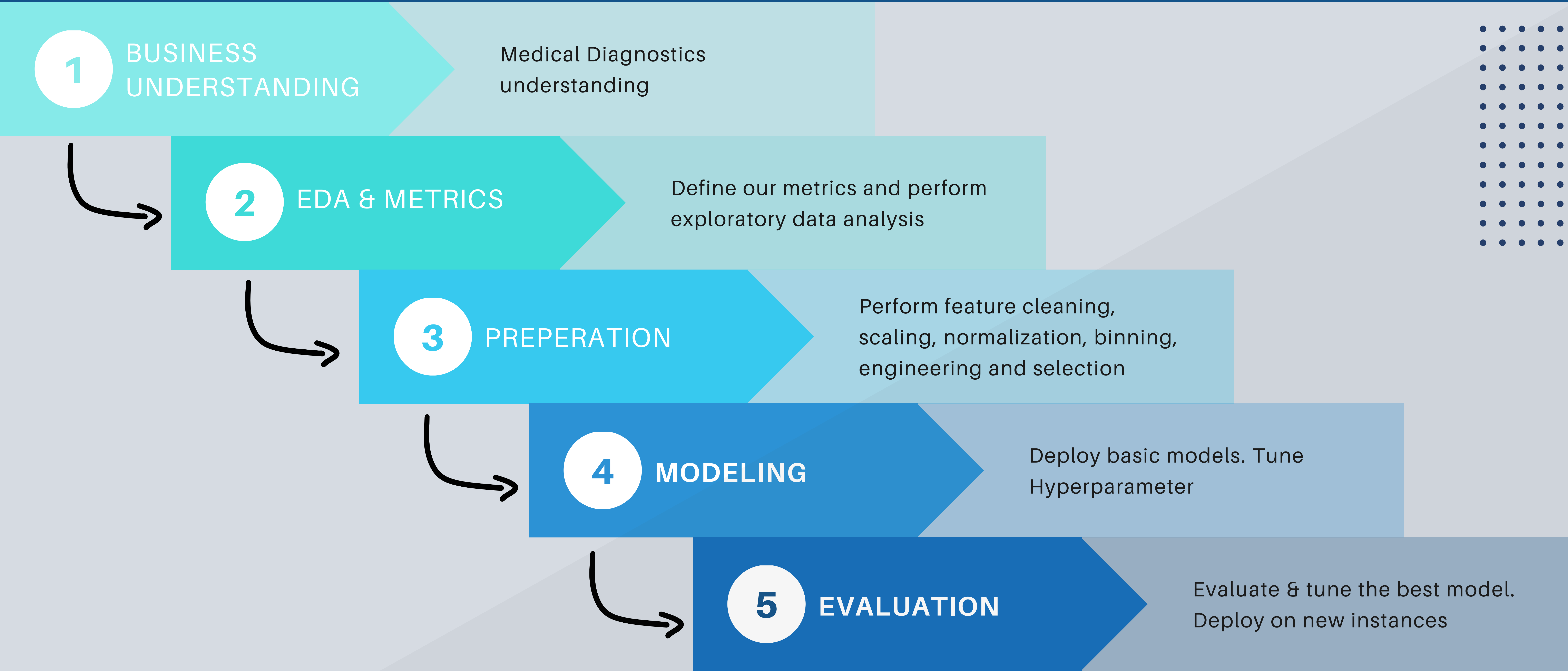
DIABETES PREDICTION

USING THE PIMA INDIAN DATA SET

A medical test to alert the possibility of having the disease.

Ben Tobin
Tomer Treidel

PROJECT STAGES



1 BUSINESS UNDERSTANDING

Medical Diagnostics understanding

2 EDA & METRICS

Define our metrics and perform exploratory data analysis

3 PREPERATION

Perform feature cleaning, scaling, normalization, binning, engineering and selection

4 MODELING

Deploy basic models. Tune Hyperparameter

5 EVALUATION

Evaluate & tune the best model. Deploy on new instances

DIABETES GENERAL UNDESTANDING

TYPE 1 ➤ *Problem Creating Insulin*

TYPE 2 ➤ *Problem Using Insulin*

GESTATIONAL ➤ *Creates Risk to Develop Type 2*

*https://www.cdc.gov/diabetes/library/spotlights/diabetes-facts-stats.html

FIRST LOOK AT DATA SET

Population features:

- Woman of Pima Indian heritage
- Over the age of 21

Data Set:

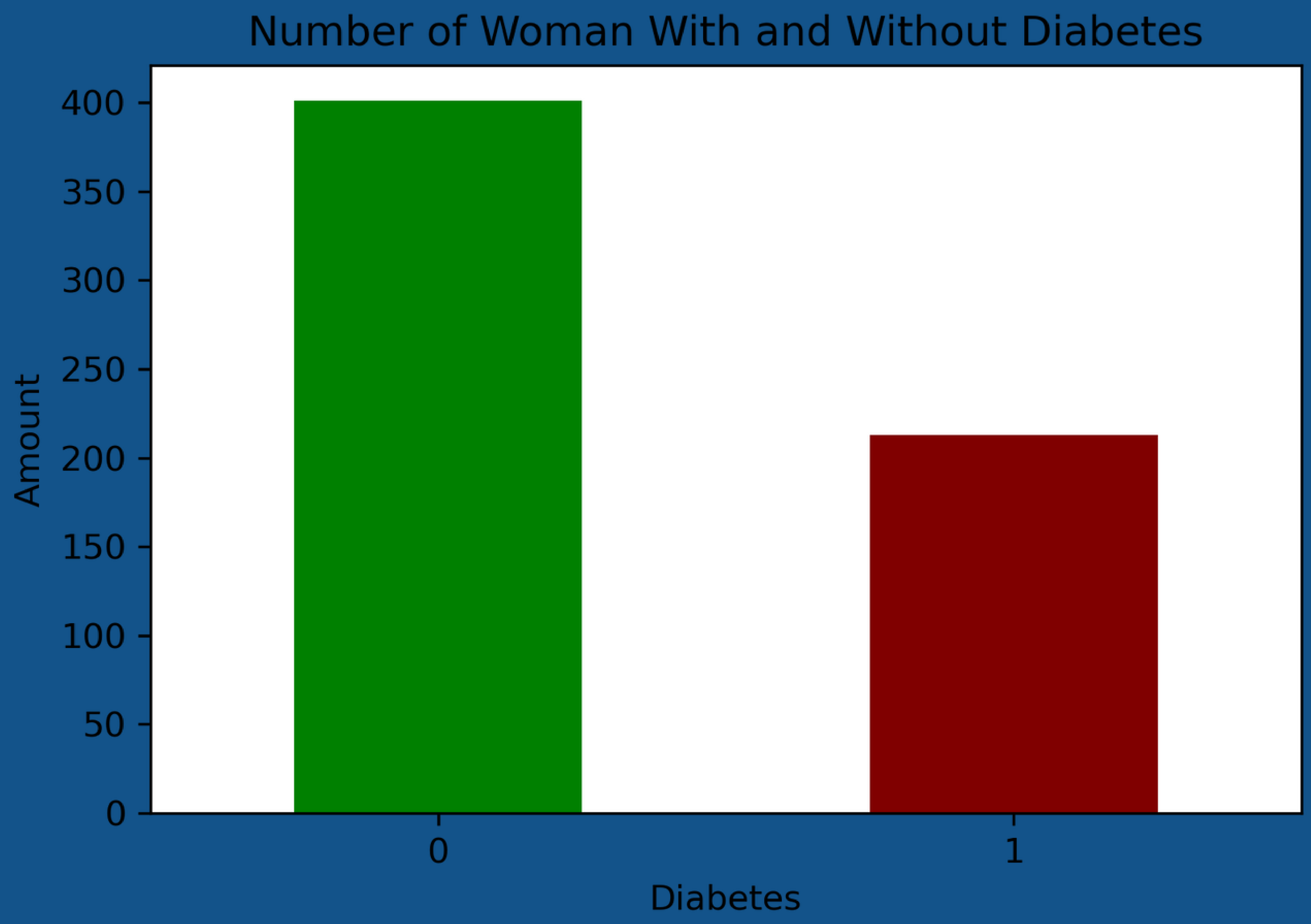
Features - 9

Instances - 768

Target Variable:

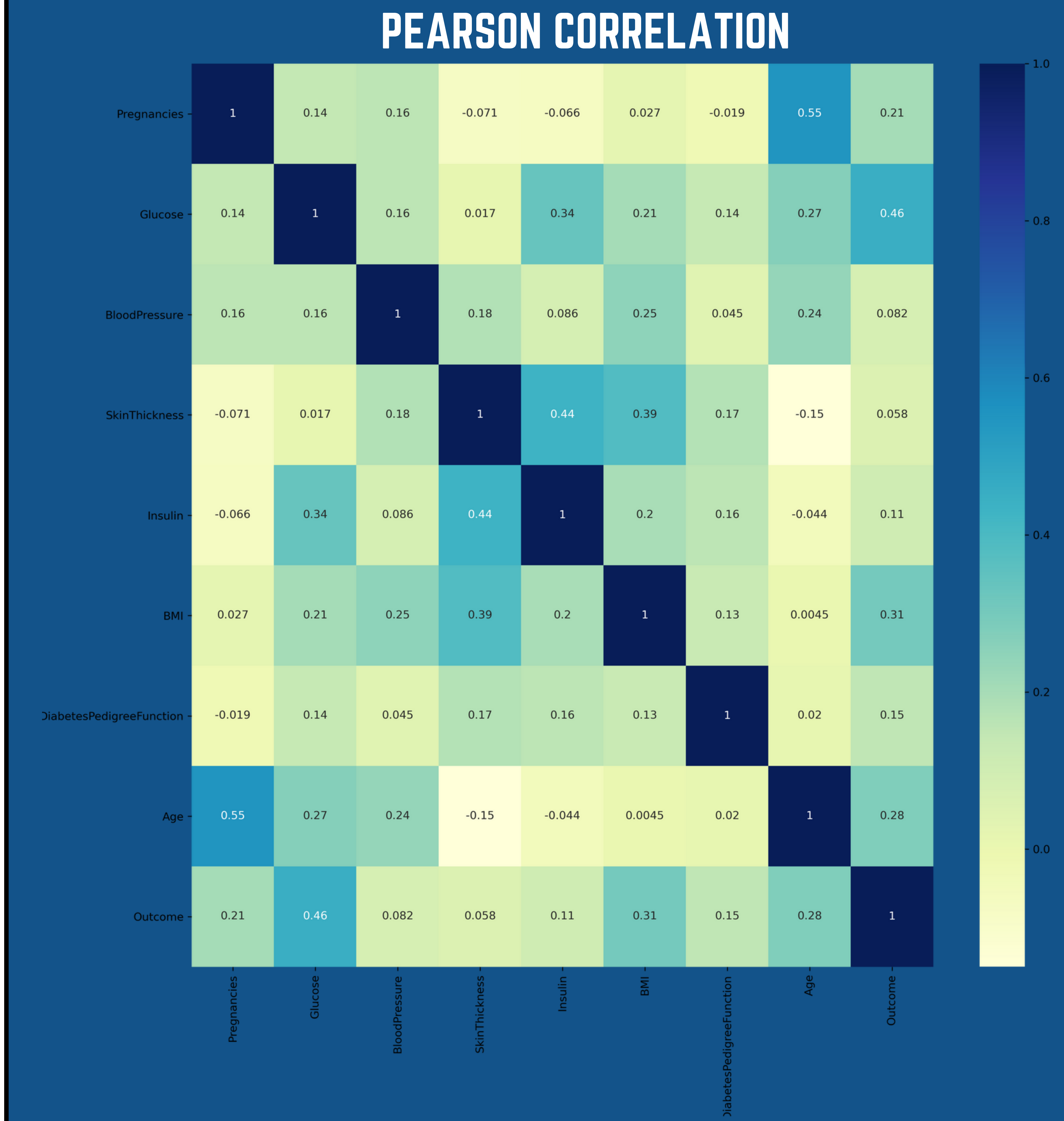
Positive - 35%

Negative- 65%



INITIAL CORRELATION BETWEEN FEATURES

- Low correlations between most features
- Spearman correlation not better
- Somethings wrong



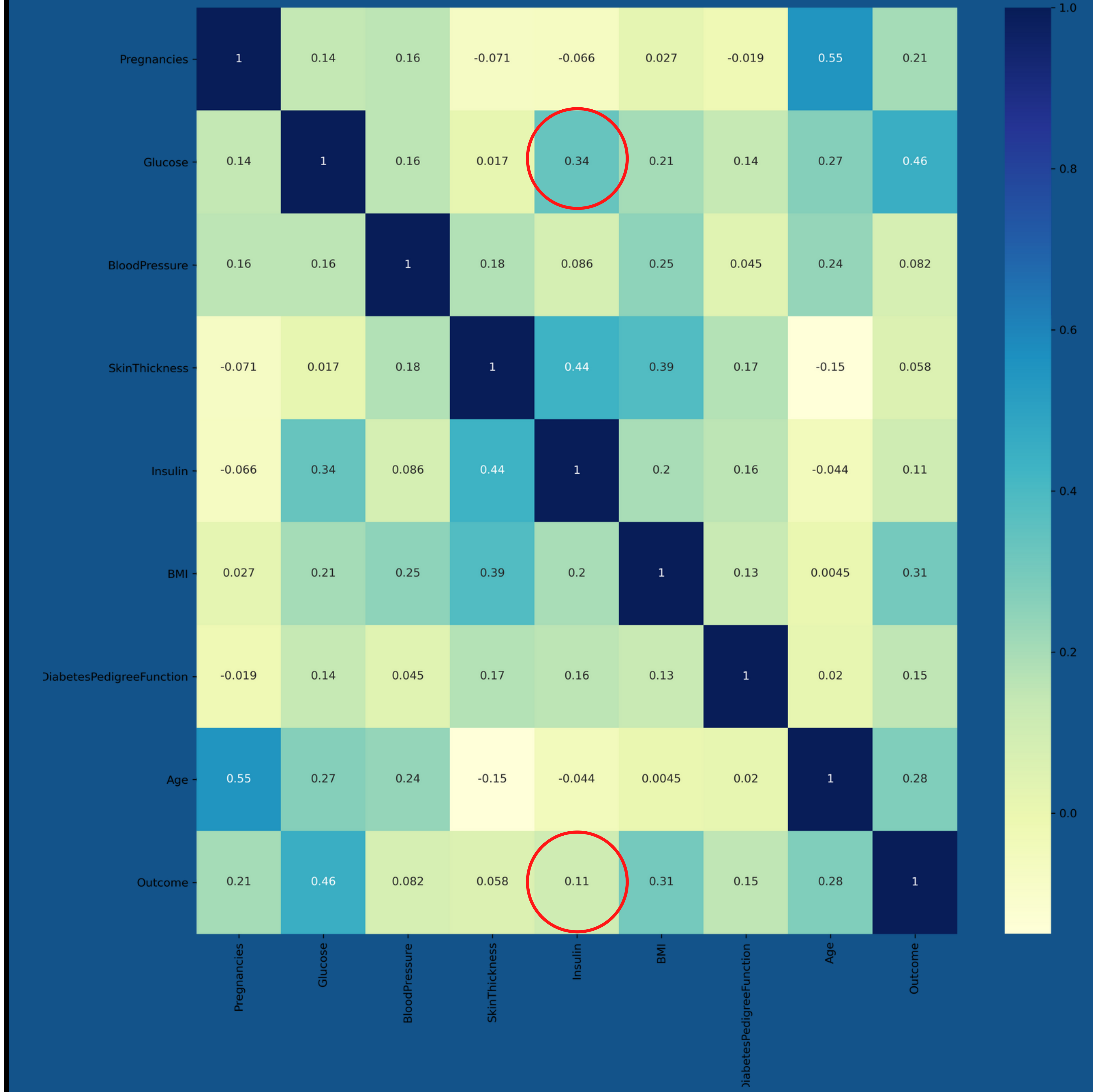
PEARSON CORRELATION

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.14	0.16	-0.071	-0.066	0.027	-0.019	0.55	0.21
Glucose	0.14	1	0.16	0.017	0.34	0.21	0.14	0.27	0.46
BloodPressure	0.16	0.16	1	0.18	0.086	0.25	0.045	0.24	0.082
SkinThickness	-0.071	0.017	0.18	1	0.44	0.39	0.17	-0.15	0.058
Insulin	-0.066	0.34	0.086	0.44	1	0.2	0.16	-0.044	0.11
BMI	0.027	0.21	0.25	0.39	0.2	1	0.13	0.0045	0.31
DiabetesPedigreeFunction	-0.019	0.14	0.045	0.17	0.16	0.13	1	0.02	0.15
Age	0.55	0.27	0.24	-0.15	-0.044	0.0045	0.02	1	0.28
Outcome	0.21	0.46	0.082	0.058	0.11	0.31	0.15	0.28	1

INITIAL CORRELATION BETWEEN FEATURES

- Low correlations between most features
- Spearman correlation not better
- Somethings wrong

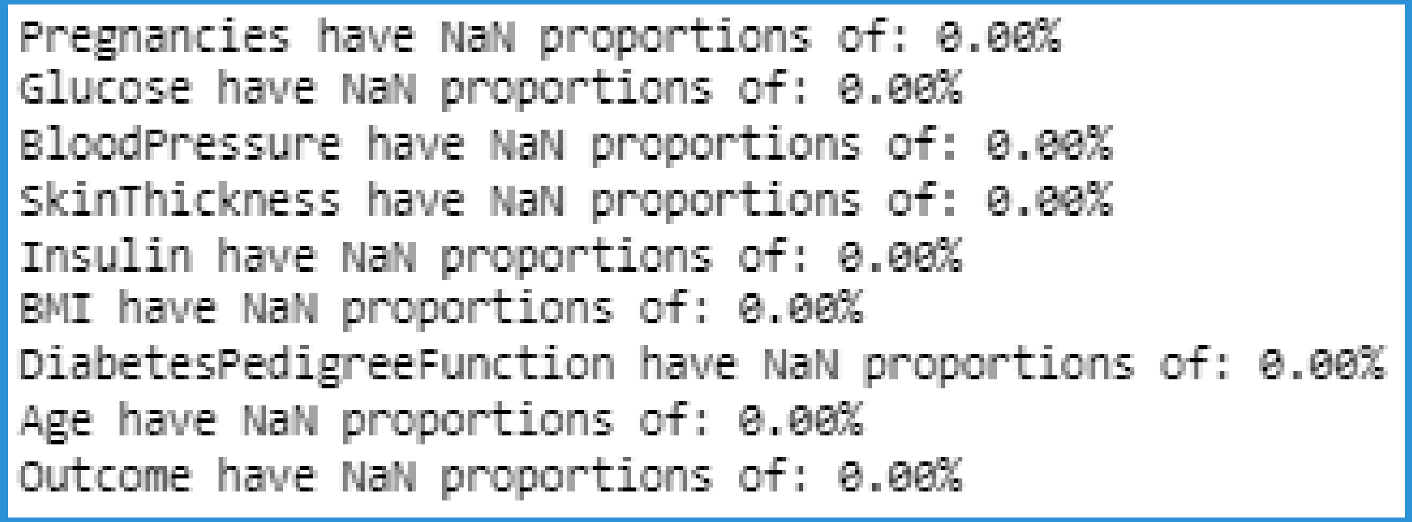
PEARSON CORRELATION



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.14	0.16	-0.071	-0.066	0.027	-0.019	0.55	0.21
Glucose	0.14	1	0.16	0.017	0.34	0.21	0.14	0.27	0.46
BloodPressure	0.16	0.16	1	0.18	0.086	0.25	0.045	0.24	0.082
SkinThickness	-0.071	0.017	0.18	1	0.44	0.39	0.17	-0.15	0.058
Insulin	-0.066	0.34	0.086	0.44	1	0.2	0.16	-0.044	0.11
BMI	0.027	0.21	0.25	0.39	0.2	1	0.13	0.0045	0.31
DiabetesPedigreeFunction	-0.019	0.14	0.045	0.17	0.16	0.13	1	0.02	0.15
Age	0.55	0.27	0.24	-0.15	-0.044	0.0045	0.02	1	0.28
Outcome	0.21	0.46	0.082	0.058	0.11	0.31	0.15	0.28	1

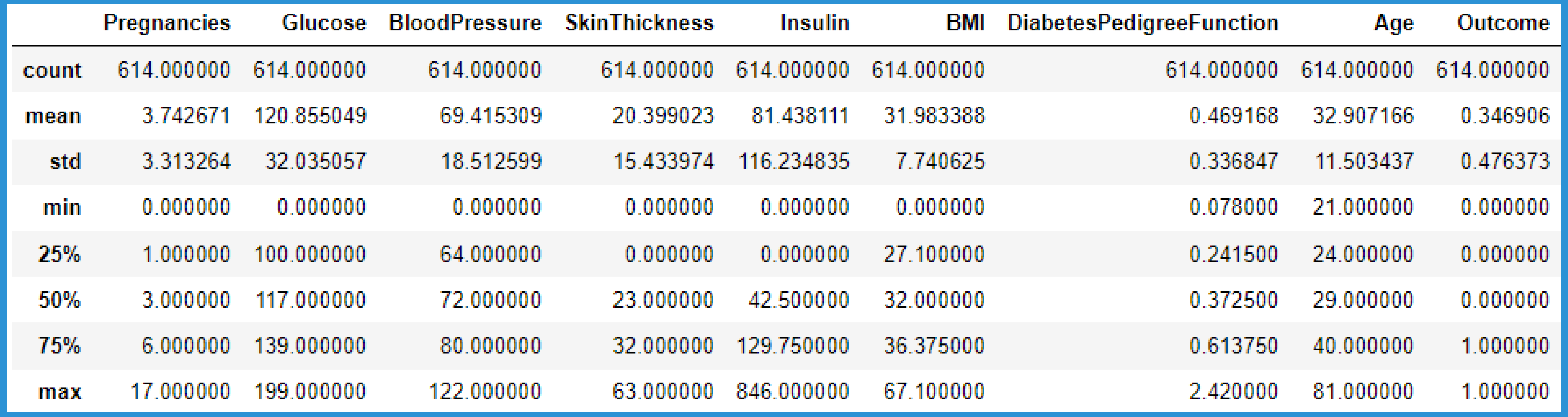
FURTHER EDA - NA & STATISTICAL DISTRIBUTIONS

NAN PROPORTION



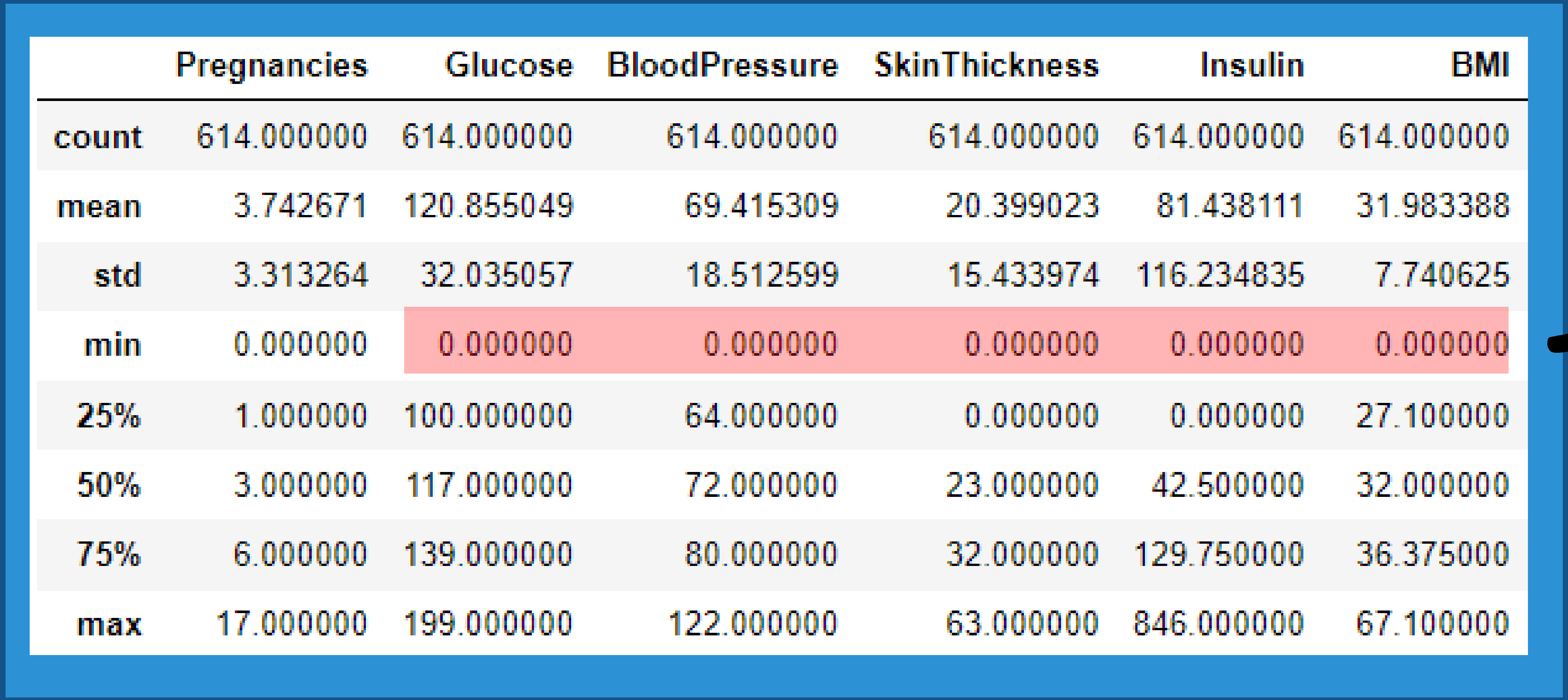
```
Pregnancies have NaN proportions of: 0.00%
Glucose have NaN proportions of: 0.00%
BloodPressure have NaN proportions of: 0.00%
SkinThickness have NaN proportions of: 0.00%
Insulin have NaN proportions of: 0.00%
BMI have NaN proportions of: 0.00%
DiabetesPedigreeFunction have NaN proportions of: 0.00%
Age have NaN proportions of: 0.00%
Outcome have NaN proportions of: 0.00%
```

STATISTICAL DISTRIBUTIONS



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	614.000000	614.000000	614.000000	614.000000	614.000000	614.000000	614.000000	614.000000	614.000000
mean	3.742671	120.855049	69.415309	20.399023	81.438111	31.983388	0.469168	32.907166	0.346906
std	3.313264	32.035057	18.512599	15.433974	116.234835	7.740625	0.336847	11.503437	0.476373
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	100.000000	64.000000	0.000000	0.000000	27.100000	0.241500	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	42.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	139.000000	80.000000	32.000000	129.750000	36.375000	0.613750	40.000000	1.000000
max	17.000000	199.000000	122.000000	63.000000	846.000000	67.100000	2.420000	81.000000	1.000000

CLOSER LOOK AT STATISTICAL ATTRIBUTES

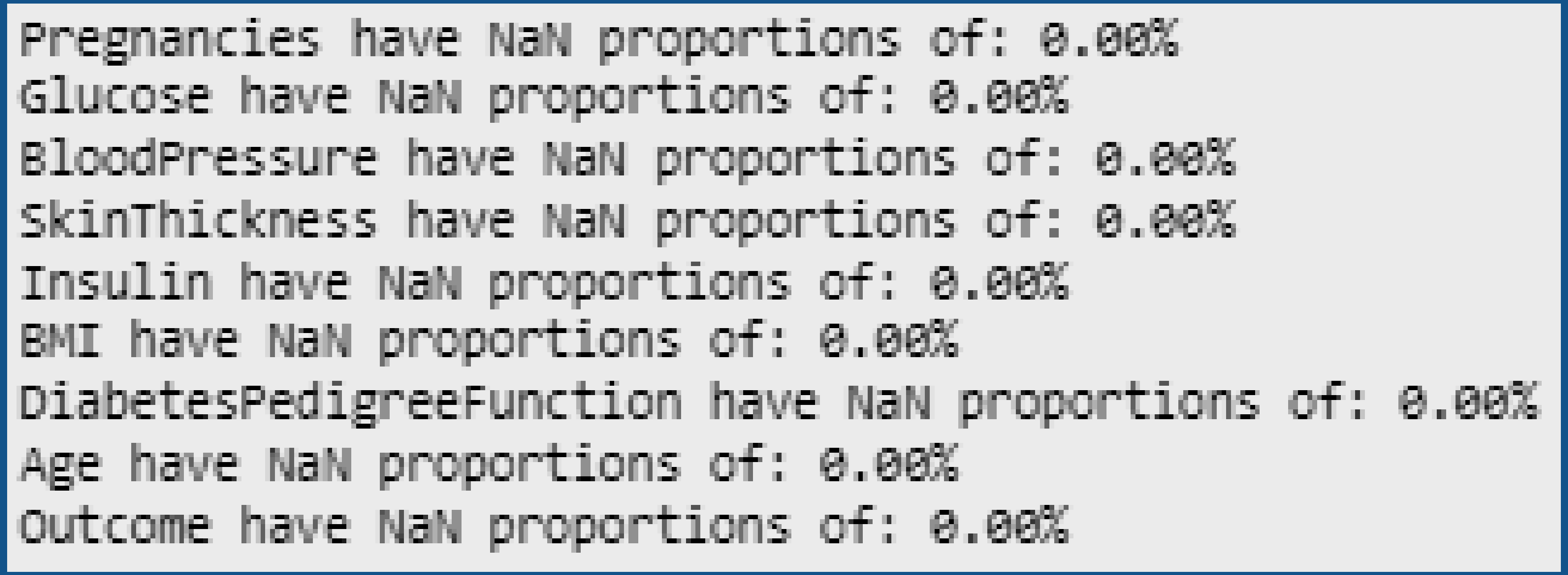


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
count	614.000000	614.000000	614.000000	614.000000	614.000000	614.000000
mean	3.742671	120.855049	69.415309	20.399023	81.438111	31.983388
std	3.313264	32.035057	18.512599	15.433974	116.234835	7.740625
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	100.000000	64.000000	0.000000	0.000000	27.100000
50%	3.000000	117.000000	72.000000	23.000000	42.500000	32.000000
75%	6.000000	139.000000	80.000000	32.000000	129.750000	36.375000
max	17.000000	199.000000	122.000000	63.000000	846.000000	67.100000

Medical tests that can't result in 0

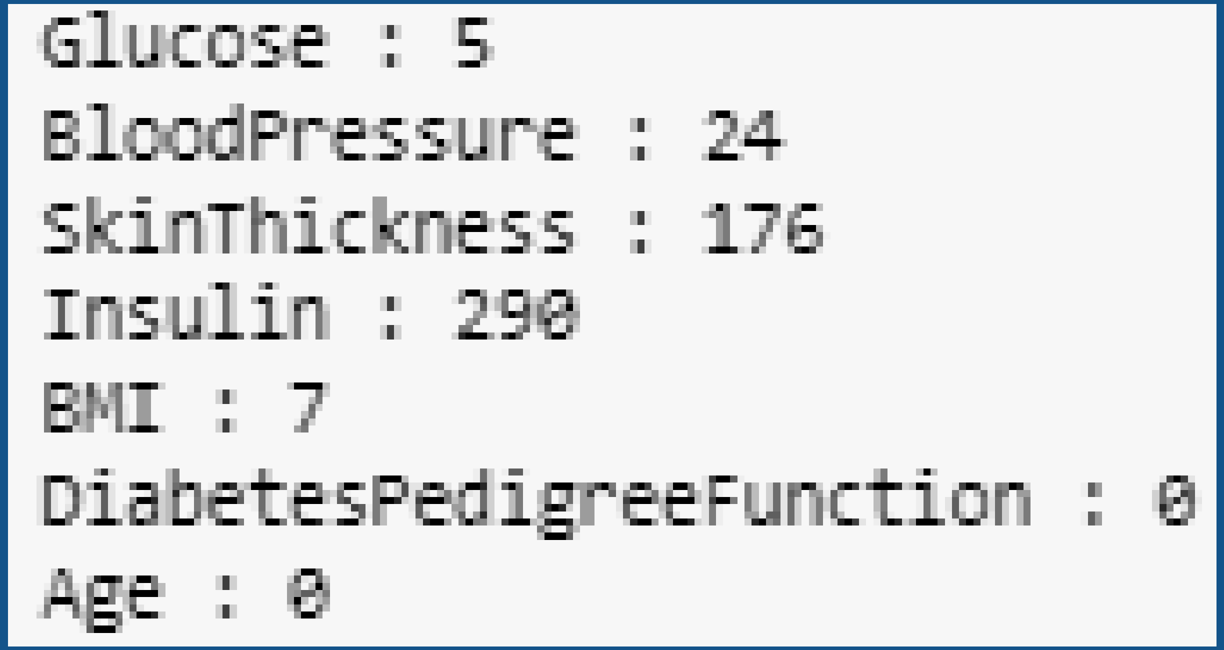
CONCLUSION:
0’S IN THOSE COLMNS ARE NA’S

NAN LEVEL



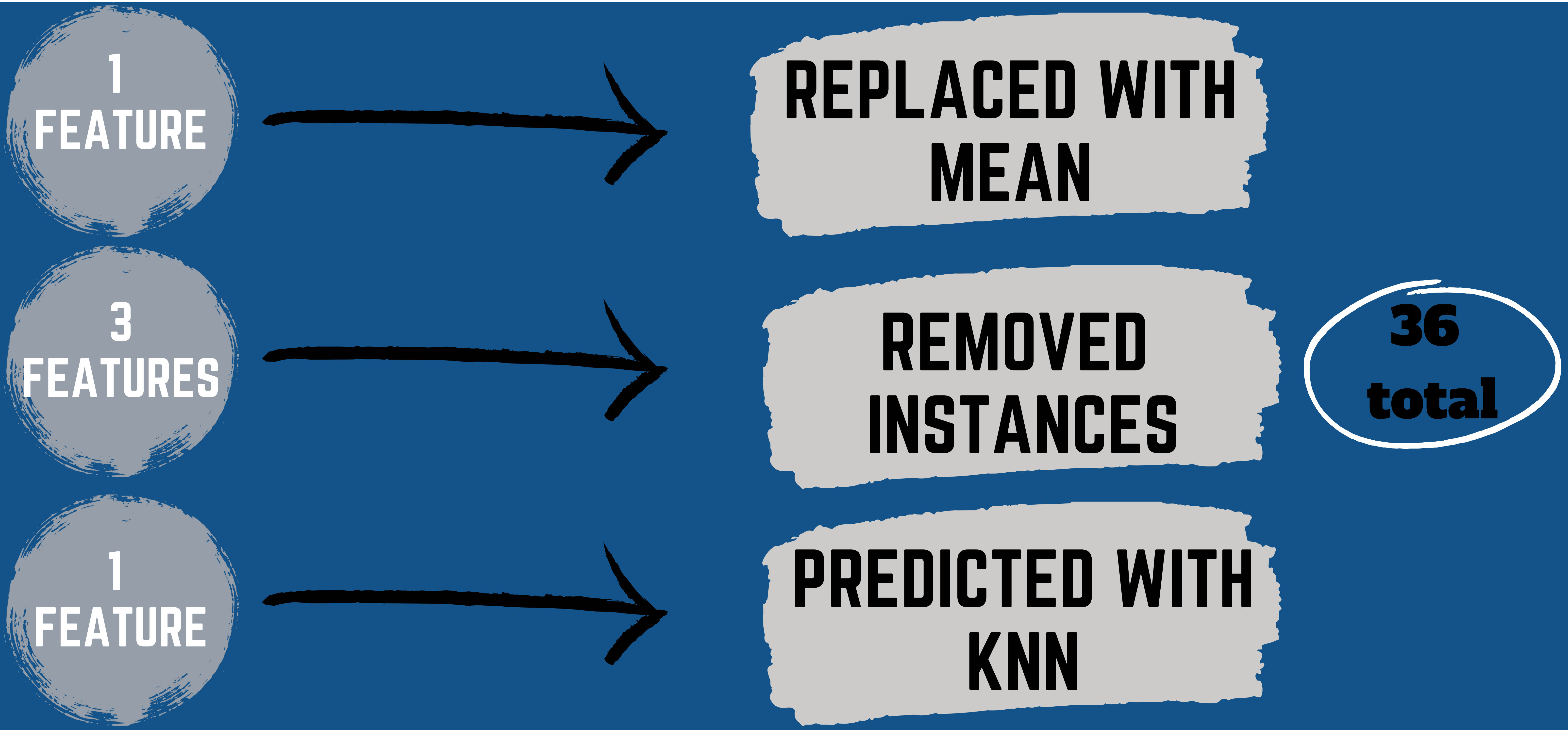
```
Pregnancies have NaN proportions of: 0.00%
Glucose have NaN proportions of: 0.00%
BloodPressure have NaN proportions of: 0.00%
SkinThickness have NaN proportions of: 0.00%
Insulin have NaN proportions of: 0.00%
BMI have NaN proportions of: 0.00%
DiabetesPedigreeFunction have NaN proportions of: 0.00%
Age have NaN proportions of: 0.00%
Outcome have NaN proportions of: 0.00%
```

ZERO’S LEVEL



```
Glucose : 5
BloodPressure : 24
SkinThickness : 176
Insulin : 290
BMI : 7
DiabetesPedigreeFunction : 0
Age : 0
```

CLEANING DATA - DEALING WITH NA'S



1 FEATURE → REPLACED WITH MEAN

3 FEATURES → REMOVED INSTANCES

36 total

1 FEATURE → PREDICTED WITH KNN

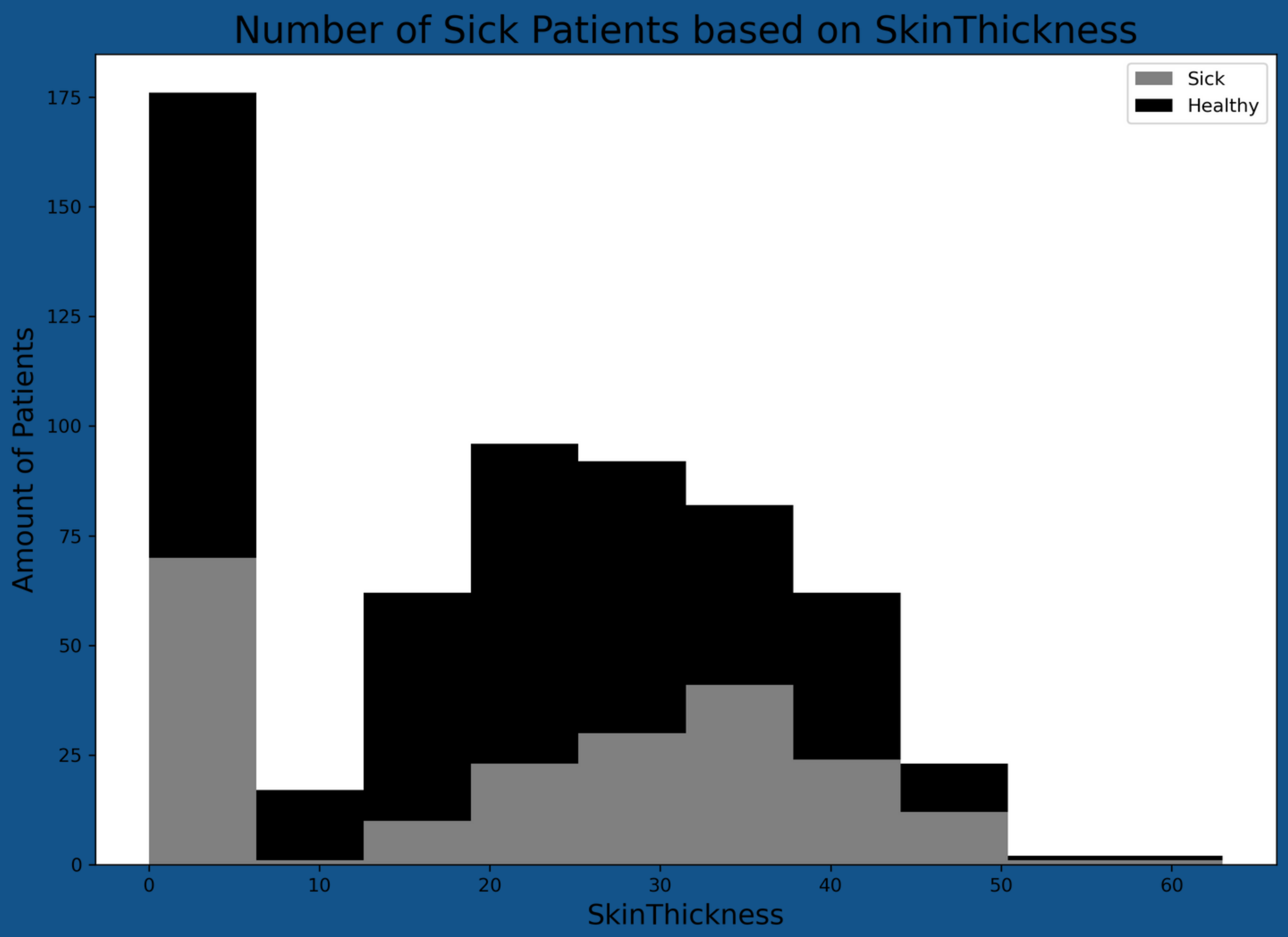
DEALING WITH NA'S
SKIN THICKNESS - BEFORE

MAX
63

MEAN
20.39

MIN
0

STD
15.43



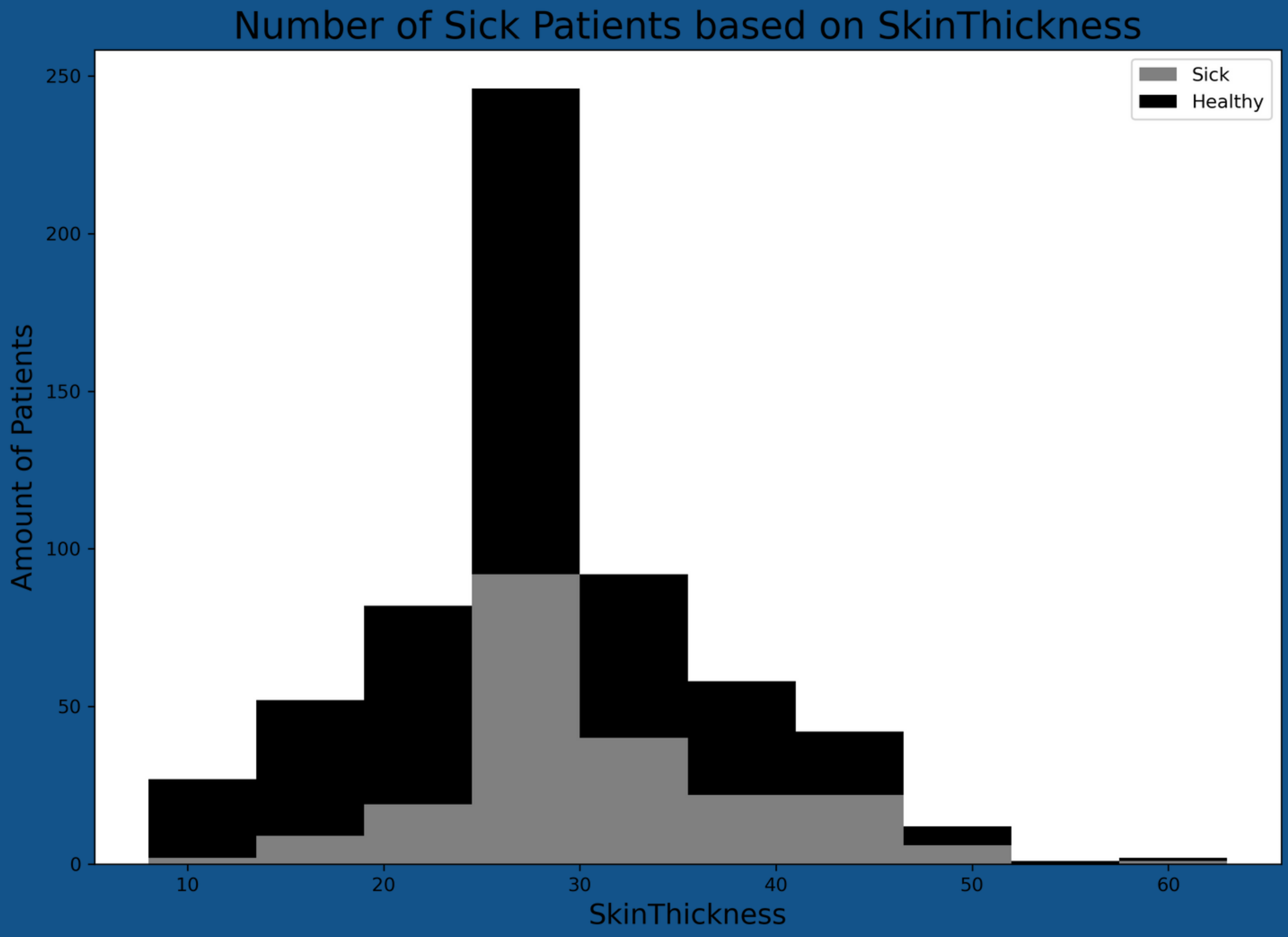
DEALING WITH NA'S
SKIN THICKNESS - AFTER (REPLACE MEAN)

MAX
63

MEAN
28.61

MIN
8

STD
8.60



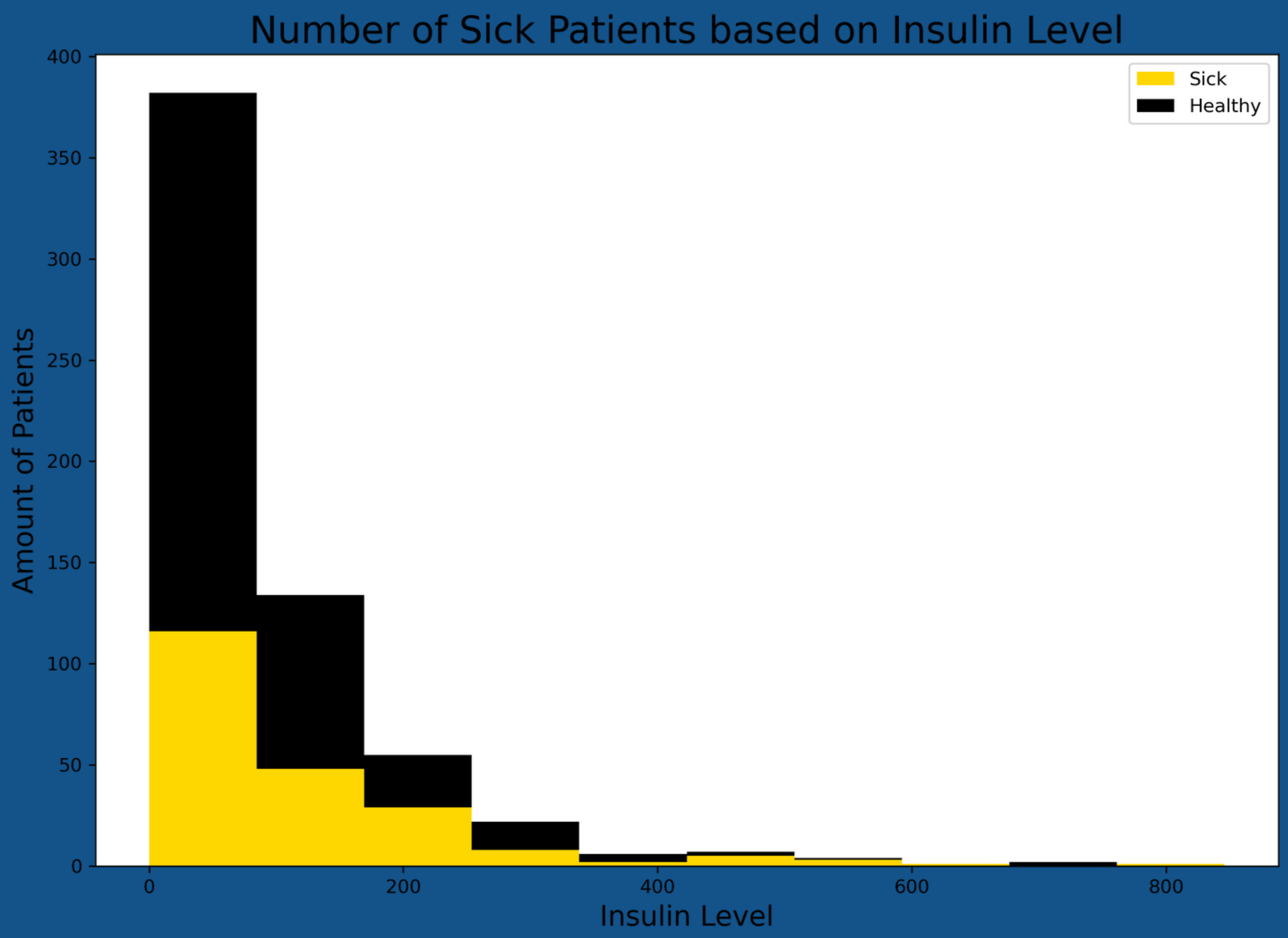
DEALING WITH NA'S INSULIN LEVEL- BEFORE

MAX
846

MEAN
81.43

MIN
0

STD
116.23



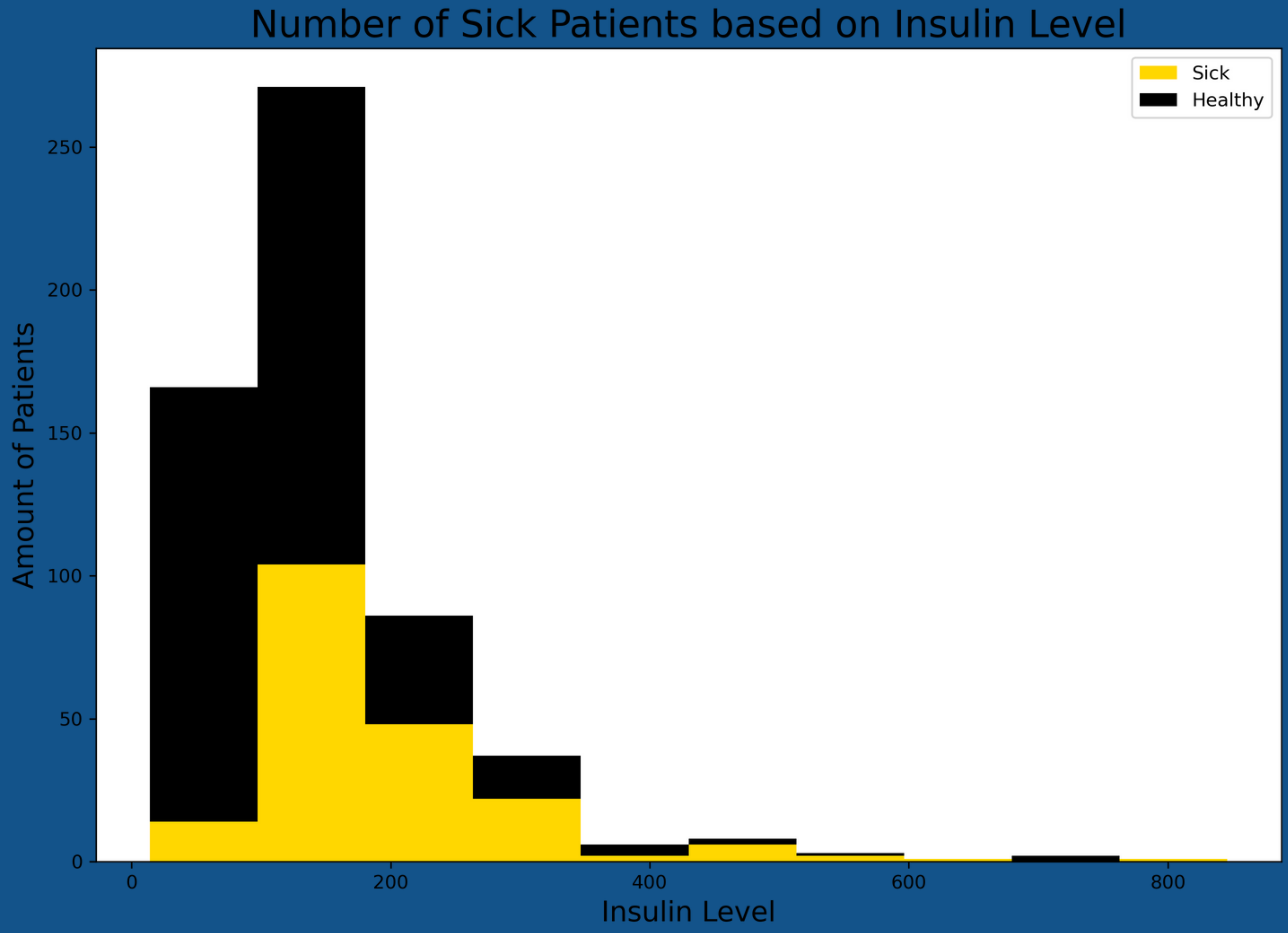
DEALING WITH NA'S
INSULIN LEVEL- AFTER (USING KNN)

MAX
846

MEAN
152.39

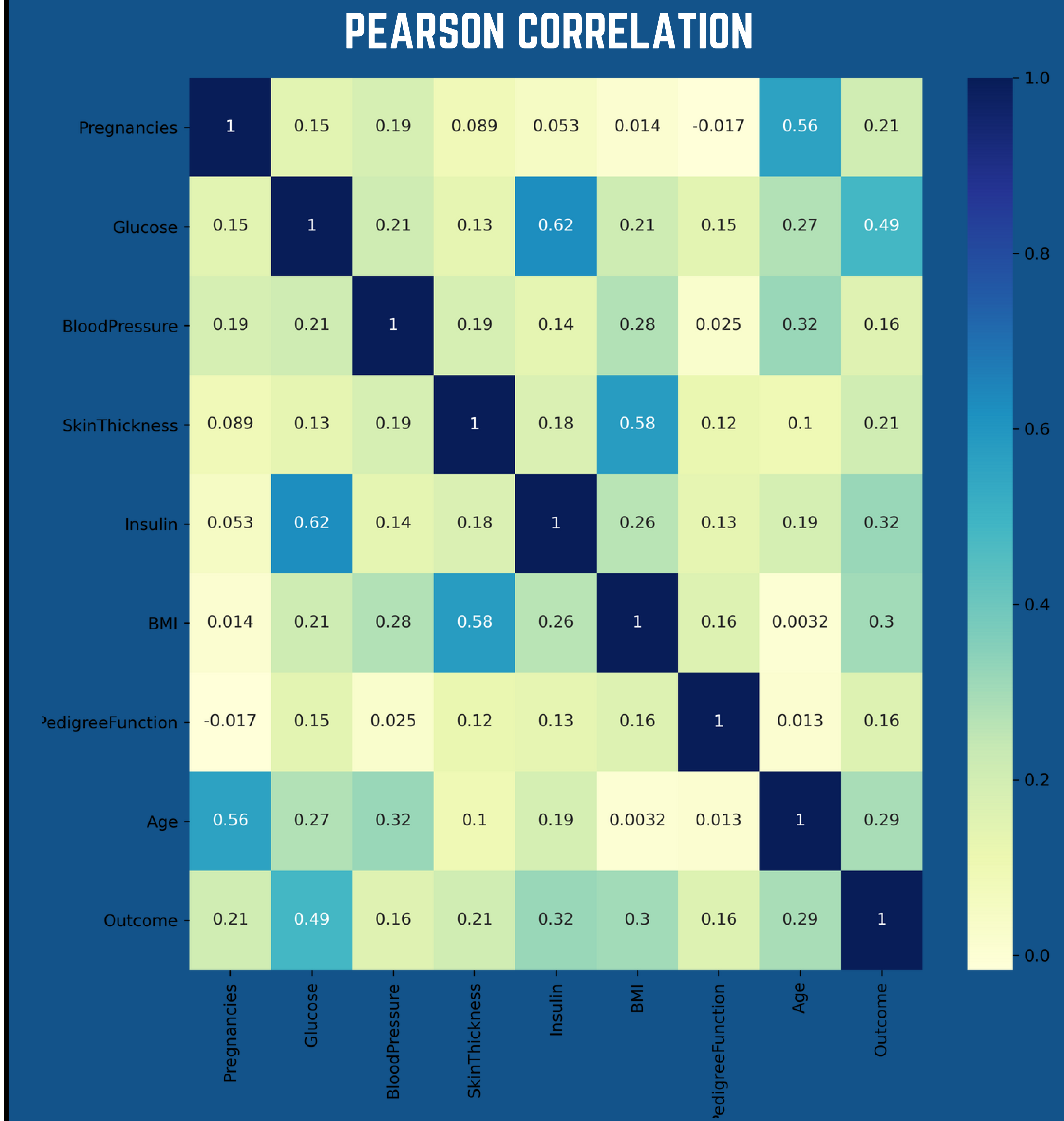
MIN
14

STD
97



NEW CORRELATION BETWEEN FEATURES

- Glucose - Insulin
- BMI - SkinThickness



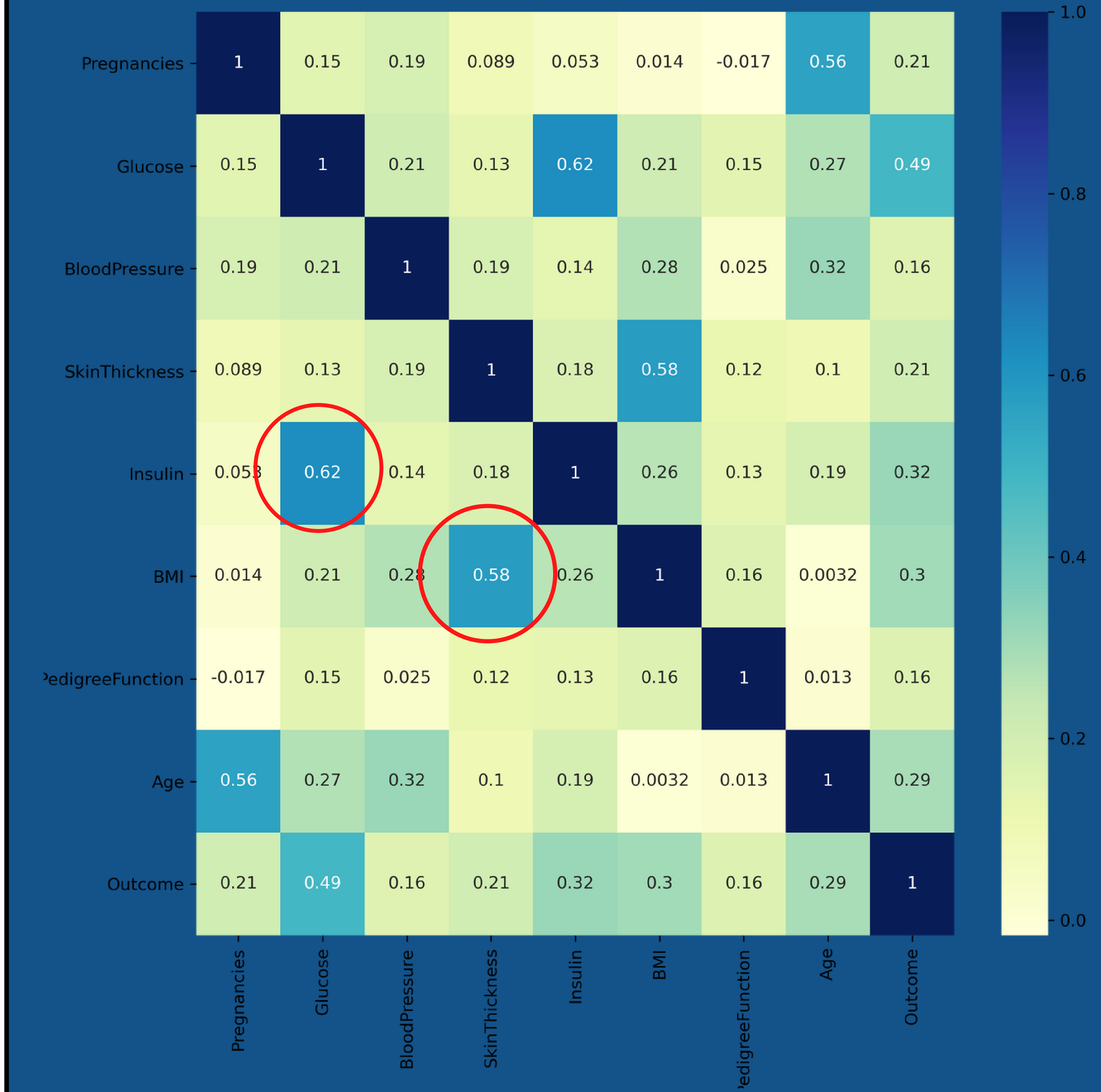
PEARSON CORRELATION

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	edigreeFunction	Age	Outcome
Pregnancies	1	0.15	0.19	0.089	0.053	0.014	-0.017	0.56	0.21
Glucose	0.15	1	0.21	0.13	0.62	0.21	0.15	0.27	0.49
BloodPressure	0.19	0.21	1	0.19	0.14	0.28	0.025	0.32	0.16
SkinThickness	0.089	0.13	0.19	1	0.18	0.58	0.12	0.1	0.21
Insulin	0.053	0.62	0.14	0.18	1	0.26	0.13	0.19	0.32
BMI	0.014	0.21	0.28	0.58	0.26	1	0.16	0.0032	0.3
edigreeFunction	-0.017	0.15	0.025	0.12	0.13	0.16	1	0.013	0.16
Age	0.56	0.27	0.32	0.1	0.19	0.0032	0.013	1	0.29
Outcome	0.21	0.49	0.16	0.21	0.32	0.3	0.16	0.29	1

NEW CORRELATION BETWEEN FEATURES

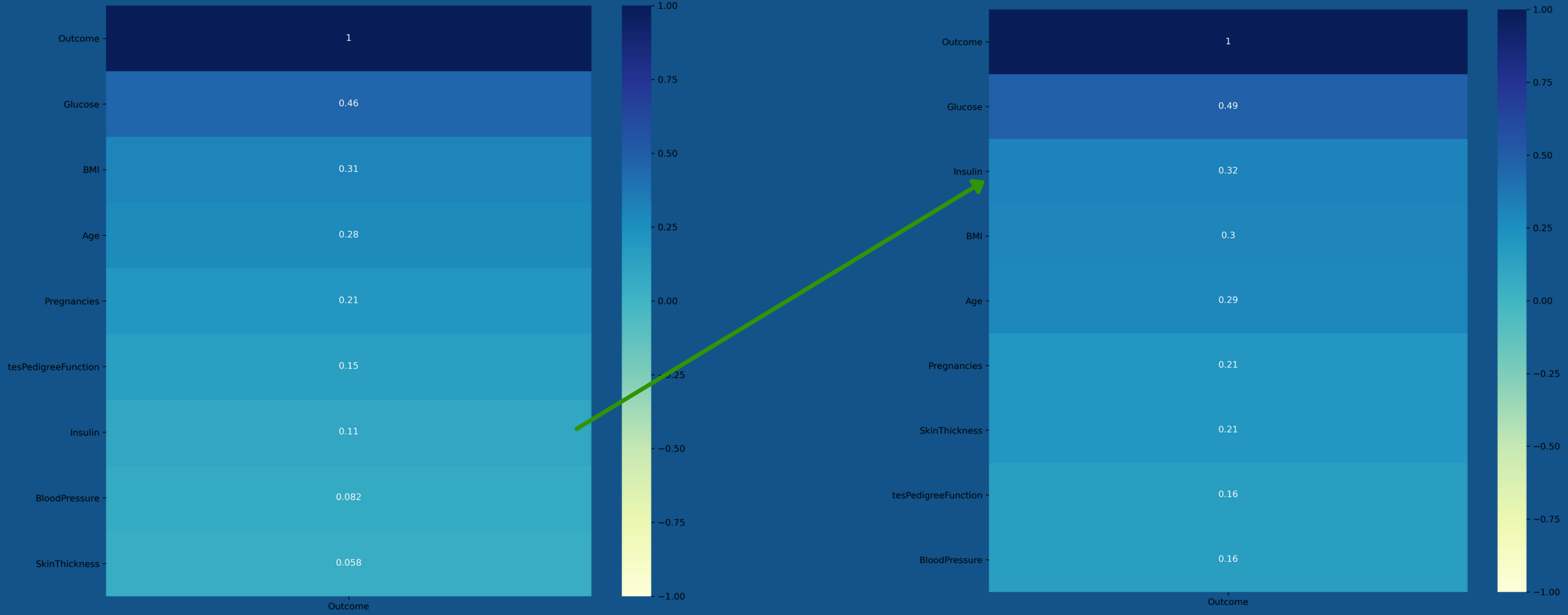
- Glucose - Insulin
- BMI - SkinThickness

PEARSON CORRELATION

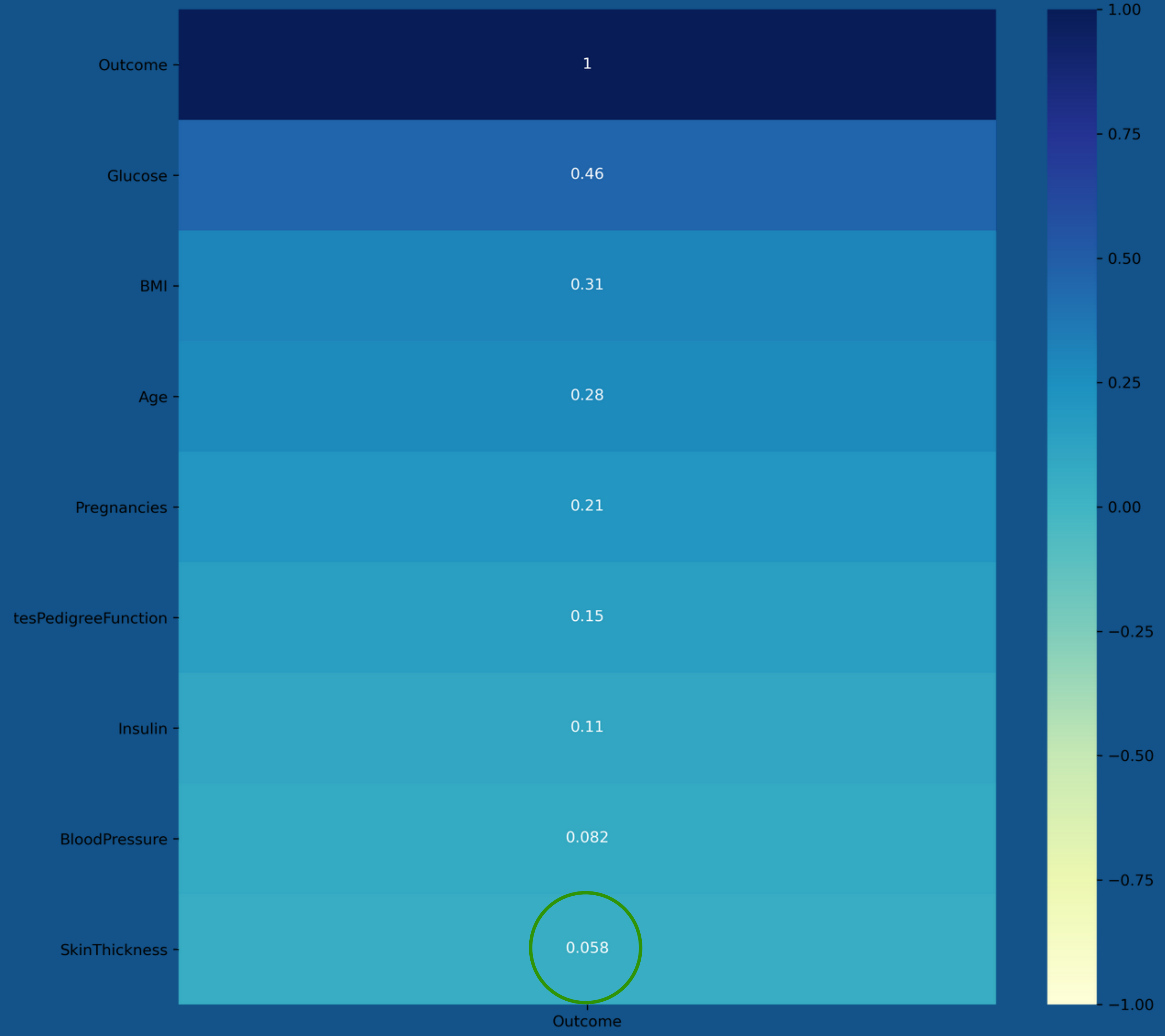


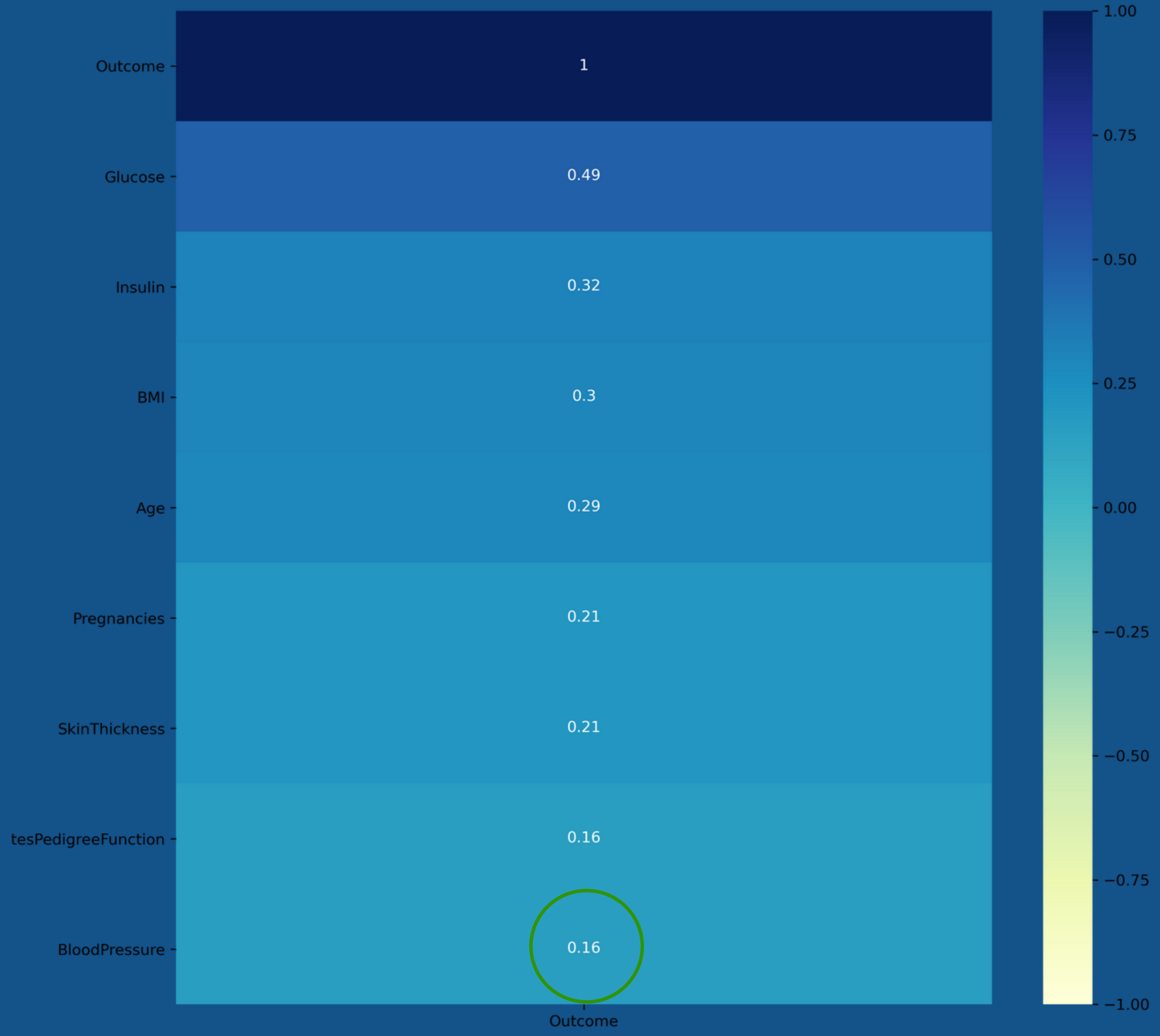
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	edigreeFunction	Age	Outcome
Pregnancies	1	0.15	0.19	0.089	0.053	0.014	-0.017	0.56	0.21
Glucose	0.15	1	0.21	0.13	0.62	0.21	0.15	0.27	0.49
BloodPressure	0.19	0.21	1	0.19	0.14	0.28	0.025	0.32	0.16
SkinThickness	0.089	0.13	0.19	1	0.18	0.58	0.12	0.1	0.21
Insulin	0.053	0.62	0.14	0.18	1	0.26	0.13	0.19	0.32
BMI	0.014	0.21	0.28	0.58	0.26	1	0.16	0.0032	0.3
edigreeFunction	-0.017	0.15	0.025	0.12	0.13	0.16	1	0.013	0.16
Age	0.56	0.27	0.32	0.1	0.19	0.0032	0.013	1	0.29
Outcome	0.21	0.49	0.16	0.21	0.32	0.3	0.16	0.29	1

CORRELATIONS TO TARGET VARIABLE

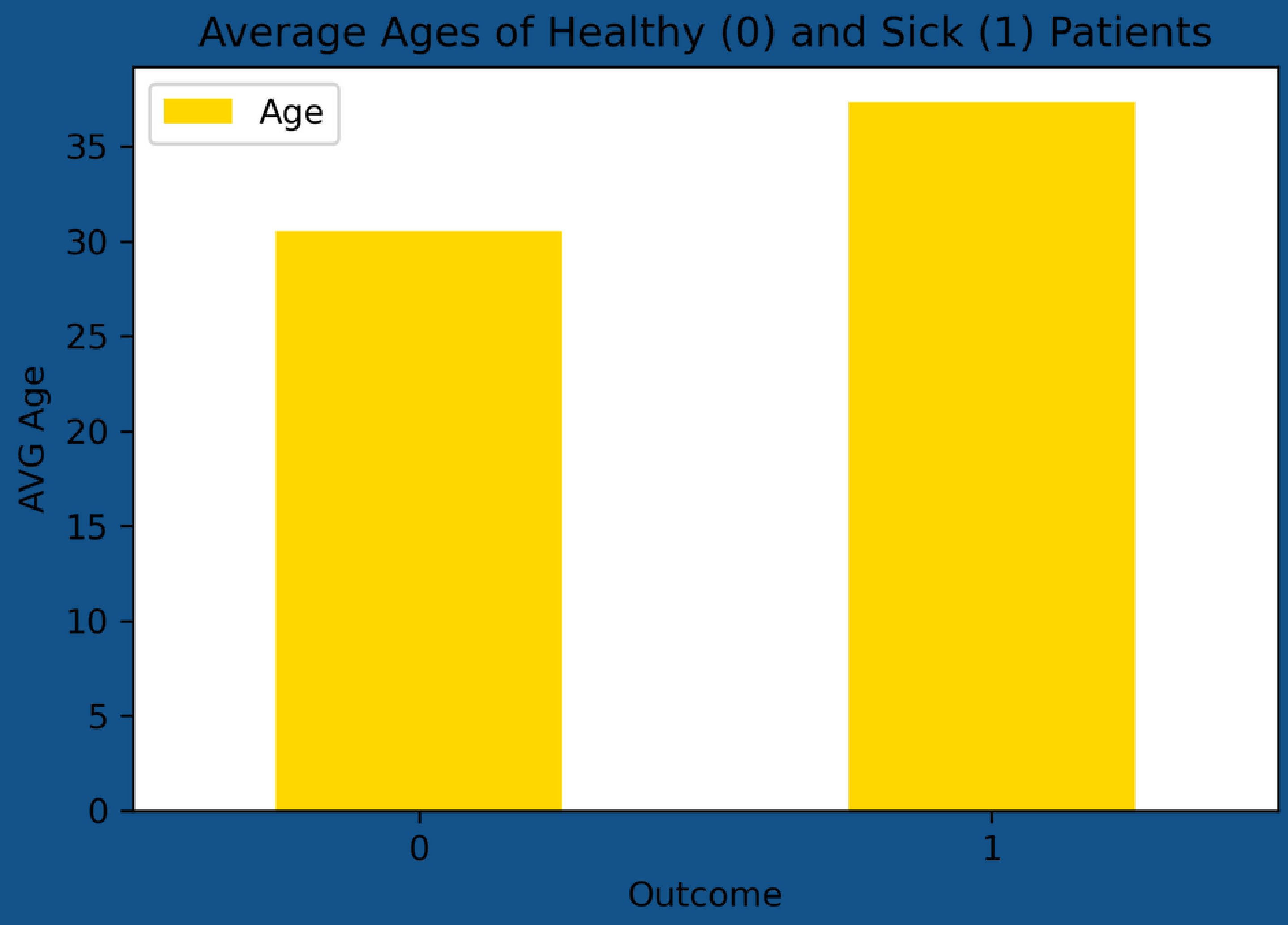


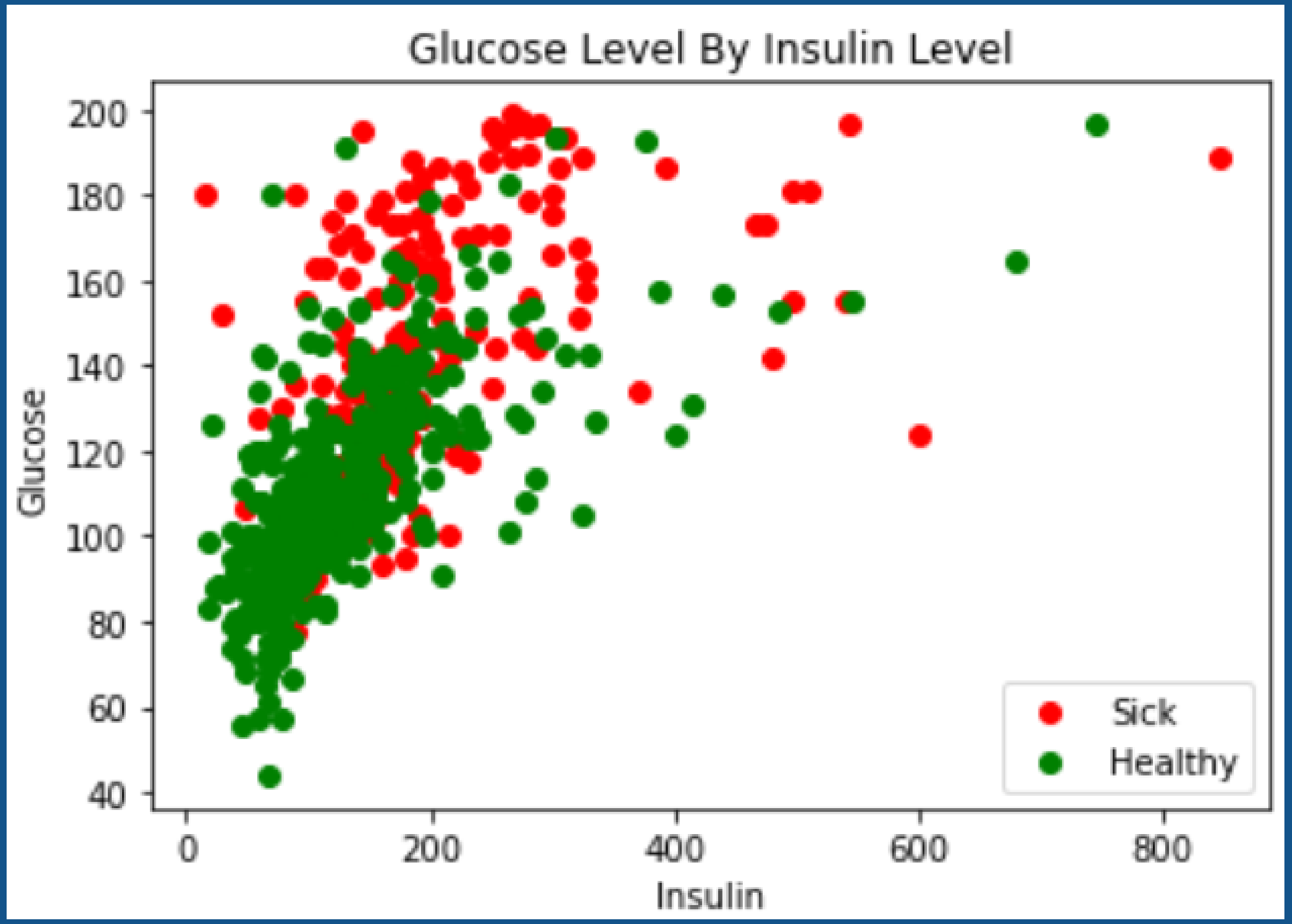
CORRELATIONS TO TARGET VARIABLE





A FEW INTERESTING RELATIONSHIPS





NEW DATA SET - DF1

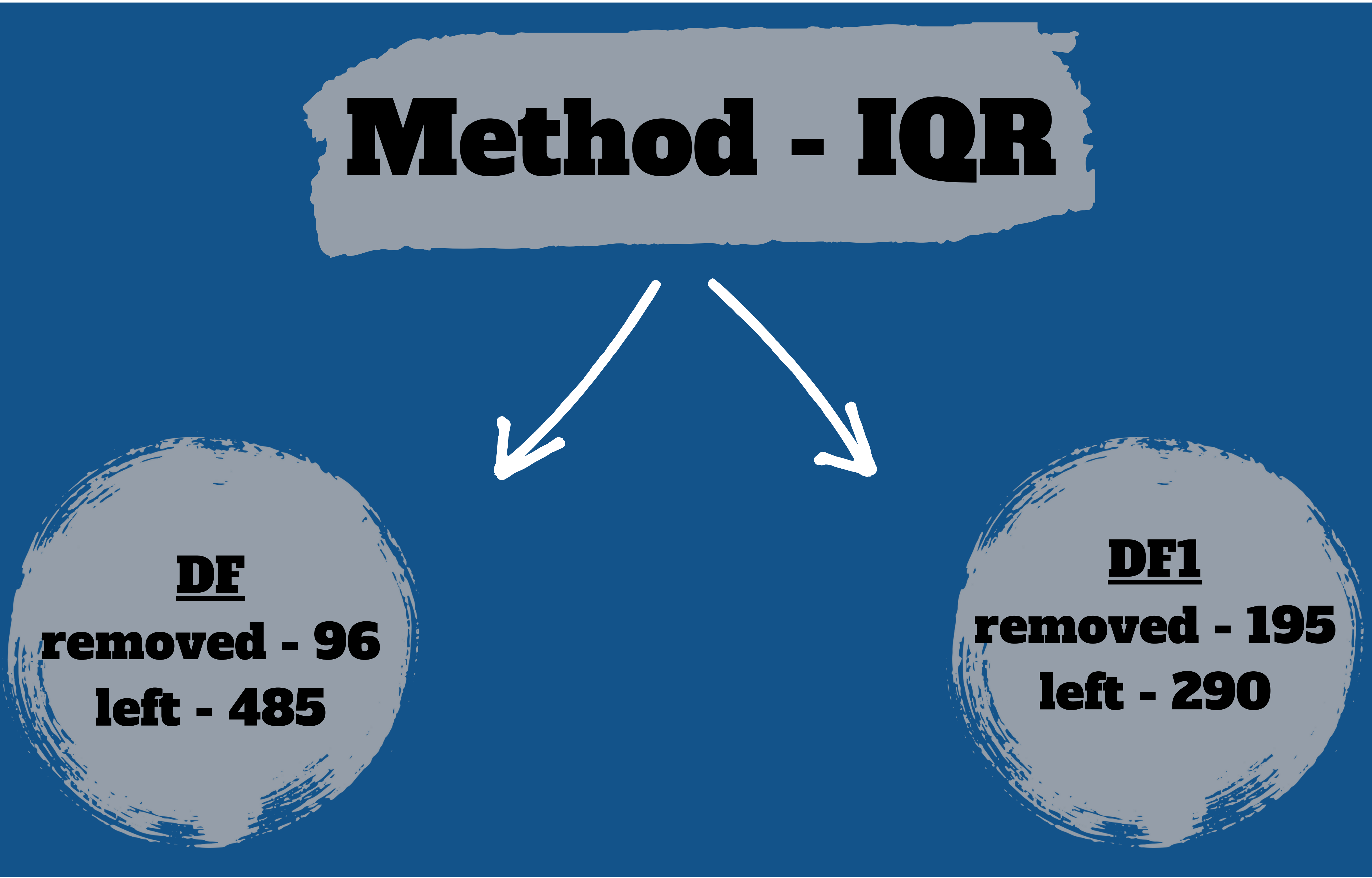
Features
25

Instances
290

Feature Manipulation and Engineering Features:

- Glucose
- BloodPresure
- Insulin
- DPF
- Age
- IG_ratio
- Glucose^2
- BP^2
- Insulin^2
- BMI^2
- DPF^2
- Age^2
- Glucose^3
- BP^3
- ST^3
- Insulin^3
- BMI^3
- DPF^3
- Age^3
- BMI
- Preg
- Preg^2
- Preg^3
- ST
- ST^2

CLEANING DATA -OUTLIER REMOVAL



Method - IQR

DF

removed - 96

left - 485

DF1

removed - 195

left - 290

MODEL + METRIC

MODEL

Random Forest

Best performer on base model with data

METRIC

Accuracy

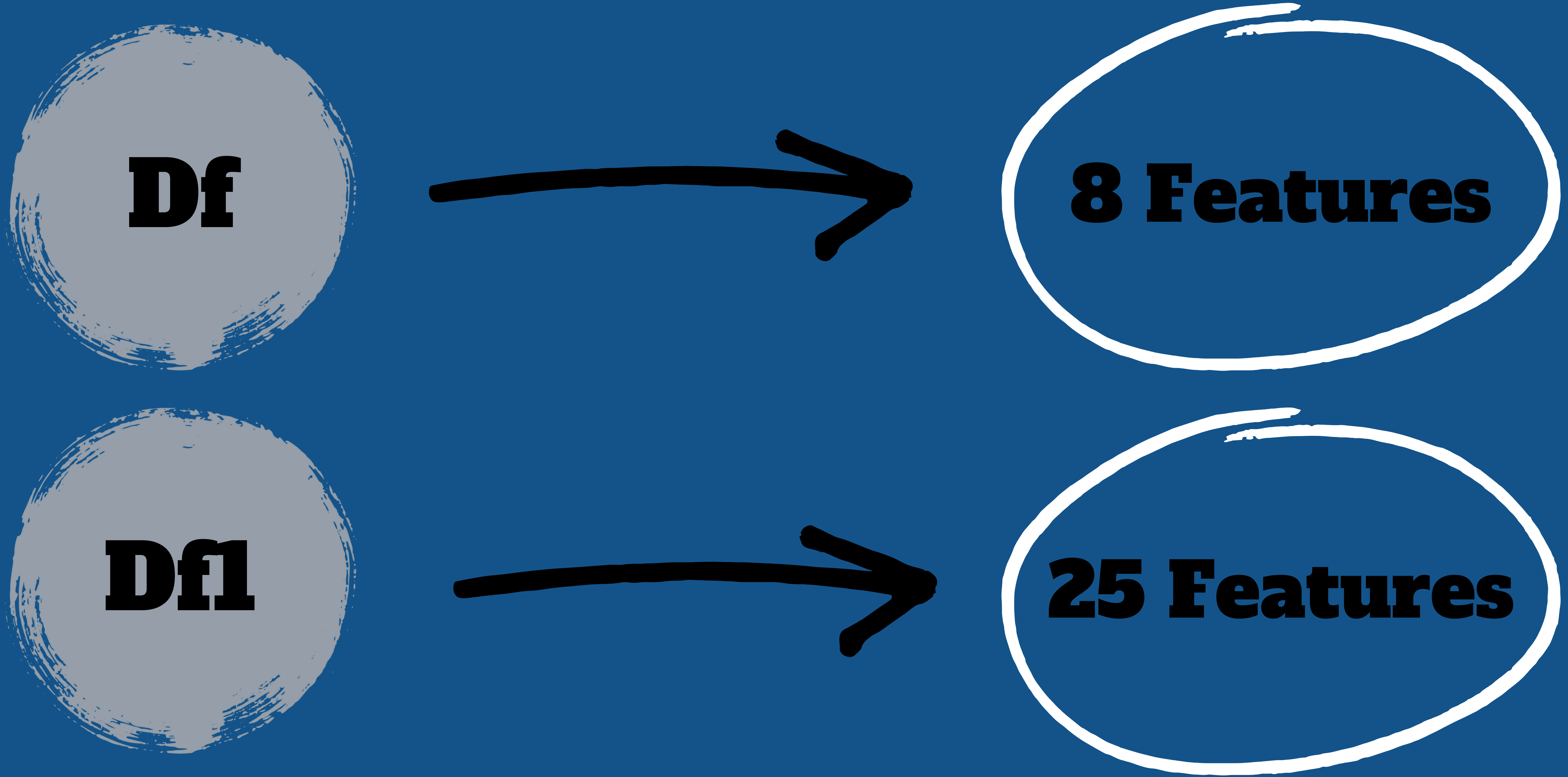
we chose this because in our eyes false positive is just as dangerous as false negative

BASLINE MODEL

65% Accuracy

Based on majority of initial target variable

DATA SETS



Df → 8 Features

Df1 → 25 Features

DF1 - FEATURE SELECTED

Features
25

Instances
290

Feature Selection Method - RFECV

Features:

- Glucose
- BloodPresure
- Insulin
- DPF
- Age
- IG_ratio
- Glucose^2
- BP^2
- Insulin^2
- BMI^2
- DPF^2
- Age^2
- Glucose^3
- BP^3
- ST^3
- Insulin^3
- BMI^3
- DPF^3
- Age^3
- BMI
- Preg
- Preg^2
- Preg^3
- ST
- ST^2

DATA OPTIMIZATION - DATA TYPES

df

Features - 8

Instances - 485

df1

Features - 25

Instances - 290

Data Set	Normalized	Feature Selection	Accuracy
df	✗	✗	76.29%
df1	✗	✗	81.03%
df	✓	✗	76.29%
df1	✓	✗	81.03%
df1	✗	✓	84.48%
df1	✓	✓	84.48%

DATA OPTIMIZATION - CHOSEN DATA

Normalization had no effect on performance

Data Set	Normalized	Feature Selection	Accuracy
df	✗	✗	76.29%
df1	✗	✗	81.03%
df	✓	✗	76.29%
df1	✓	✗	81.03%
df1	✗	✓	84.48%
df1	✓	✓	84.48%

HYPER PARAMETER TUNING

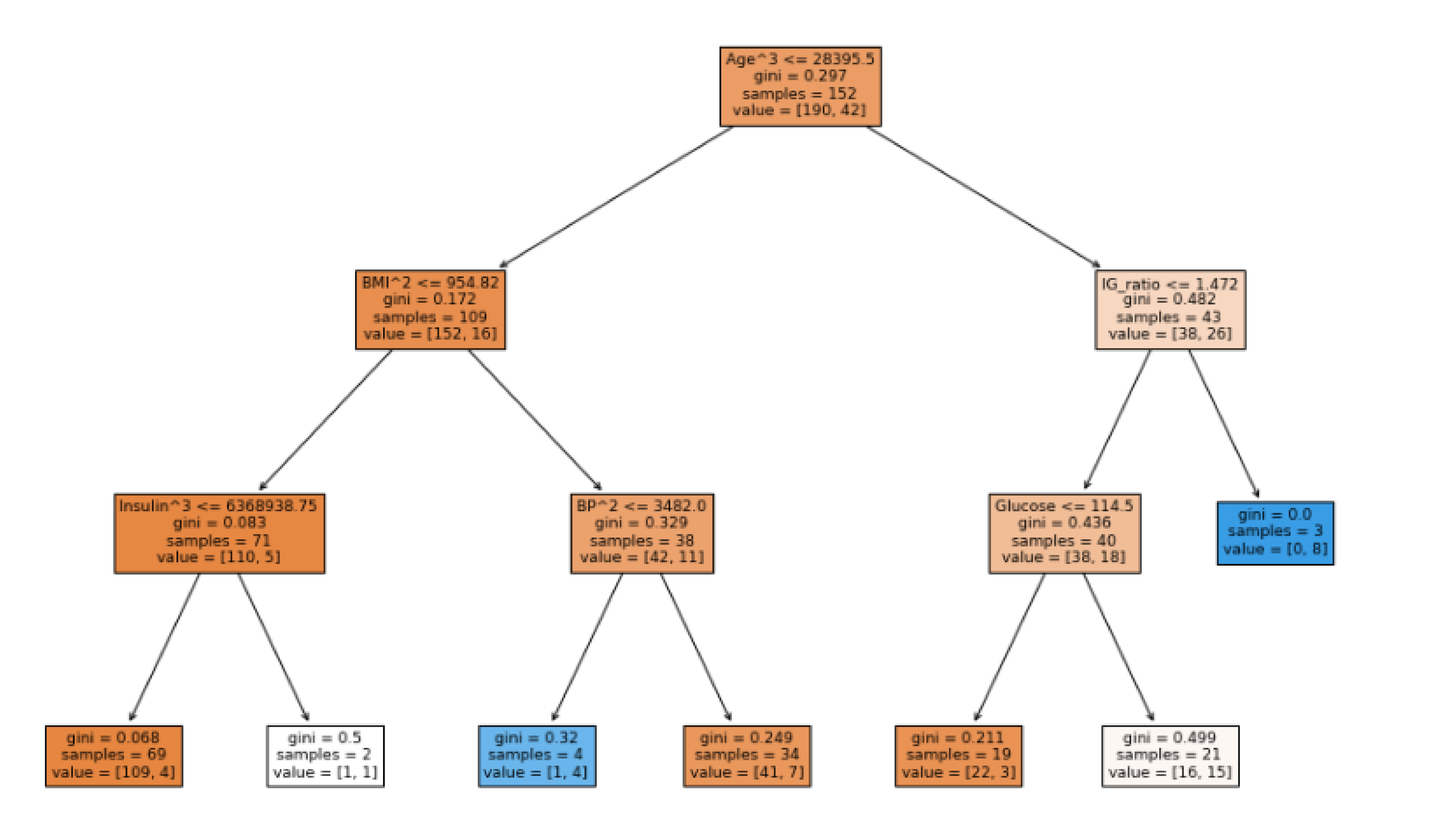
Method	n_estimators	Min Sample Split	Min Sample Leaf	Max Features	Max Depth	Accuracy
Random Search	20	3	1	sqrt	30	84.48%
Grid Search	90	12	1	sqrt	4	82.75%

HYPER PARAMETER TUNING

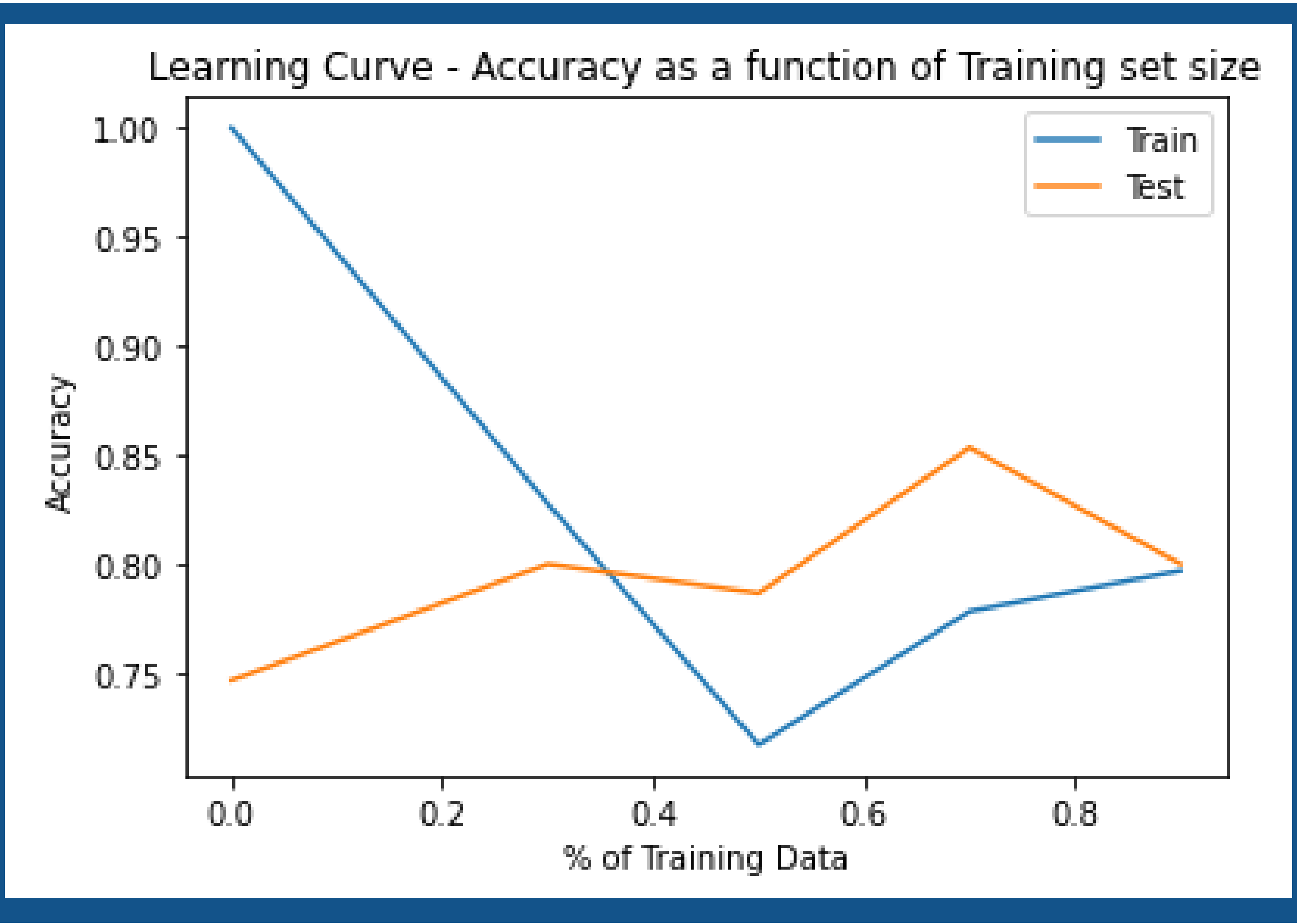
Method	n_estimators	Min Sample Split	Min Sample Leaf	Max Features	Max Depth	Accuracy
Random Search	20	3	1	sqrt	30	84.48%
Grid Search	90	12	1	sqrt	4	82.75%

RANDOM SEARCH

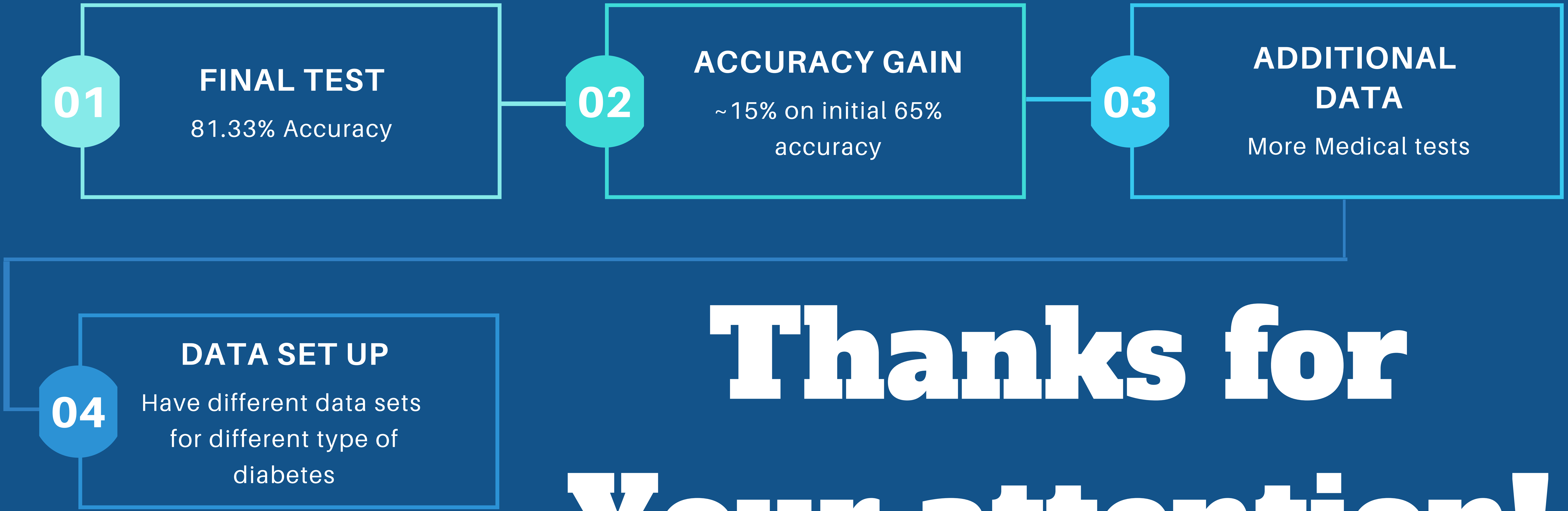
MODEL INTROSPECTION



LEARNING CURVE



FINAL THOUGHTS



01 FINAL TEST

81.33% Accuracy

02 ACCURACY GAIN

~15% on initial 65% accuracy

03 ADDITIONAL DATA

More Medical tests

04 DATA SET UP

Have different data sets for different type of diabetes

Thanks for Your attention!