# Diabetes Risk Indicator

**(Youths)**

Should you be worried?

**SC1015 Mini-Project** by

Belvedere

Malcolm

Raphael

# Why does it concern you?


https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

https://www.straitstimes.com/life/d-diabetes-in-teens

## TODAY..

"Approximately **half a billion** adults (20-79 years) are living with diabetes and is projected to rise to 783 million by 2045"

"Diabetes have caused **6.7 million** deaths"

"Type 2 diabetes - usually more common in people who are older than 40 - has been **increasingly diagnosed among teenagers"**

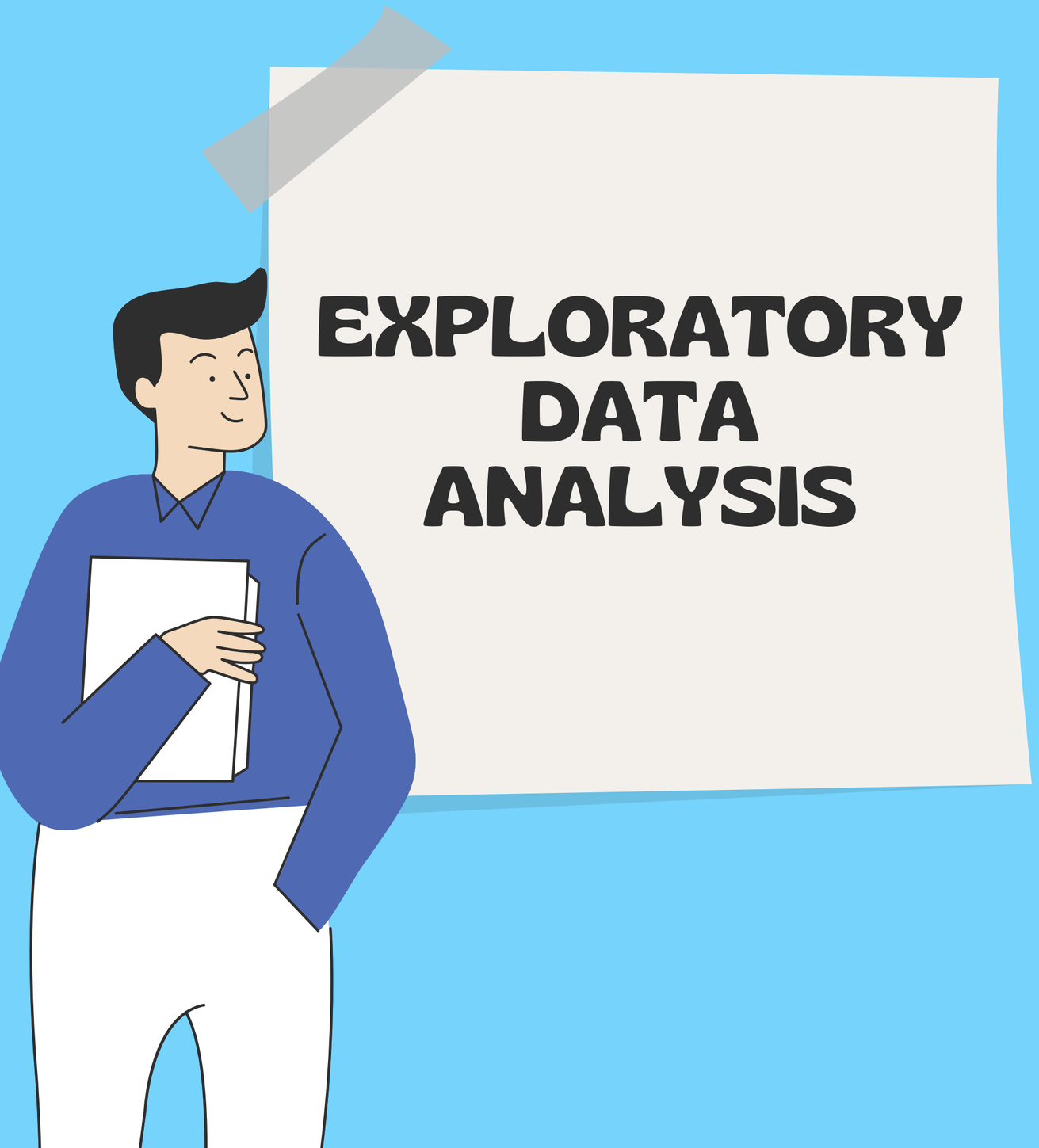
# What can we do?

- **Identifying risk factors** that are most predictive of diabetes risk can aid in early detection, prevention, and targeted interventions. For example, prediabetes detection.

- AIM: To analyze a dataset on diabetes health indicators to **uncover key risk factors that contribute to diabetes risk in Youth.**

# Problem Definition

"What risk factors are most predictive of diabetes or **which risk factors** can be used to accurately predict **whether an individual (age 18-24) has diabetes**?"



# EXPLORATORY DATA ANALYSIS

1. Data Cleaning & Preparation
2. Data Visualization

## "NO MISSING ENTRIES"

```
Number of Null entries in each column:
Diabetes_binary          0
HighBP                   0
HighChol                 0
CholCheck                0
BMI                      0
Smoker                   0
Stroke                   0
HeartDiseaseorAttack     0
PhysActivity             0
Fruits                   0
Veggies                  0
HvyAlcoholConsump        0
AnyHealthcare            0
NoDocbcCost              0
GenHlth                  0
MentHlth                 0
PhysHlth                 0
DiffWalk                 0
Sex                      0
Age                      0
Education                0
Income                   0
dtype: int64
```

## "ALL FLOAT (EVEN FOR CATEGORICAL)"

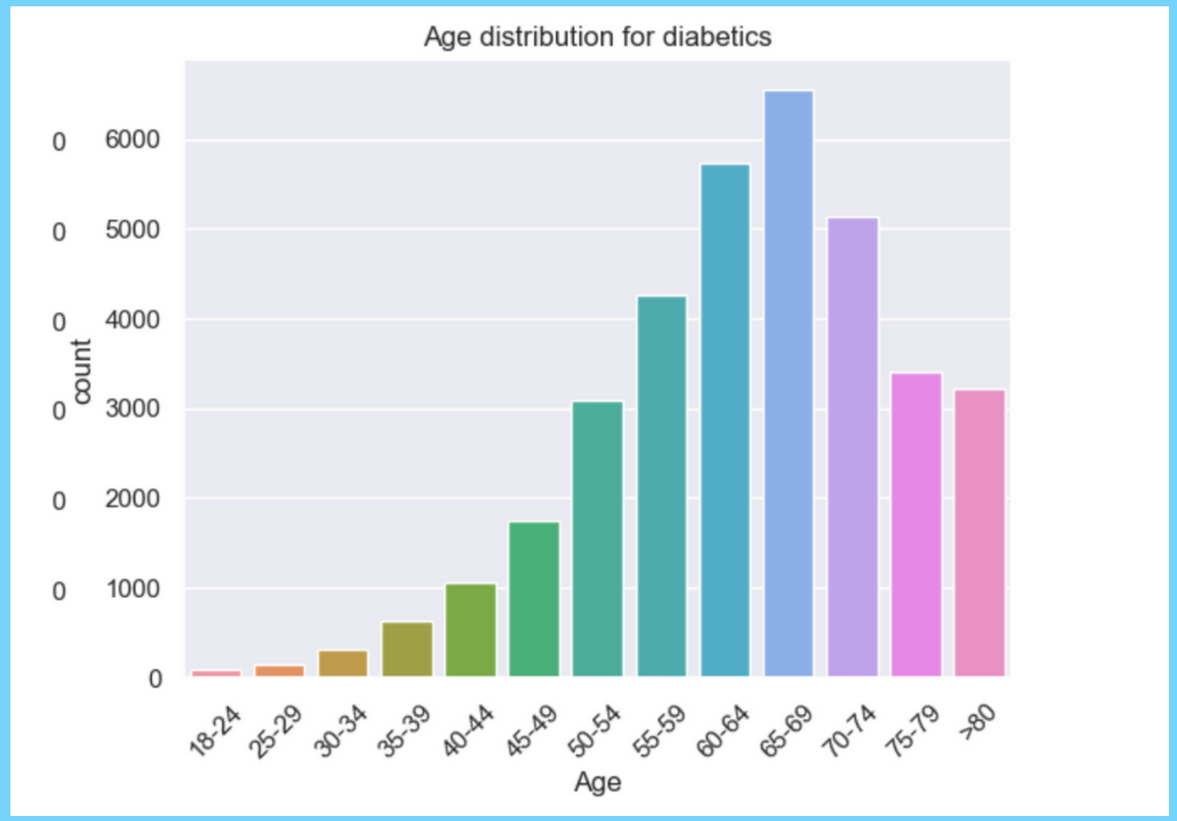
```
Date type for each column:
Diabetes_binary          float64
HighBP                   float64
HighChol                 float64
CholCheck                float64
BMI                      float64
Smoker                   float64
Stroke                   float64
HeartDiseaseorAttack     float64
PhysActivity             float64
Fruits                   float64
Veggies                  float64
HvyAlcoholConsump        float64
AnyHealthcare            float64
NoDocbcCost              float64
GenHlth                  float64
MentHlth                 float64
PhysHlth                 float64
DiffWalk                 float64
Sex                      float64
Age                      float64
Education                float64
Income                   float64
dtype: object
```

With simple satistical tool employed, we found our data to be clean and well-prepared for analysis.

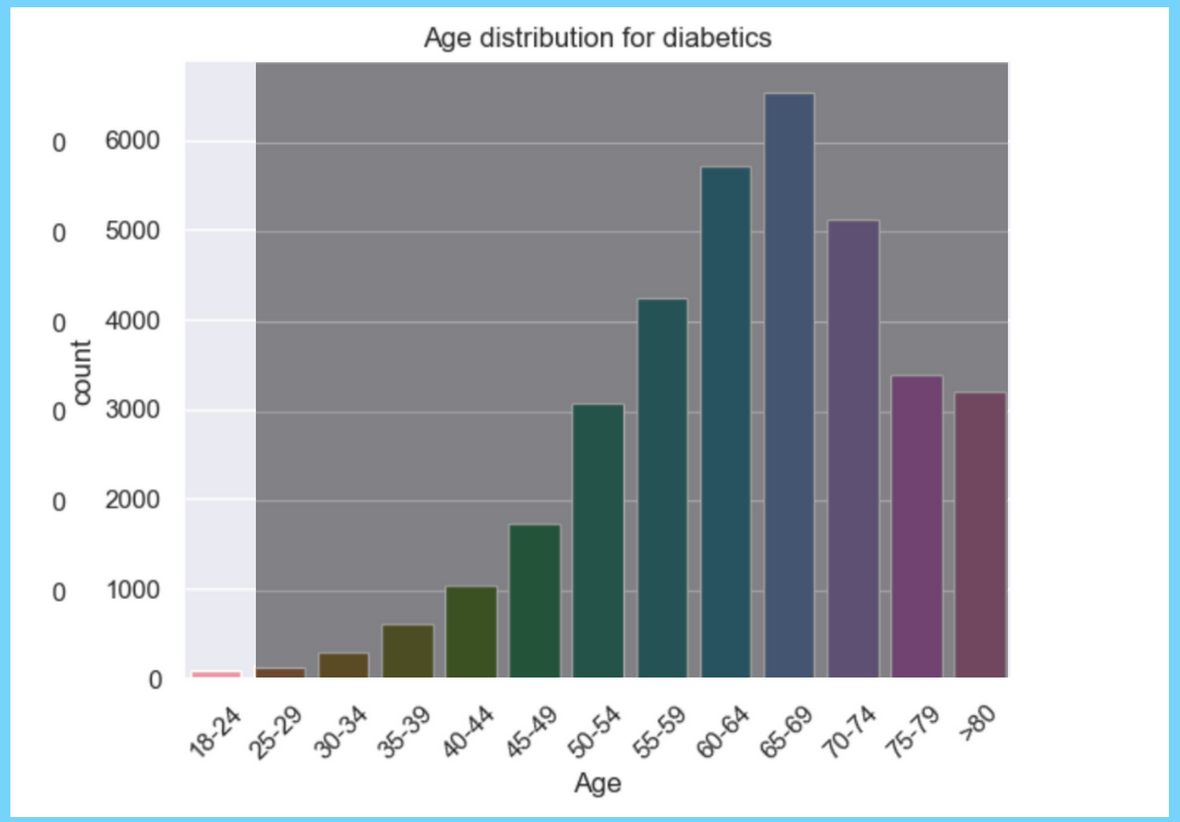
```
Number of Null entries in each column:
Diabetes_binary           0
HighBP                    0
HighChol                  0
CholCheck                 0
BMI                       0
Smoker                    0
Stroke                    0
HeartDiseaseorAttack      0
PhysActivity              0
Fruits                    0
Veggies                   0
HvyAlcoholConsump         0
AnyHealthcare             0
NoDocbcCost               0
GenHlth                   0
MentHlth                  0
PhysHlth                  0
DiffWalk                  0
Sex                       0
Age                       0
Education                 0
Income                    0
dtype: int64
```

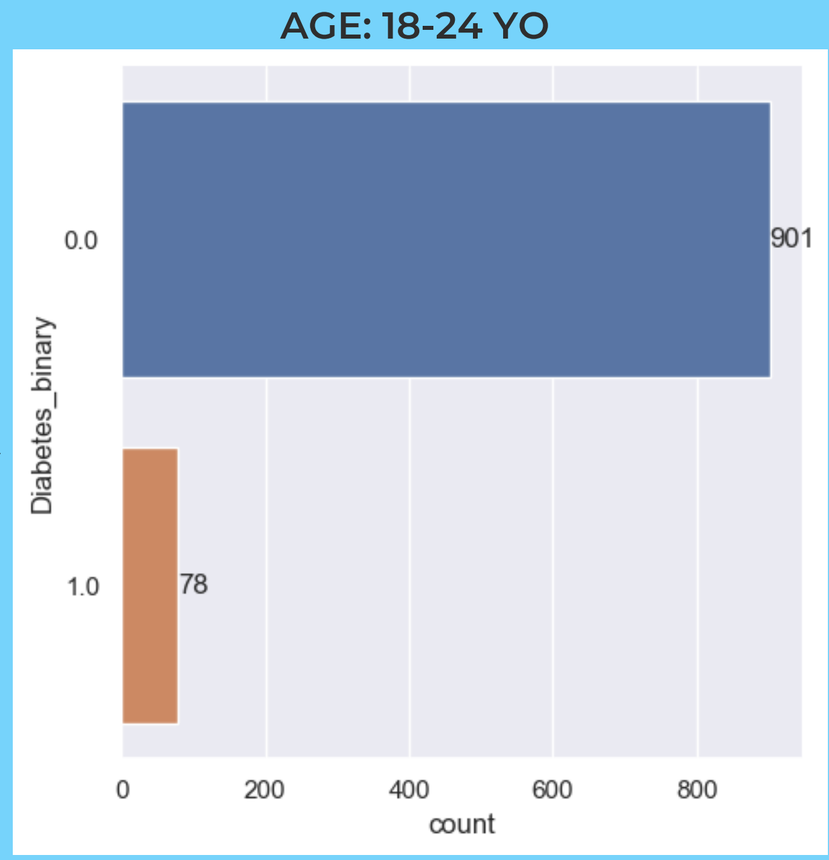
After the relevant preliminary data exploration, we have narrowed down our dataset to these 18 Columns

# Targeted Age Group

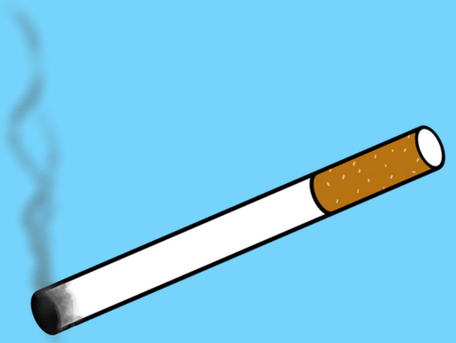


# Targeted Age Group

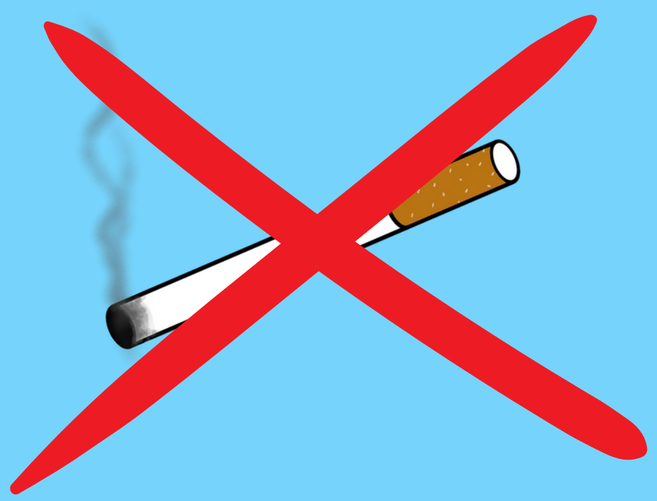


# UNHEALTHY NORMS AMONG YOUTHS

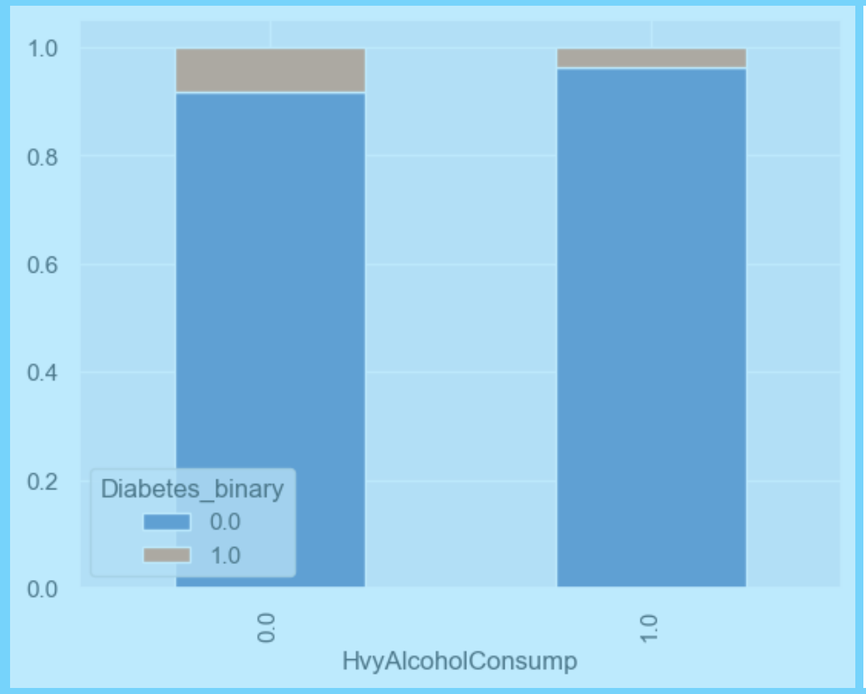


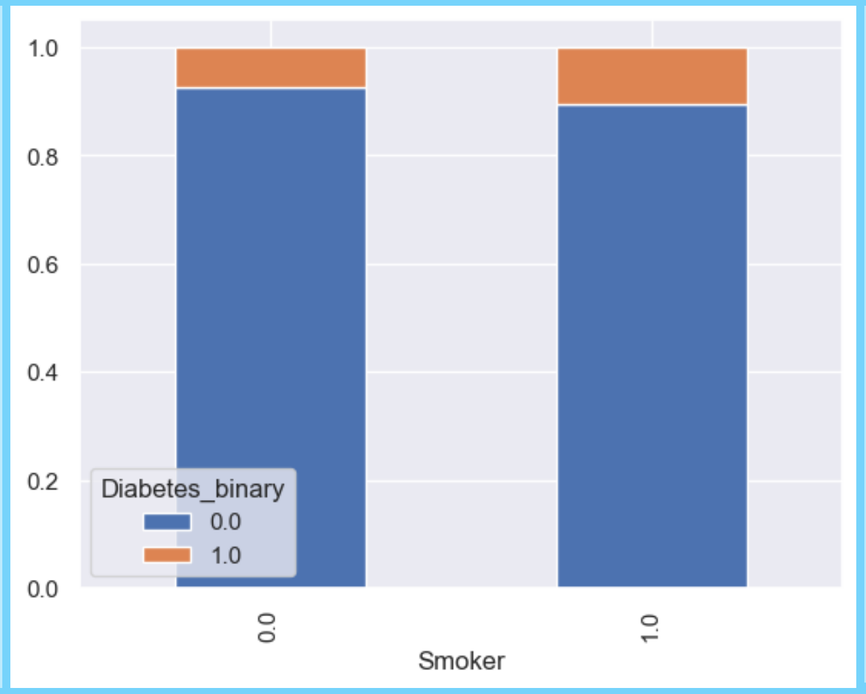
# UNHEALTHY NORMS AMONG YOUTHS

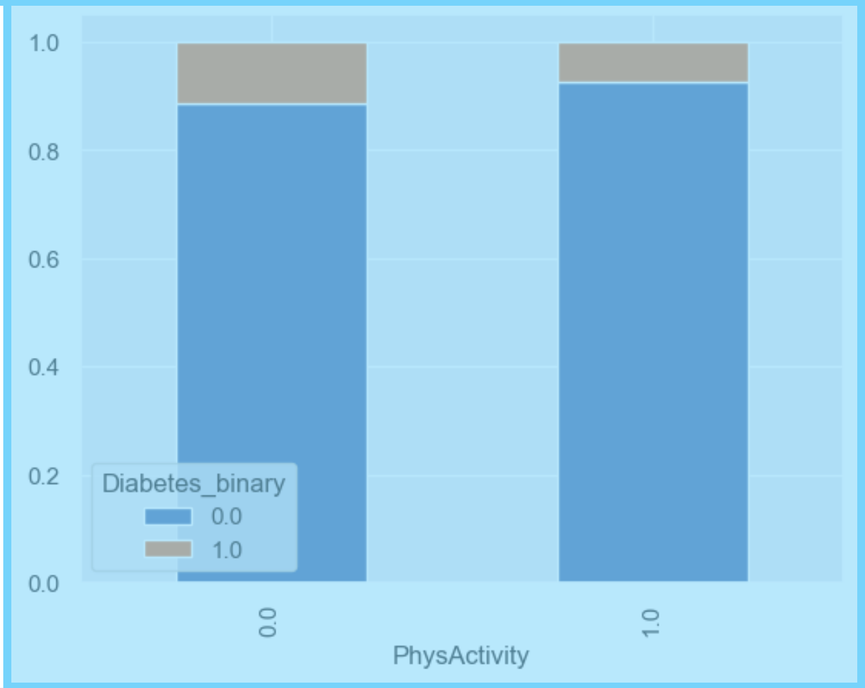


# Bi-variate Visualization

## Lifestyle habit indicator



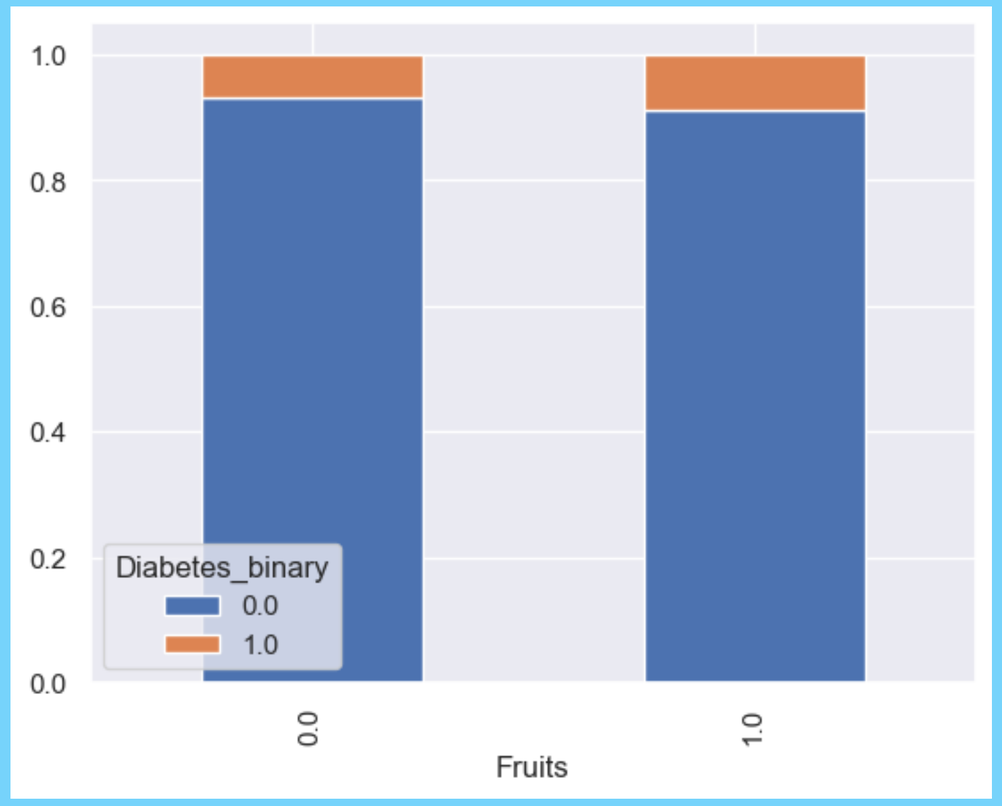



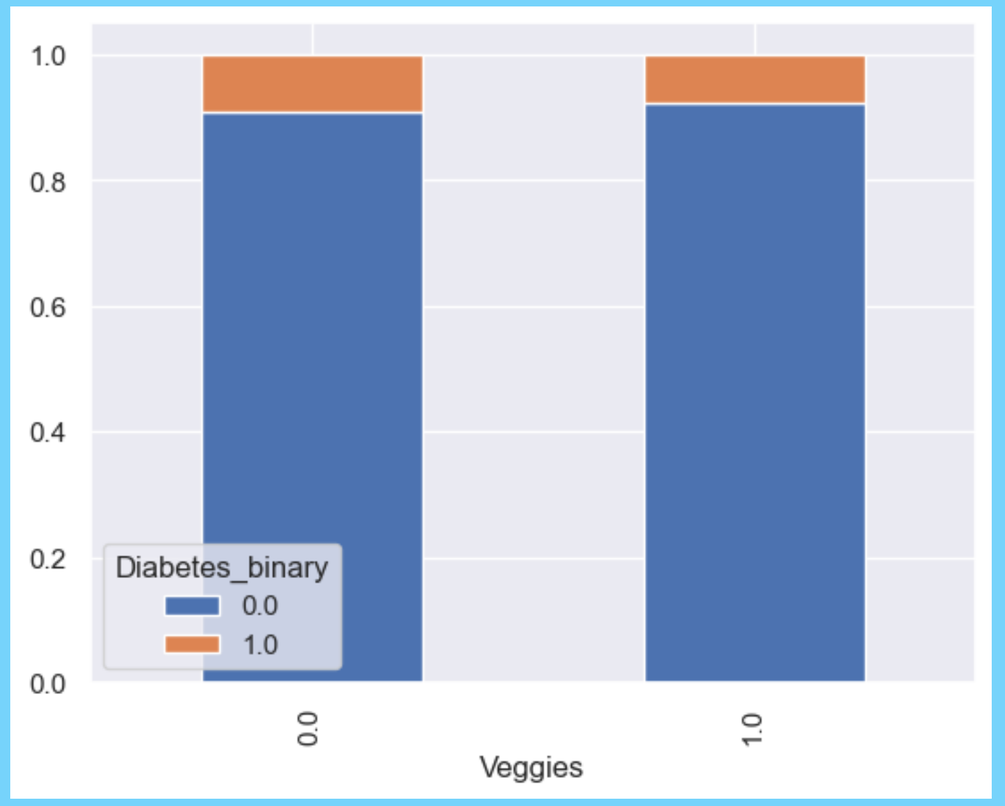




OBSERVATION: SIMILAR INCIDENCE OF DIABETES FOR SMOKERS,

# Bi-variate Visualization

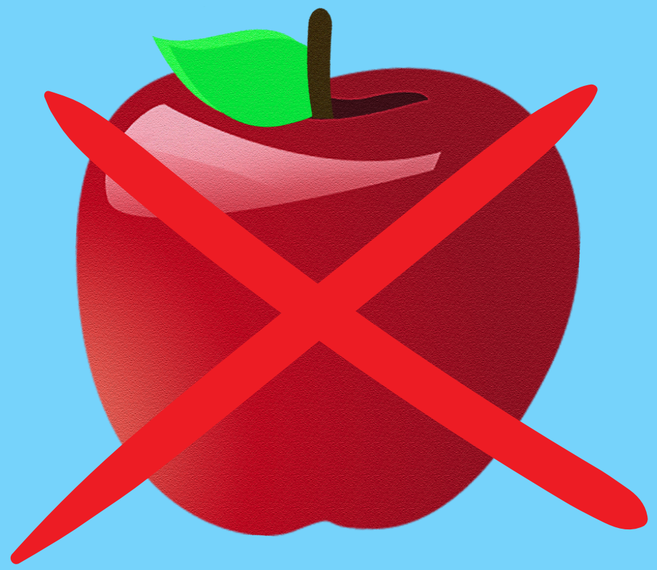
## Diet indicator



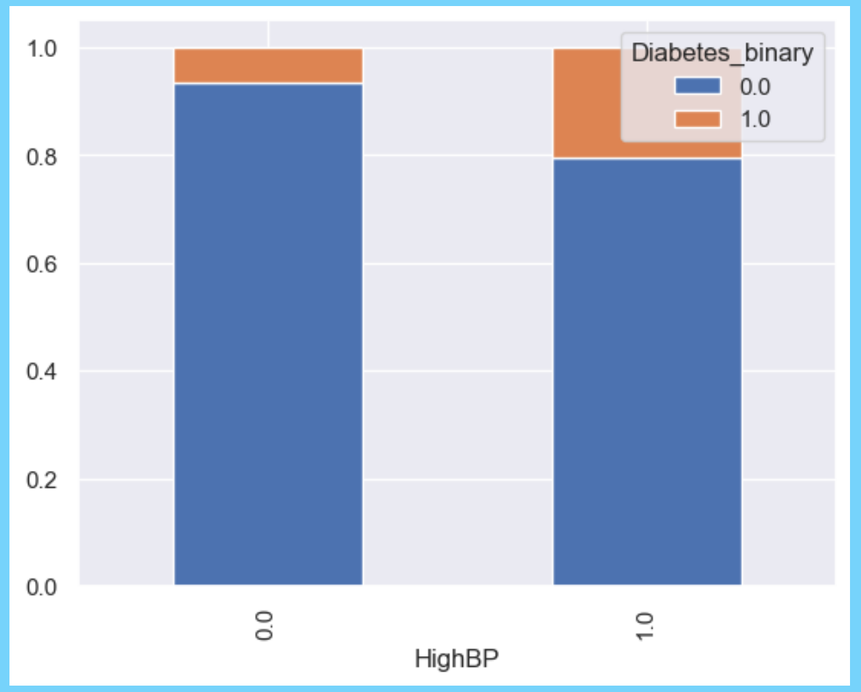

Observation: percentage of people with pre-diabetes are roughly the same for both Fruits & Non-Fruits eater and Veggies & Non-Veggies eater

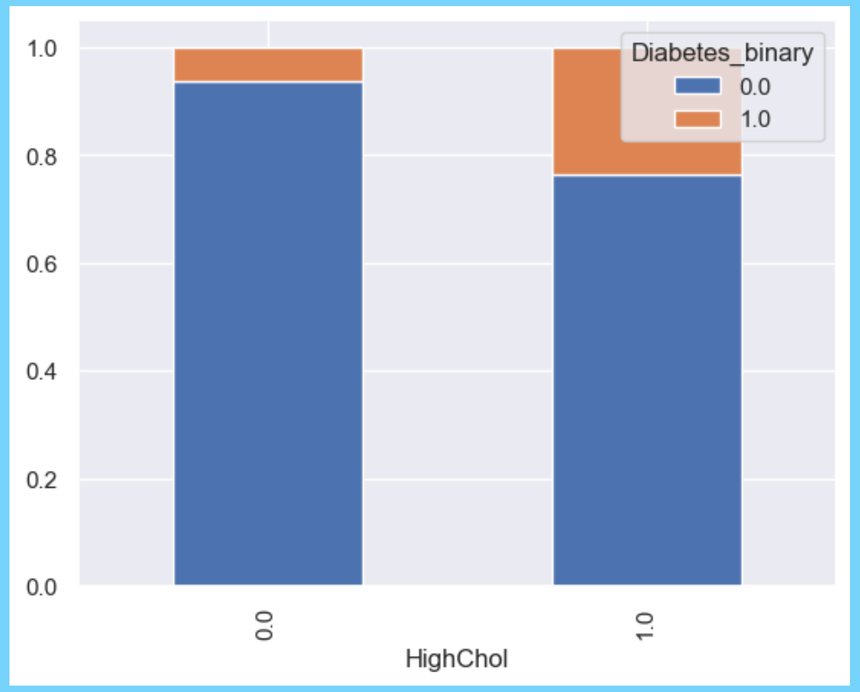
# DIET AMONG YOUTHS

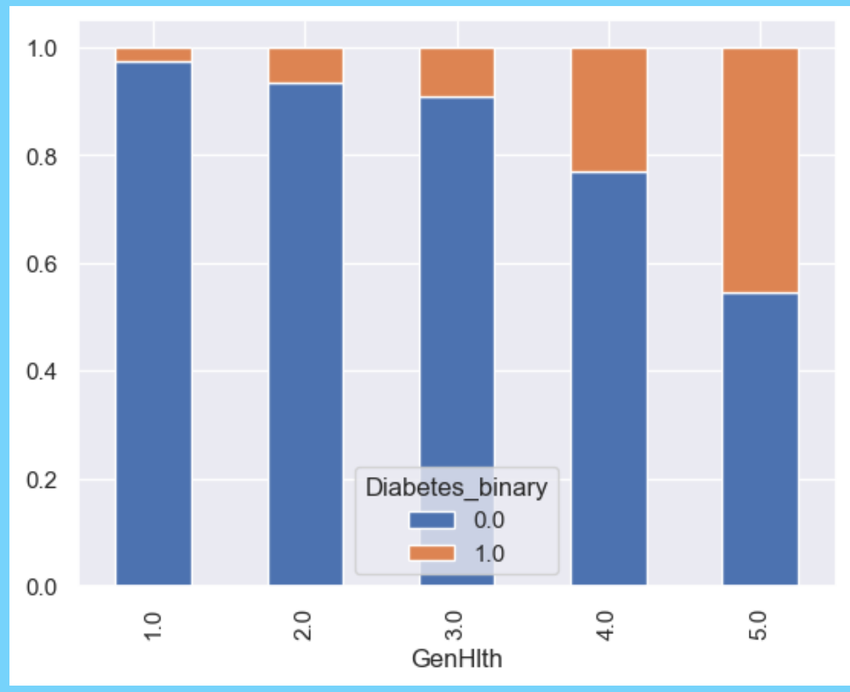


# Bi-variate Visualization

## Health Status Indicators

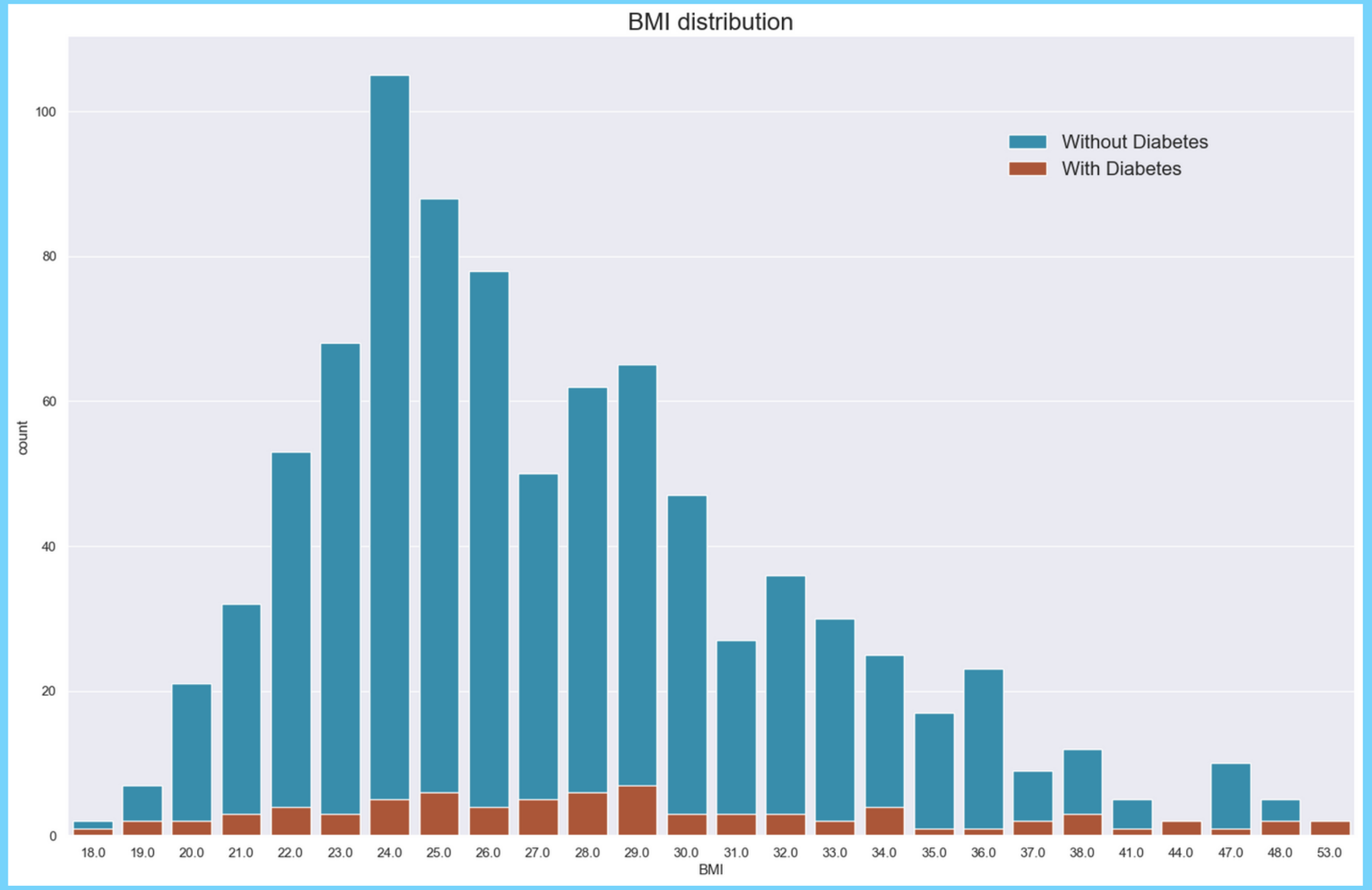



Observation: Higher incidence of prediabetes in people who experience HighBP, HighChol and Lower level of well-being

# Bi-variate Visualization




Observation: BMI has no impact on a person's diabetic status

# INSIGHT FROM EDA

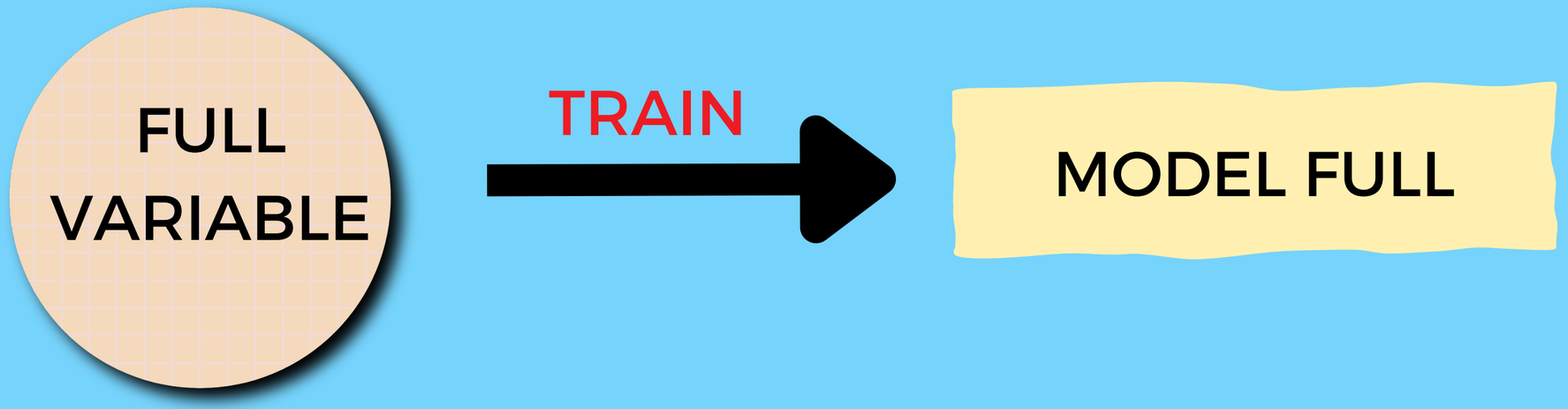
From Bi-Variate Exploration

- no variable has strong correlation with one's diabetic status

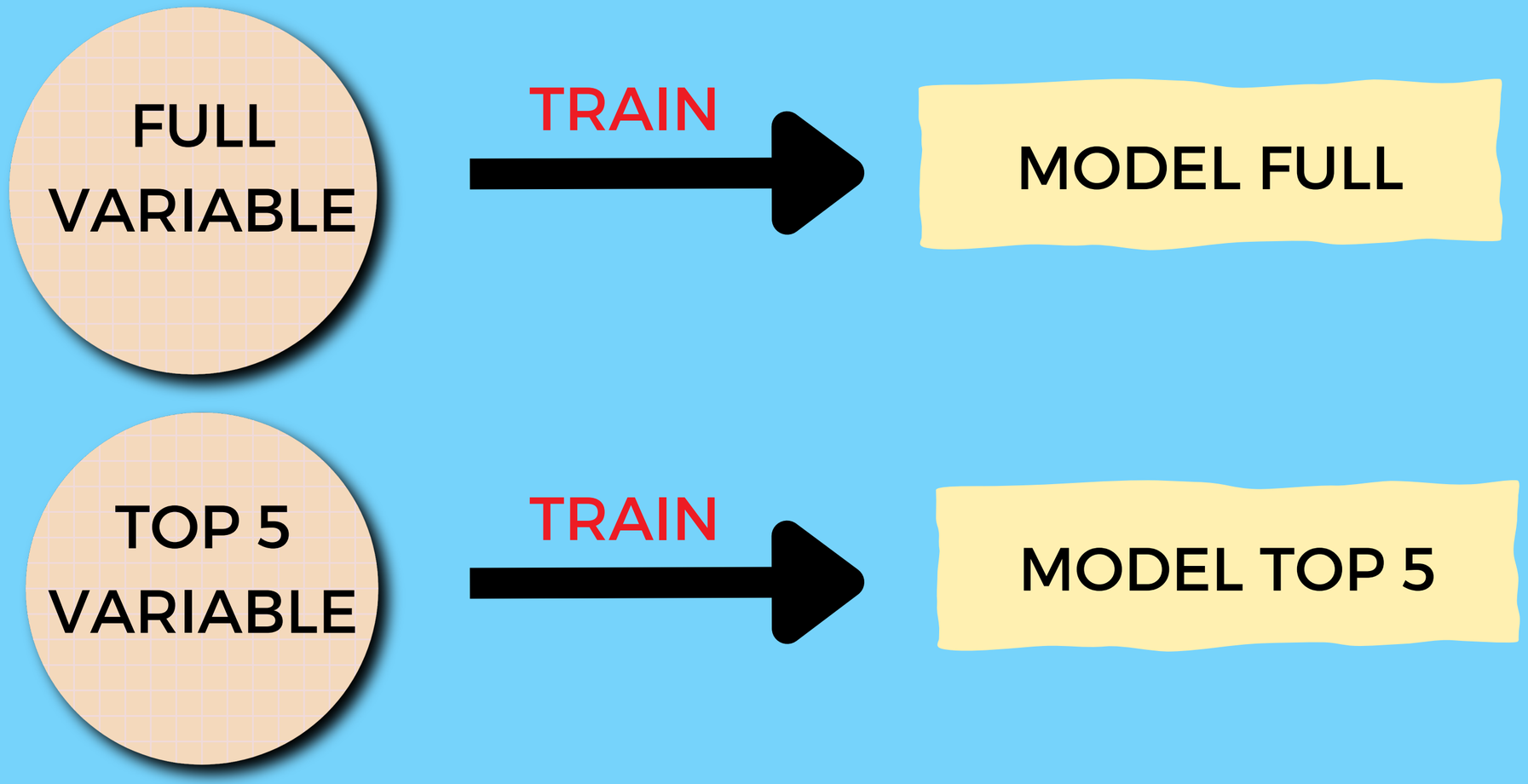
# MACHINE LEARNING MODEL

"What risk factors are most predictive of diabetes  or which risk factors can be used to accurately predict whether an individual (age 18-24) has diabetes?"

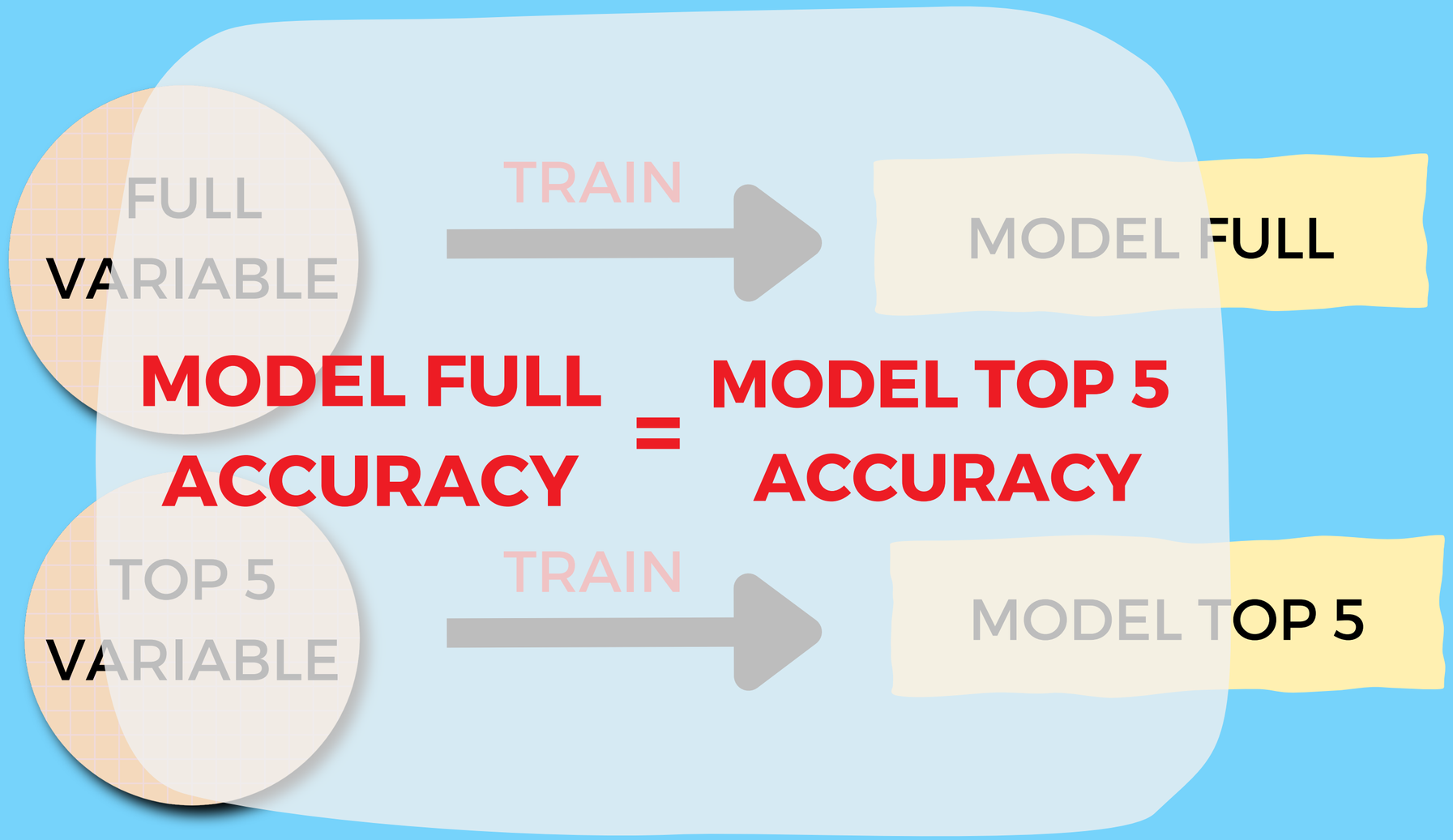
# MACHINE LEARNING MODEL



# MACHINE LEARNING MODEL



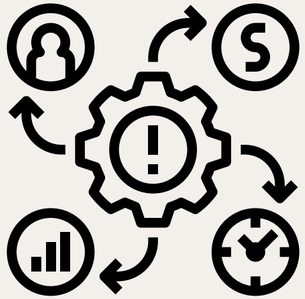
# MACHINE LEARNING MODEL



# Machine Learning

Machine learning methods

**Logistic Regression**

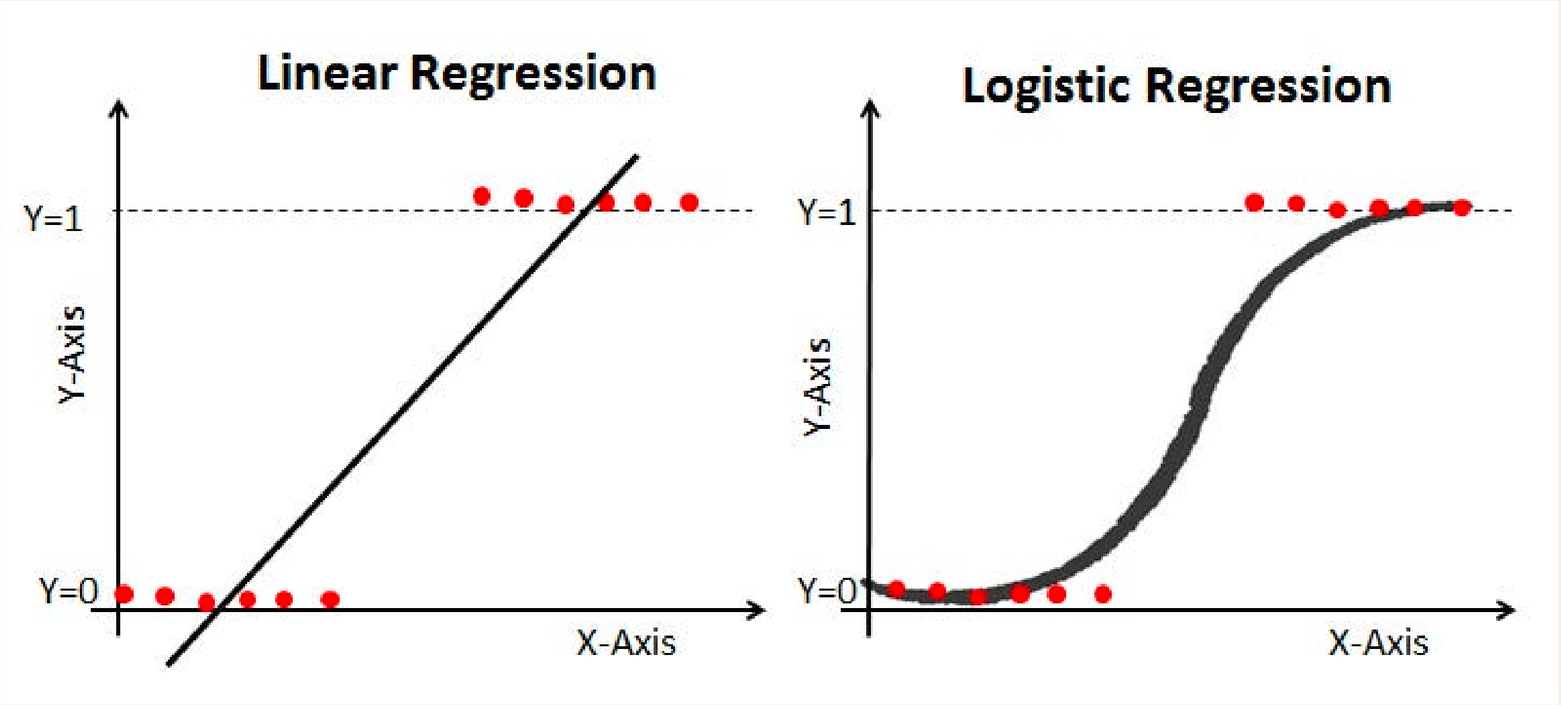


**One Class SVM**

**Isolation Forest**

# Logistic Regression

## Introduction




learning model used to find relationship between categorical data

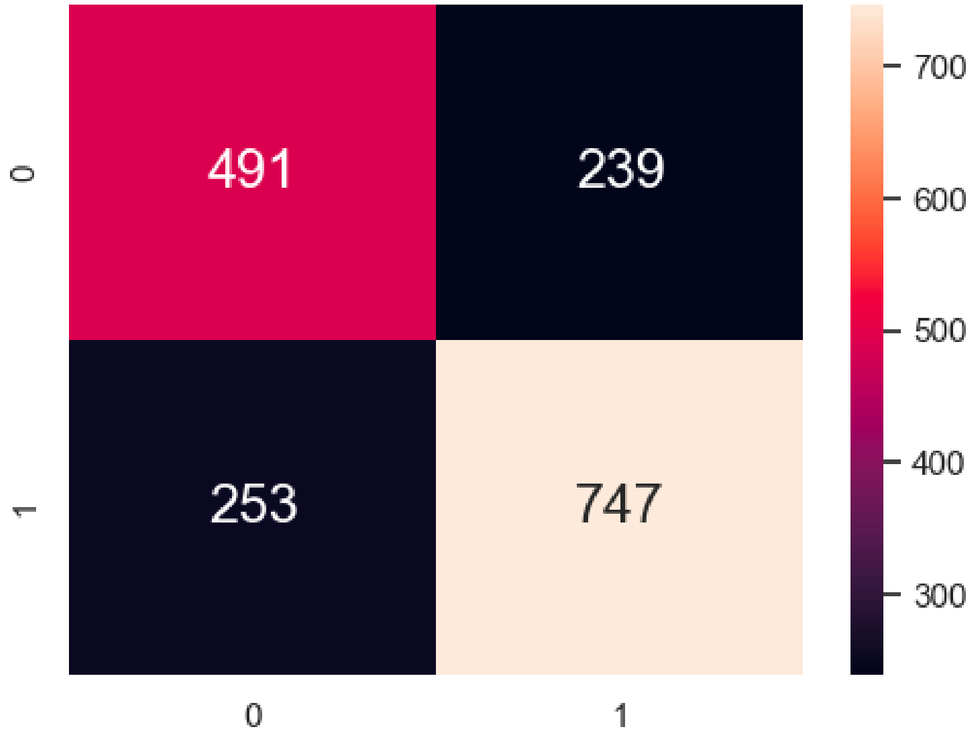
# Logistic Regression

## Data preperation

- Only 8% of dataset are pre-diabetic
- Bias dataset
- Make use of SMOTE()  to balance dataset
- SMOTE() increases the dataset by adding new data that falls within the same cluster as our dataset
- Machine Learning help to predict more accurate results

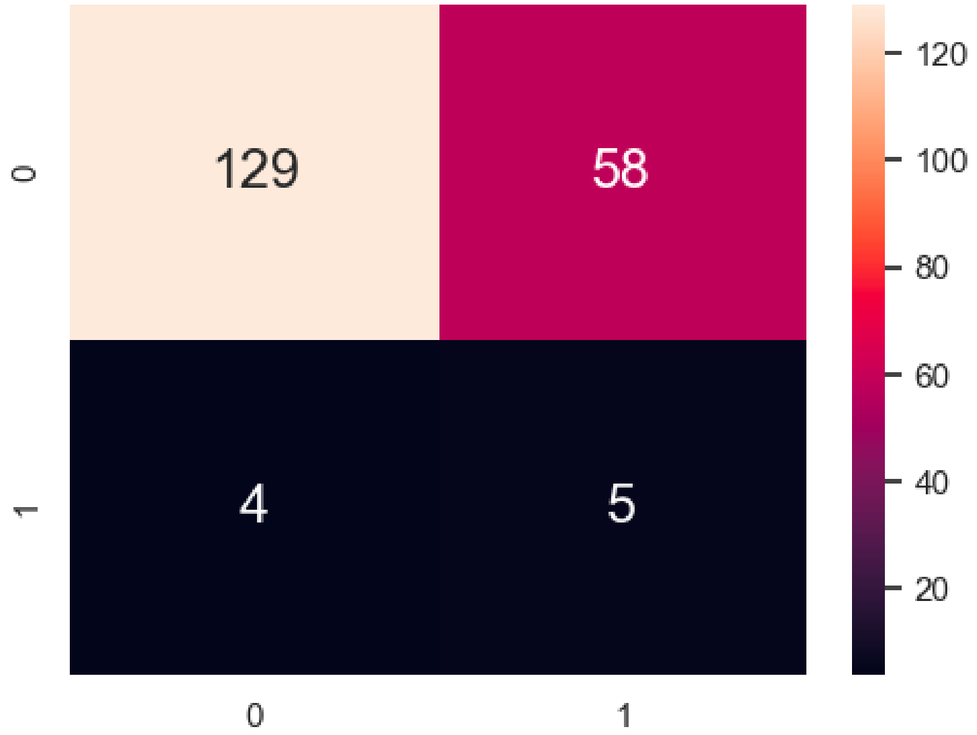
# Logistic Regression Results

## Train Dataset




```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.715606936416185
Mean Squared Error (MSE)        : 0.28439306358381505
Classification Accuracy         : 0.715606936416185
```

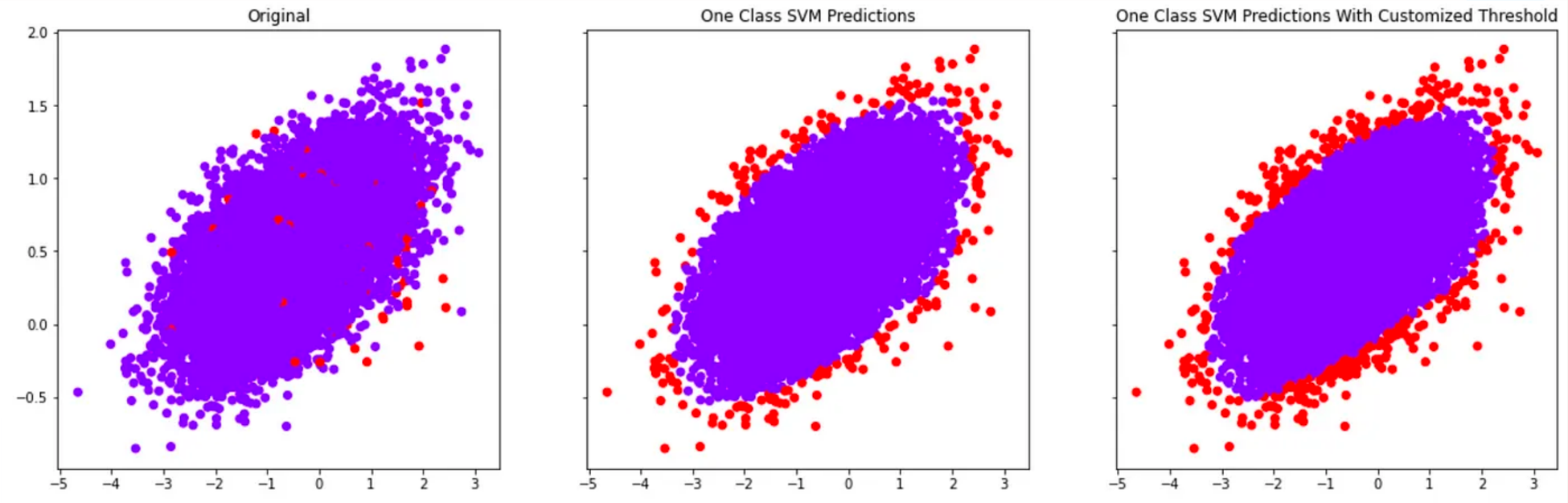
## Train Dataset




```
Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.6836734693877551
Mean Squared Error (MSE)        : 0.3163265306122449
Classification Accuracy         : 0.6836734693877551
```

# One Class SVM

## Introduction




Unsupervised machine learning model that identify anomalies from the dataset

# One Class SVM

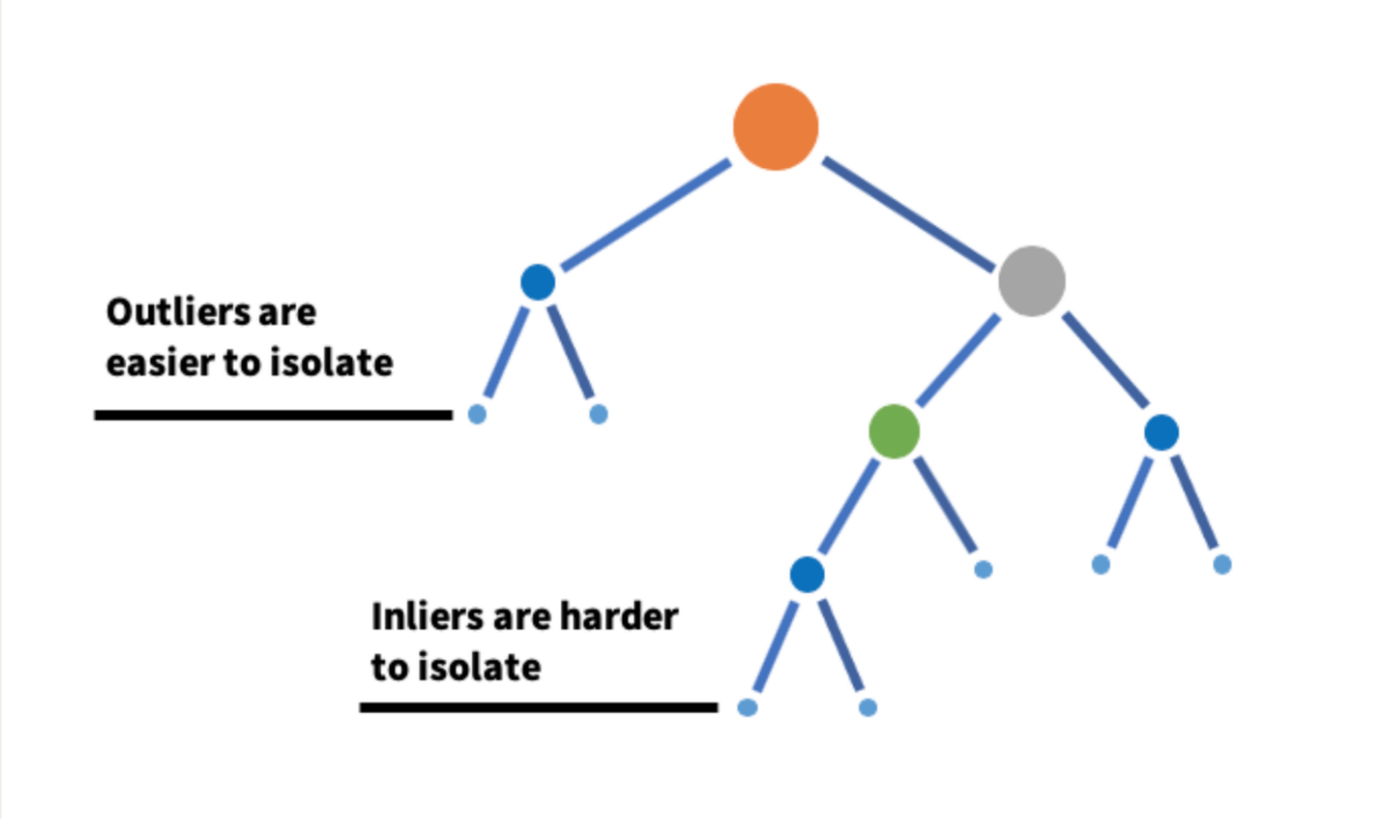
## Results

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.98 | 0.63 | 0.77 | 187 |
| 1.0 | 0.08 | 0.67 | 0.14 | 9 |
| accuracy | | | 0.63 | 196 |
| macro avg | 0.53 | 0.65 | 0.45 | 196 |
| weighted avg | 0.93 | 0.63 | 0.74 | 196 |

67% chance of correctly predicting one's diabetic status

# Isolation Forest

## Introduction




unsupervised machine learning model that identify anomaly datas from the dataset

# Isolation Forest

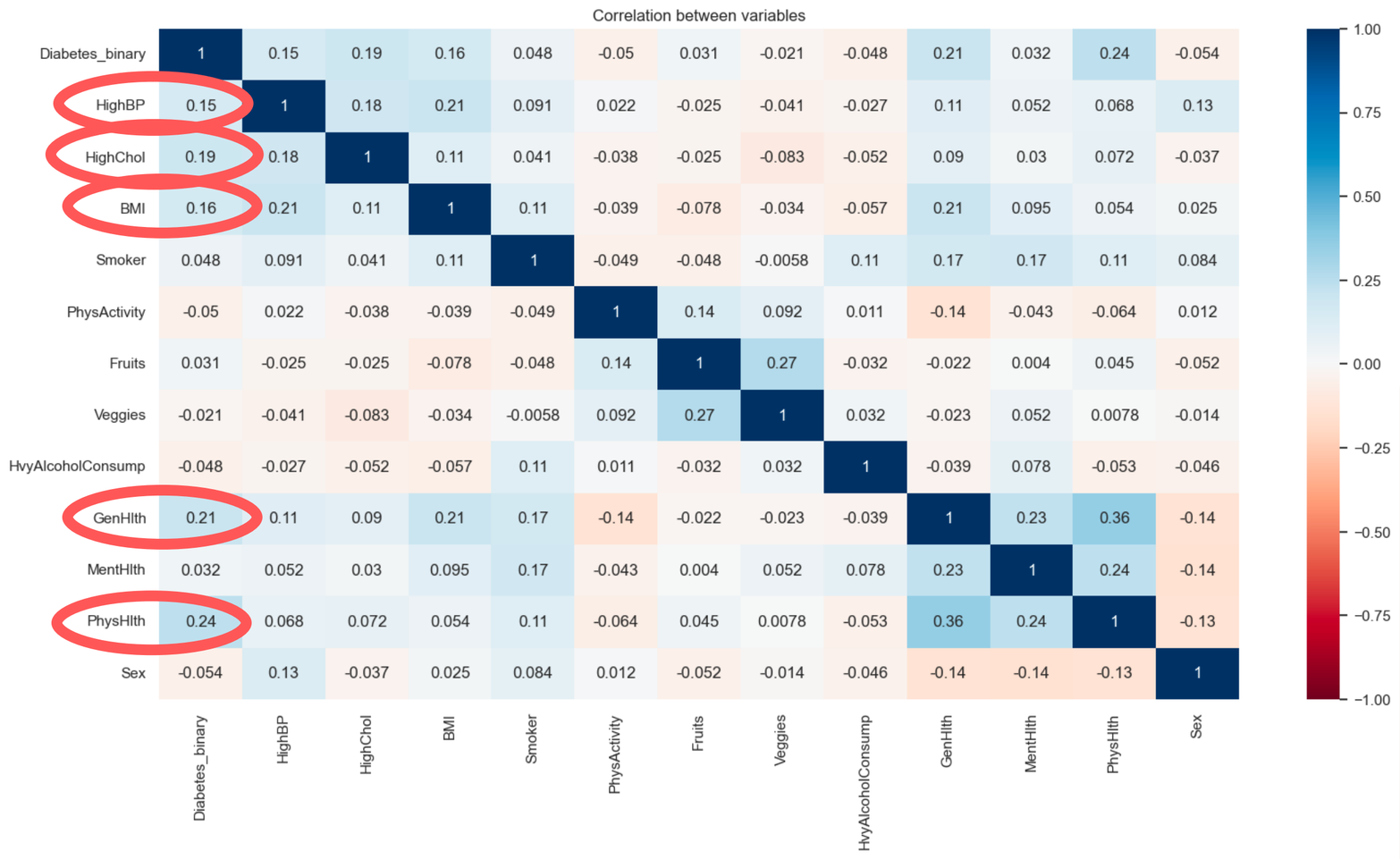
## Results

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.98 | 0.69 | 0.81 | 187 |
| 1.0 | 0.09 | 0.67 | 0.16 | 9 |
| accuracy | | | 0.69 | 196 |
| macro avg | 0.54 | 0.68 | 0.49 | 196 |
| weighted avg | 0.94 | 0.69 | 0.78 | 196 |

67% chance of correctly predicting one's diabetic status

# Machine Learning

## Machine learning methods



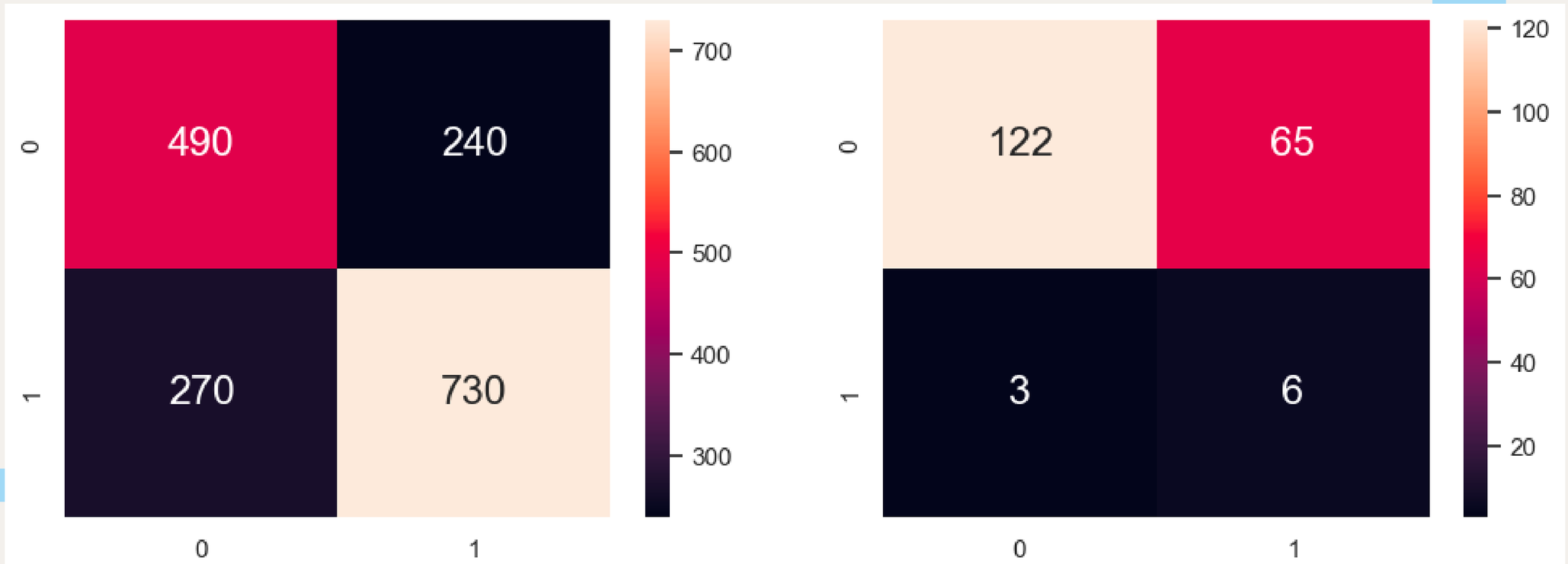

```
TOP 5 Variable with highest correlation

Diabetes_binary     1.000000
PhysHlth            0.237864
GenHlth             0.208867
HighChol            0.190936
BMI                 0.162042
HighBP              0.153232
Name: Diabetes_binary, dtype: float64
```

Using the top 5 correlated variables for Machine learning model

# Logistic Regression

## Results (Top 5 Variables)




```
Goodness of Fit of Model        Train Dataset
Explained Variance (R^2)        : 0.7052023121387283
Mean Squared Error (MSE)        : 0.2947976878612717
Classification Accuracy         : 0.7052023121387283

Goodness of Fit of Model        Test Dataset
Explained Variance (R^2)        : 0.6530612244897959
Mean Squared Error (MSE)        : 0.3469387755102041
Classification Accuracy         : 0.6530612244897959
```

# One Class SVM

## Results (Top 5 Variables)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.98 | 0.71 | 0.82 | 187 |
| 1.0 | 0.10 | 0.67 | 0.17 | 9 |
| accuracy | | | 0.71 | 196 |
| macro avg | 0.54 | 0.69 | 0.50 | 196 |
| weighted avg | 0.94 | 0.71 | 0.79 | 196 |

67% chance of correctly predicting one's prediabetic status

# Isolation Forest

## Results (Top 5 Variables)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.97 | 0.74 | 0.84 | 187 |
| 1.0 | 0.09 | 0.56 | 0.16 | 9 |
| accuracy | | | 0.73 | 196 |
| macro avg | 0.53 | 0.65 | 0.50 | 196 |
| weighted avg | 0.93 | 0.73 | 0.81 | 196 |

56% chance of correctly predicting one's prediabetic status

# Machine Learning Results

## Logistic Regression

similar accuracy result compared to model used for all variables

## One Class SVM

similar accuracy result compared to model used for all variables

## Isolation Forest

Lesser accuracy compared to model used for all variables.
Forest is build with a maximum depth of 5, which might not be that extensive resulting in a less accurate value

# Conclusion

- Individuals aged 18-24 less likely to be diagnosed with pre-diabetes, but risk is not negligible
- 3 different Machine Learning model incorporating Classification & Anomaly detection
- Used supervised and unsupervised learning for predicting an individual's diabetic status
- 5 Key Health Indicators
    - Physical Health
    - General Health
    - High Cholesterol
    - BMI Level
    - High Blood Pressure