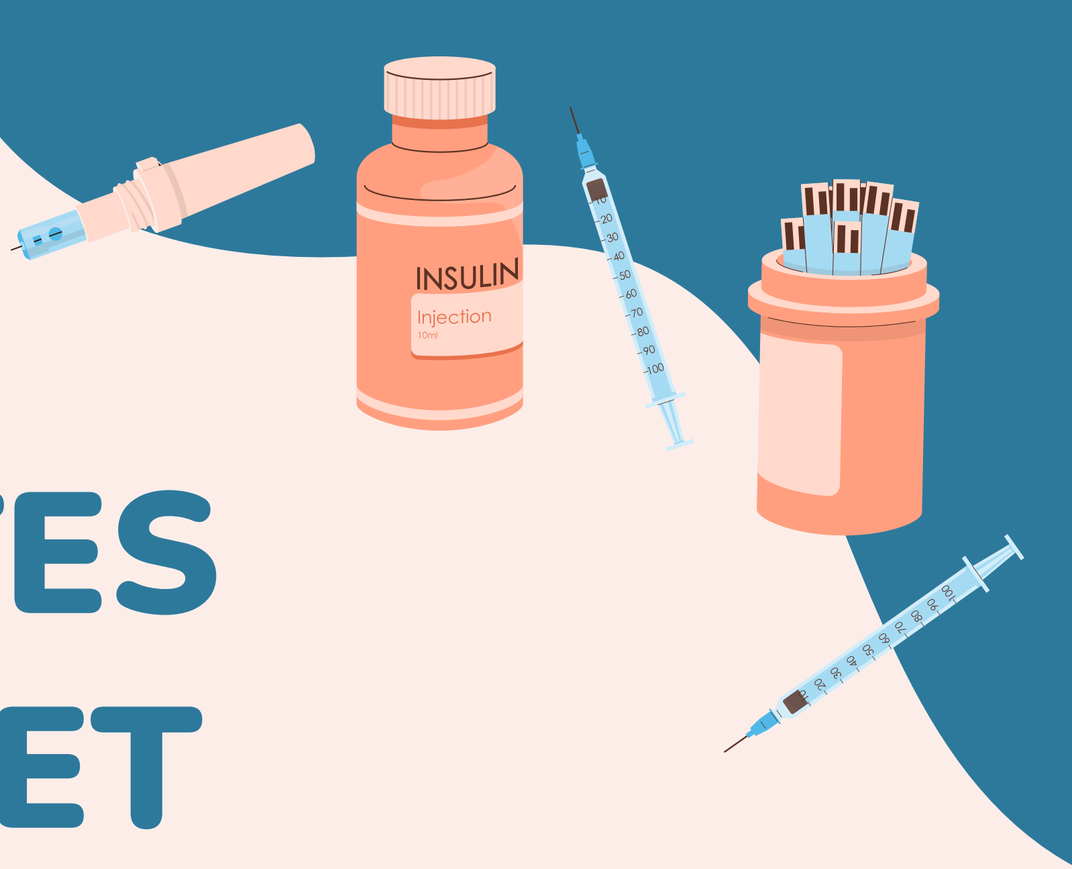
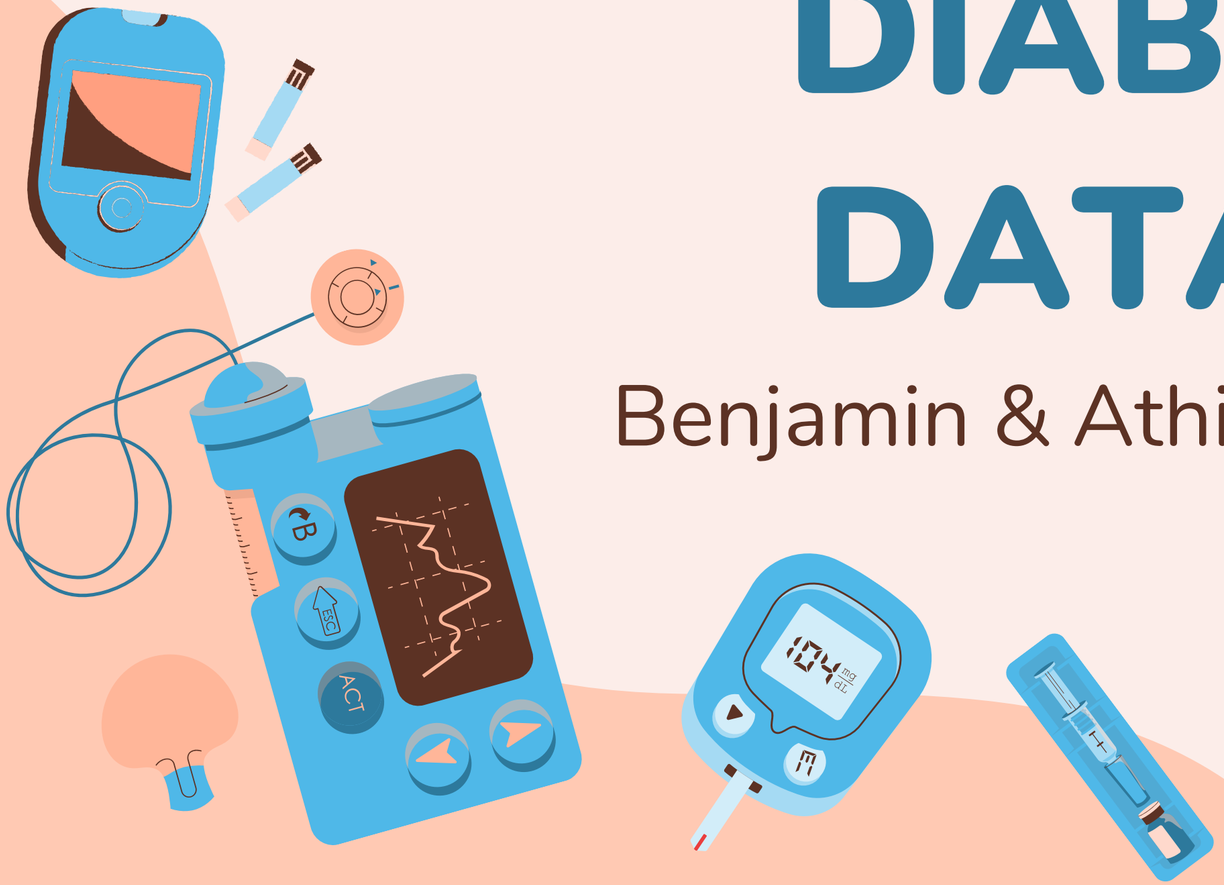
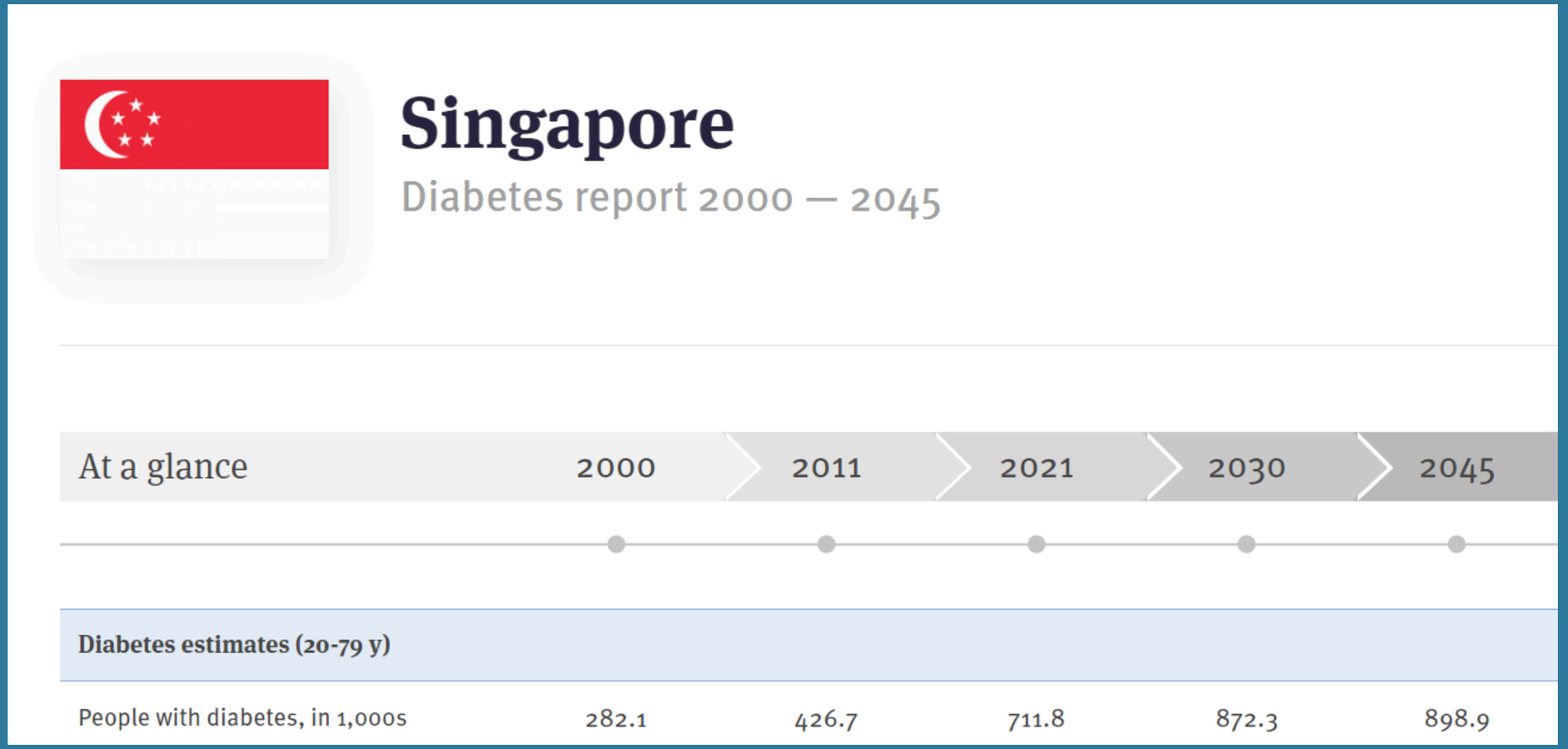


DIABETES DATASET

Benjamin & Athithiya (Team 11)



MOTIVATION



Singapore

Diabetes report 2000 — 2045

At a glance	2000	2011	2021	2030	2045
Diabetes estimates (20-79 y)					
People with diabetes, in 1,000s	282.1	426.7	711.8	872.3	898.9

- The figure shows an increasing trend (in ***thousands***) about the number of people getting or predicted having diabetes

TABLE OF CONTENTS

- Our Dataset and Problem Statement.
- EDA
- Models Implemented
- Analysis of findings and models
- Conclusion

PROBLEM STATMENT

Which is more important in predicting diabetes, family history or current health conditions?

OUR DATASET

- Pregnancies - No. of weeks of pregnancy during which a pregnant women can get diabetic (Numerical)
- Glucose - Amount of glucose content present (Numerical)
- Blood Pressure - Range of blood pressure in a person (Numerical)
- Skin Thickness - Thickness of skin(mm) in a person (Numerical)
- Insulin - Amount of Insulin injected into a person (Numerical)
- Body Mass Index - BMI of a person (Numerical)
- Diabetes Pedigree Function - Function range predicting if a person is diabetic (Numerical)
- Age - Age of a person (Numerical)
- Outcome - Predicting if one is diabetic ; 1 (Yes) 0 (No) (Categorical)

WHAT IS WRONG?

Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0

- SOME VALUES IN THE COLUMNS WERE RECORDED TO BE ZERO
- THE VALUE 0 IN SOME COLUMNS ARE NOT POSSIBLE IN THE HUMAN BODY

WHAT DID WE DO?

	Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	Age	Outcom
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1

Data mixed between diabetic and non-diabetic

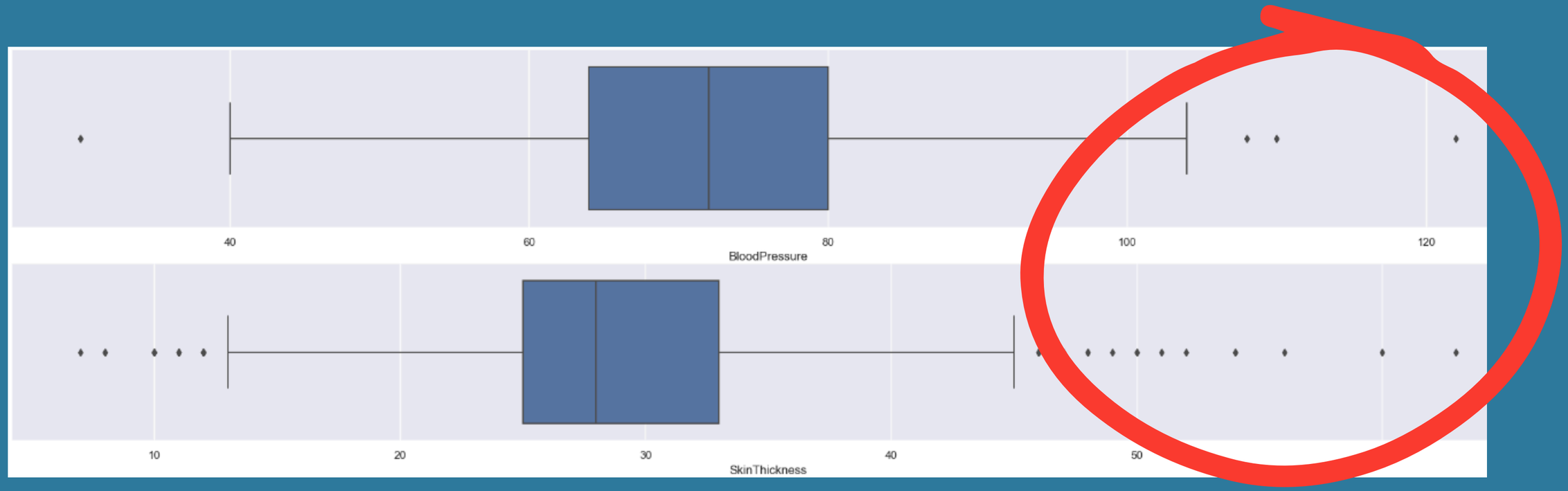
Cleaning the dataset (Exploratory Data Analysis) (Separating them into different datasets (with) and (without) diabetes)

```
diabetes_with = diabetes[diabetes["Outcome"] == 1].copy()
diabetes_without = diabetes[diabetes["Outcome"] == 0].copy()
```

Separating mixed data

- SEPARATED THE MIXED DATA INTO THEIR RESPECTIVE DATAFRAME (WITH & WITHOUT DIABETES)

WHAT IS WRONG?



- FILTERING OUT OUTLIERS SO THAT THE MEAN USED TO FILL THE NULL VALUES ARE NOT BIASED

```python
# Function for outliers
def outliers(df):
    q1 = df.quantile(0.25)
    q3 = df.quantile(0.75)
    iqr = q3 - q1
    upper_limit = q3 + 1.5 * iqr
    lower_limit = q1 - 1.5 * iqr
    return (lower_limit, upper_limit)
```

WHAT DID WE DO?

```
In [142]: mean_glucose_filtered = glucose_filtered["Glucose"].mean()
          mean_blood_pressure_filtered = blood_pressure_filtered["BloodPressure"].mean()
          mean_skinthickness_filtered = skinthickness_filtered["SkinThickness"].mean()
          mean_bmi_filtered = bmi_filtered["BMI"].mean()

          (
          mean_glucose_filtered,
          mean_blood_pressure_filtered,
          mean_skinthickness_filtered,
          mean_bmi_filtered,
          )

Out[142]: (122.33535353535353, 71.97649572649573, 28.73654390934844, 32.1654958677686)

In [143]: diabetes_filtered["Glucose"].fillna(mean_glucose_filtered.mean(), inplace = True)
          diabetes_filtered["BloodPressure"].fillna(mean_blood_pressure_filtered.mean(), inplace = True)
          diabetes_filtered["SkinThickness"].fillna(mean_skinthickness_filtered.mean(), inplace = True)
          diabetes_filtered["BMI"].fillna(mean_bmi_filtered.mean(), inplace = True)

In [144]: (
          diabetes_filtered["Glucose"].isnull().sum(),
          diabetes_filtered["BloodPressure"].isnull().sum(),
          diabetes_filtered["SkinThickness"].isnull().sum(),
          diabetes_filtered["BMI"].isnull().sum()
          )

Out[144]: (0, 0, 0, 0)
```

using the .mean() function

Making sure all NULL values are filled

- **REPLACED THE ZERO VALUES WITH THE MEAN OF THE NON-ZERO ENTRIES OF THE COLUMNS**

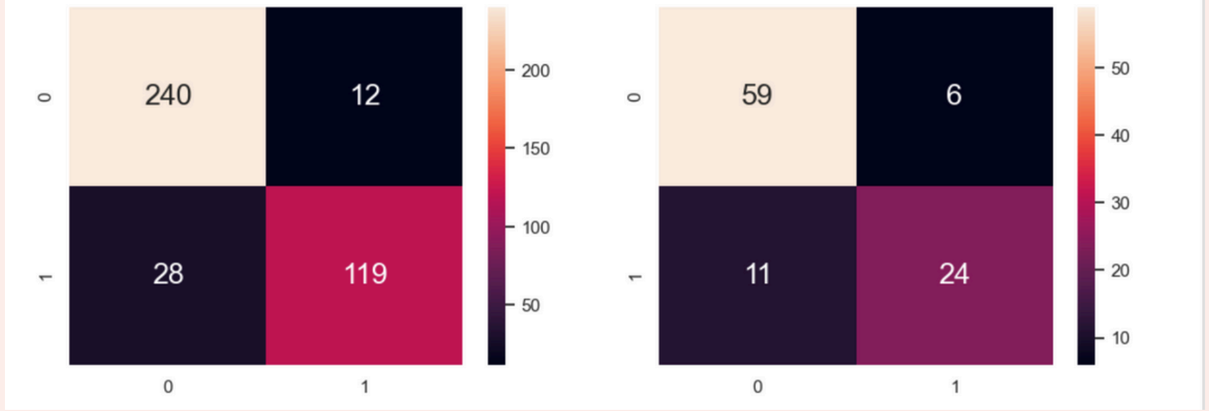
Classification Decision Tree

Without Pedigree function

```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.899749373433584

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.83

: <Axes: >
```

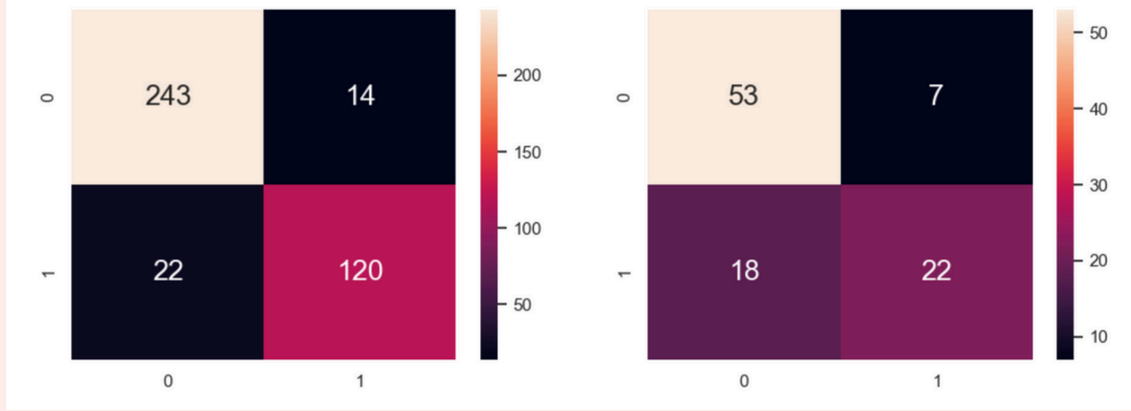


With Pedigree function

```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.9097744360902256

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.75

: <Axes: >
```



Classification Accuracy for Test Data (Without) : 0.83

Classification Accuracy for Test Data (With) : 0.75

Accuracy for Test Data (Without) >

Accuracy for Test Data (With)

Shifting Our Focus to FNR

Without Pedigree function

The TPR Train is :	0.8095238095238095
The TNR Train is :	0.9523809523809523
The FPR Train is :	0.04761904761904761 6
The FNR Train is :	0.19047619047619047
The TPR Test is :	0.6857142857142857
The TNR Test is :	0.9076923076923077
The FPR Test is :	0.09230769230769231
The FNR Test is :	0.3142857142857143

With Pedigree function

The TPR Train is :	0.8450704225352113
The TNR Train is :	0.9455252918287937
The FPR Train is :	0.054474708171206226
The FNR Train is :	0.15492957746478872
The TPR Test is :	0.55
The TNR Test is :	0.8833333333333333
The FPR Test is :	0.11666666666666667
The FNR Test is :	0.45

Findings

FNR Test (Without) : 0.31

FNR Test (With) : 0.45

FNR Test Difference : 0.14 (With Pedigree Higher)

WHY FNR?

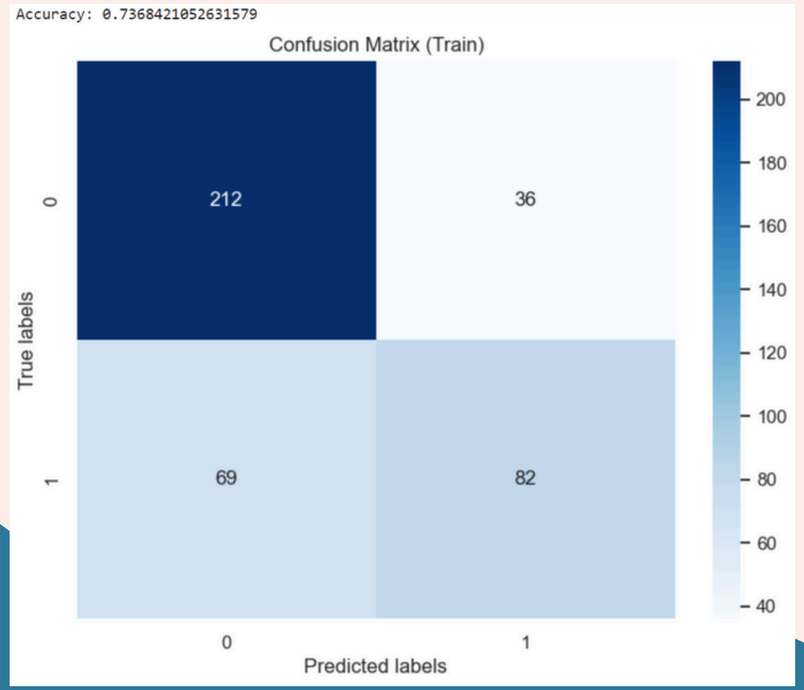
Better to be err on the safe side

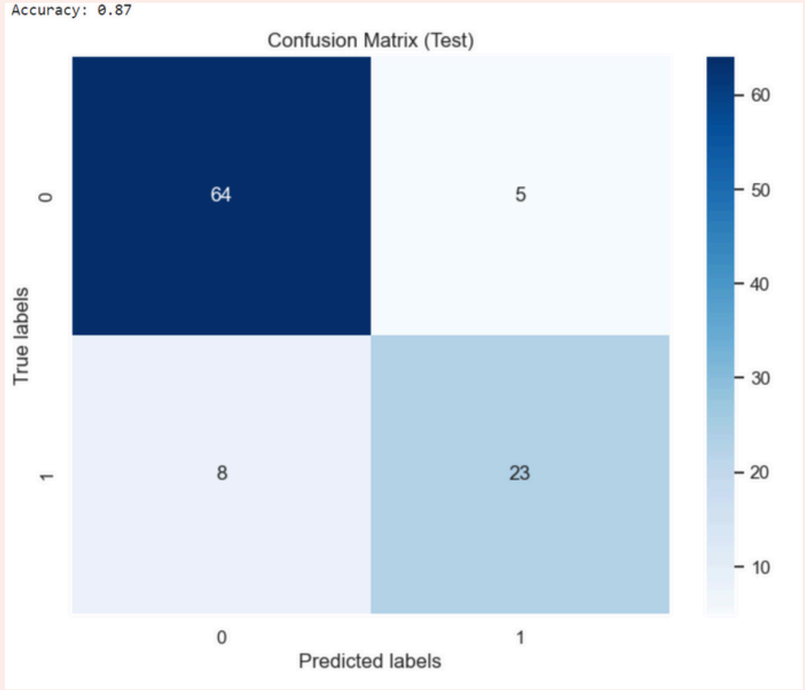
Logistic Regression

What is Logistic Regression?

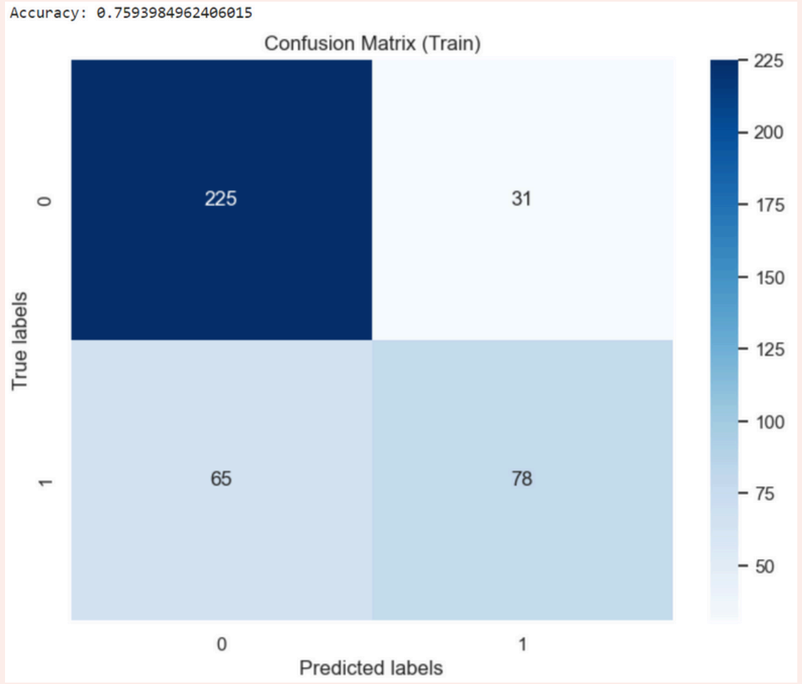
- A Model used to predict the probability of binary outcomes using 1 or more predictor variable
- Uses a logistic function to transform the output of a linear equation into a probability between 0 and 1

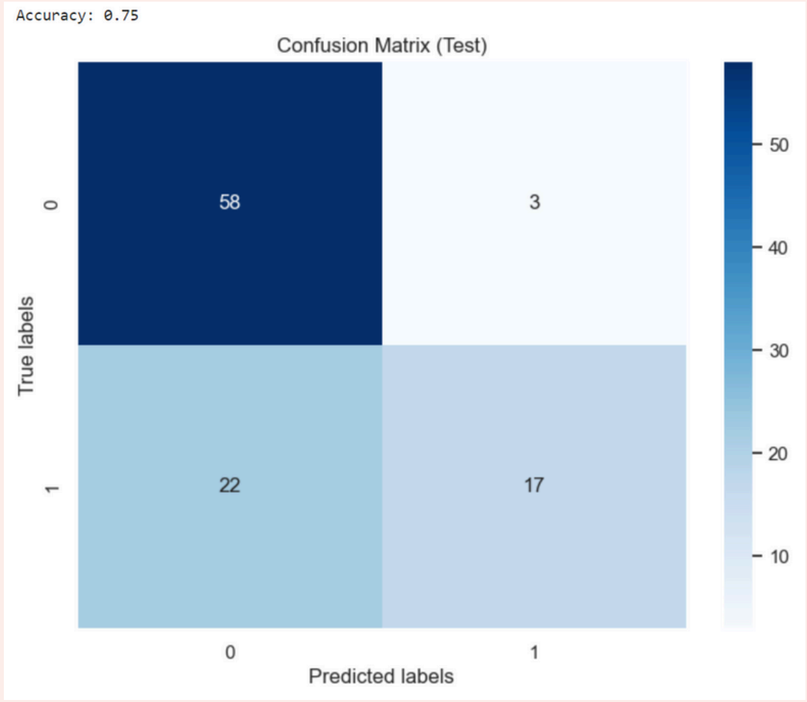
Without Pedigree function





Logistic Regression
With Pedigree function





Classification Accuracy for Test Data (Without) : 0.87

Classification Accuracy for Test Data (With) : 0.75

The model with the Diabetes Pedigree Function has a much lower classification accuracy as compared to the model that does not include it

Shifting Our Focus to FNR

Without Pedigree function

The TPR Train is :	0.543046357615894
The TNR Train is :	0.8548387096774194
The FPR Train is :	0.14516129032258066
The FNR Train is :	0.45695364238410596
The TPR Test is :	0.7419354838709677
The TNR Test is :	0.927536231884058
The FPR Test is :	0.07246376811594203
The FNR Test is :	0.25806451612903225

With Pedigree function

The TPR Train is :	0.5454545454545454
The TNR Train is :	0.87890625
The FPR Train is :	0.12109375
The FNR Train is :	0.4545454545454453
The TPR Test is :	0.4358974358974359
The TNR Test is :	0.9508196721311475
The FPR Test is :	0.04918032786885246
The FNR Test is :	0.5641025641025641

Findings

FNR Test (Without) : 0.25

FNR Test (With) : 0.56

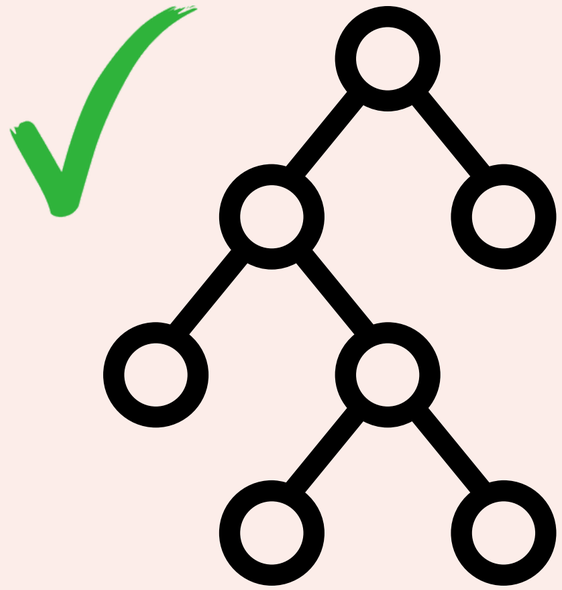
FNR Test Difference : 0.31 (With Pedigree Higher)

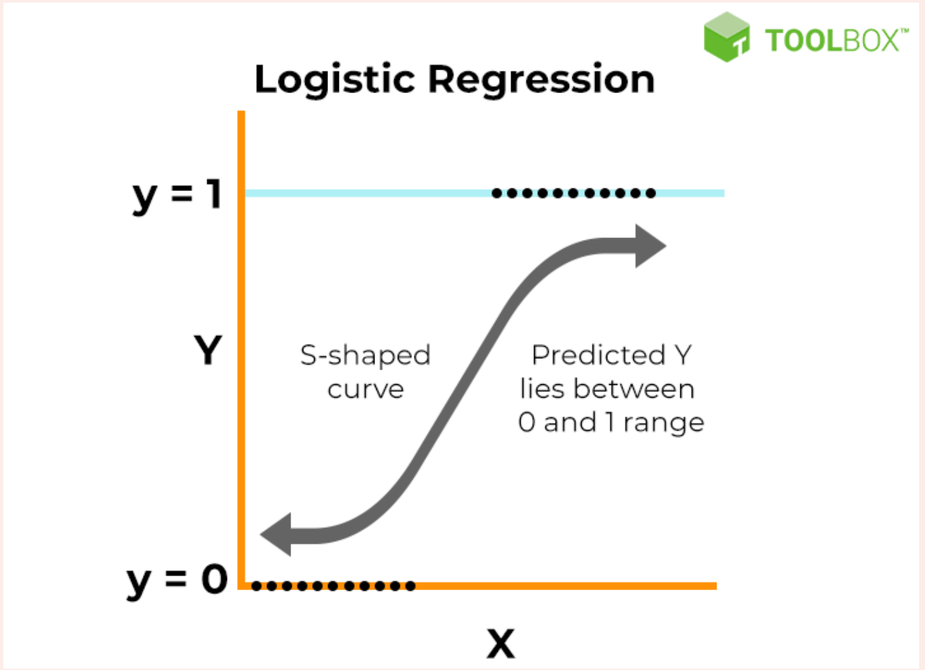
The model with the Pedigree Function returns a much higher False Negative Rate

Analysis of Models

Classification Decision Tree vs Logistic Regression

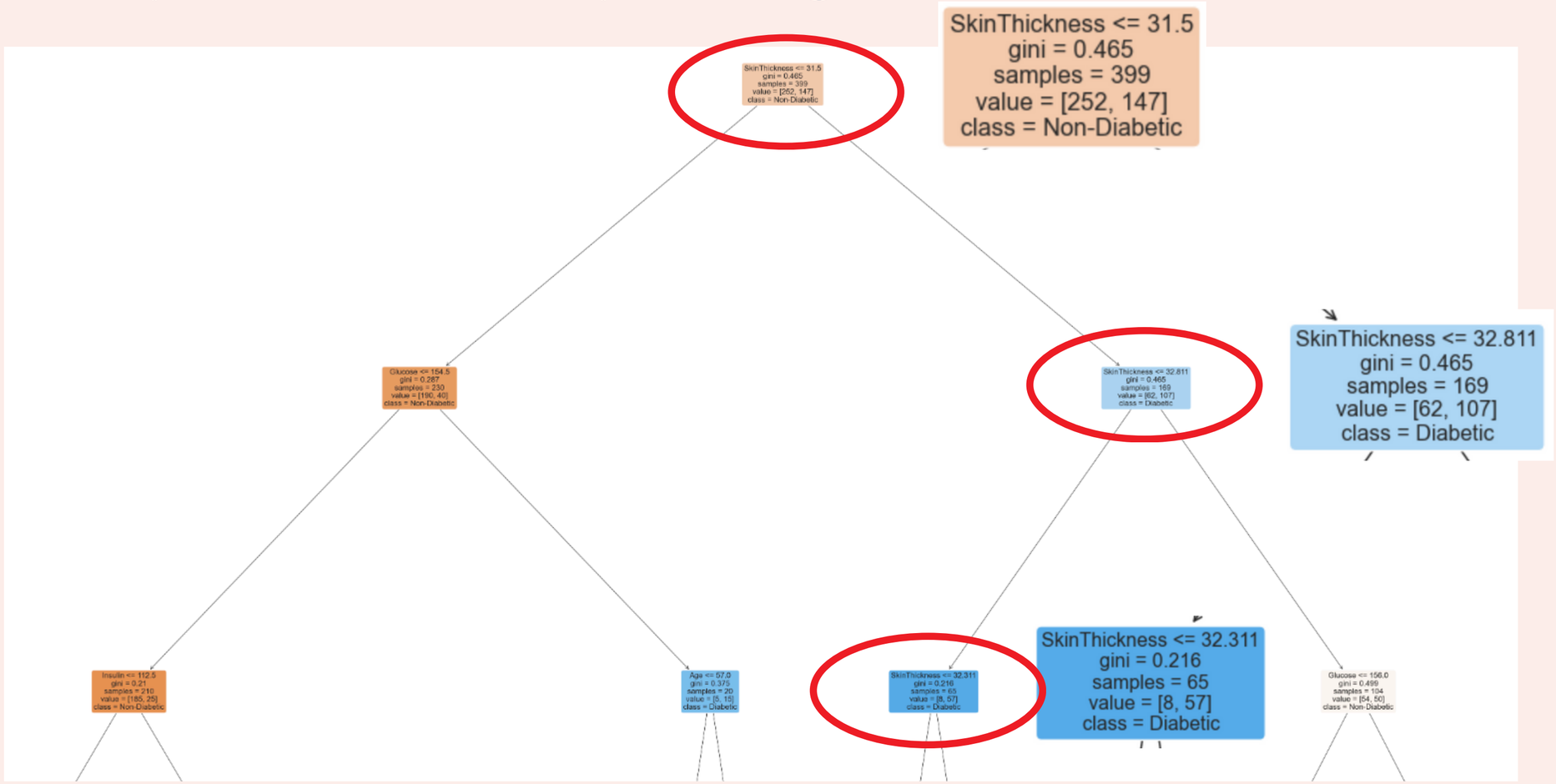
- The depth of the tree can be modified till pure leaf nodes are formed to ensure a more accurate analysis
- Decision tree can capture non-linear relationships in the data whereas logistic regression assumes a linear relationship between predictor and log-odds





Which factor holds the most influence?

- Skin Thickness is the root node in the tree and also appears frequently in the first 3 levels, indicating that it is the most important variable in predicting the outcome of diabetes.



Conclusion

Current Health vs Family History

- Both logistic regression and classification decision tree show lower classification accuracy and higher FNR rates with Diabetes Pedigree Function
- **Current Health** is the most important factor when it comes to predicting Diabetes