SC1015 Mini Project

Group 10 FCS7

Muhammad Hanif (U2320378F), Lim En Jia (U2320279L), Harikrishnan Vinod (U2321114H)

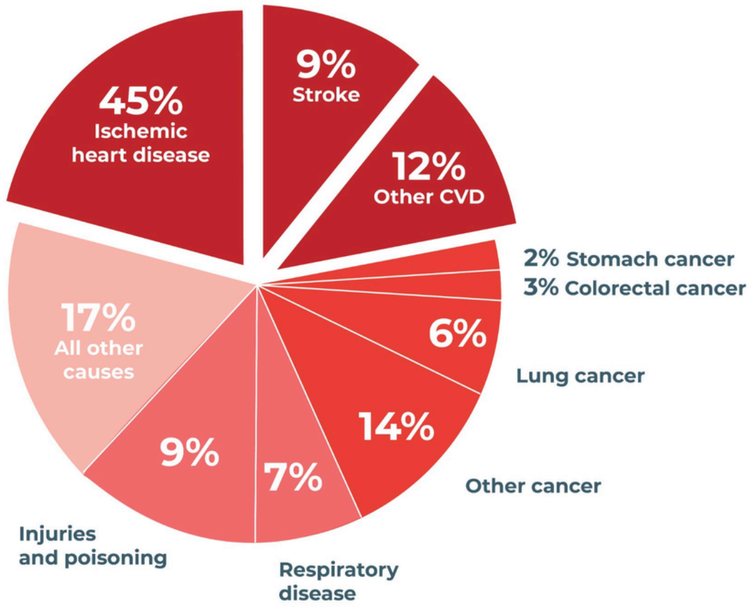
Background

World Health Organisation

Cardiovascular disease is the **leading cause of death globally**, taking an estimated **17.9 million lives** yearly, 32% of global deaths, **85% were due to heart attacks**

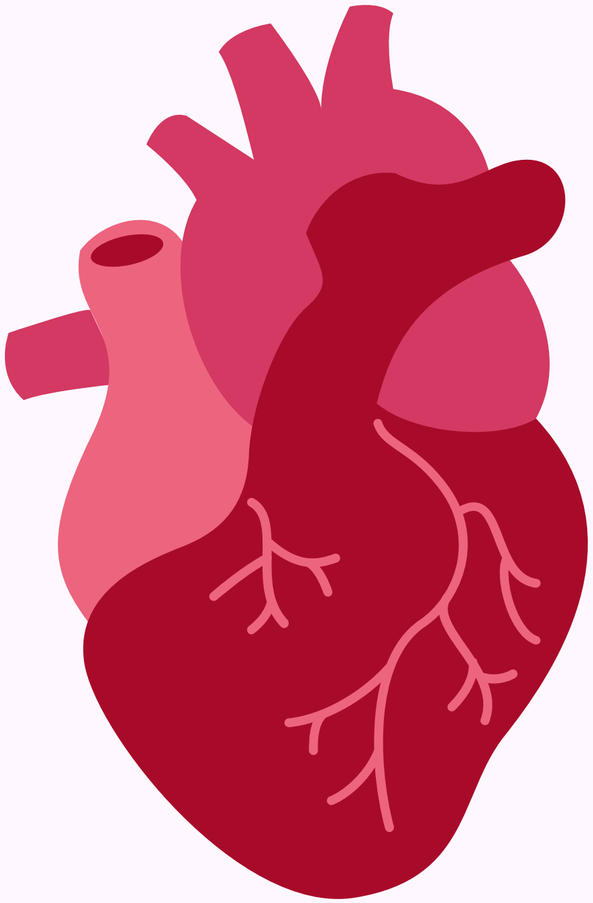
Singapore Myocardial Infarction Registry Report 2020

11631 heart attack cases per 100k people(7,344 in 2010), averaging 31 a day. Among them, **9.2% died within 30 days**



Background

Up to 80%
of premature heart attacks can be prevented with risk management.



From World Heart Report 2023

DATA SET

kaggle  SOURAV BANERJEE ·

Heart Attack Risk Prediction Dataset

Unlocking Predictive Insights with Multifaceted Synthetic Heart Attack Dataset

```
a=ol.drop(['Patient ID','Country','Hemisphere','Blood Pressure'],axis = 1)
a
```

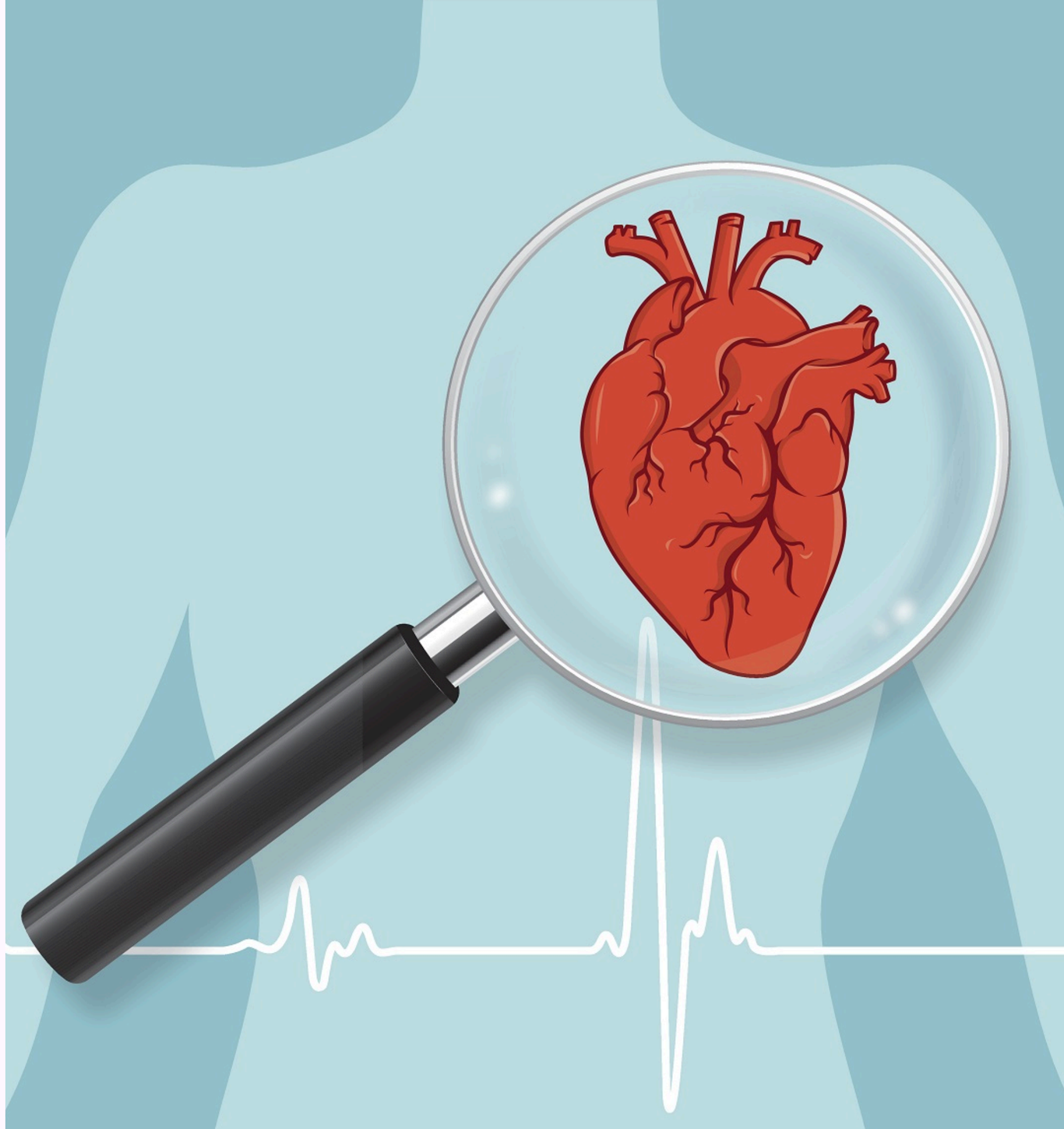
DATA SET

```
In [63]: ol.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8763 entries, 0 to 8762
Data columns (total 26 columns):
 #   Column                           Non-Null Count  Dtype  
---  ------                           --------------  -----  
 0   Patient ID                       8763 non-null   object 
 1   Age                              8763 non-null   int64  
 2   Sex                              8763 non-null   object 
 3   Cholesterol                      8763 non-null   int64  
 4   Blood Pressure                   8763 non-null   object 
 5   Heart Rate                       8763 non-null   int64  
 6   Diabetes                         8763 non-null   int64  
 7   Family History                   8763 non-null   int64  
 8   Smoking                          8763 non-null   int64  
 9   Obesity                          8763 non-null   int64  
 10  Alcohol Consumption              8763 non-null   int64  
 11  Exercise Hours Per Week          8763 non-null   float64
 12  Diet                             8763 non-null   object 
 13  Previous Heart Problems          8763 non-null   int64  
 14  Medication Use                   8763 non-null   int64  
 15  Stress Level                     8763 non-null   int64  
 16  Sedentary Hours Per Day          8763 non-null   float64
 17  Income                           8763 non-null   int64  
 18  BMI                              8763 non-null   float64
 19  Triglycerides                    8763 non-null   int64  
 20  Physical Activity Days Per Week  8763 non-null   int64  
 21  Sleep Hours Per Day              8763 non-null   int64  
 22  Country                          8763 non-null   object 
 23  Continent                        8763 non-null   object 
 24  Hemisphere                       8763 non-null   object 
 25  Heart Attack Risk                8763 non-null   int64  
```

Problem Definition

How do **different factors** affect the risk of heart attack among individuals?



Our Goal

To leverage **machine learning** to accurately **assess the risk of heart attacks** for individuals using the different **factors** from the dataset

Data Cleaning

Splitting Blood Pressure into Systolic and Dysystolic

```
split_data = ol["Blood Pressure"].str.split("/", expand = True)

ol["Systolic_BP"] = split_data[0]
ol["Diastolic_BP"] = split_data[1]

ol["Systolic_BP"] = ol["Systolic_BP"].astype("int64")
ol["Diastolic_BP"] = ol["Diastolic_BP"].astype("int64")
```

Factors Removed

Data Cleaning

Only Non-Null data

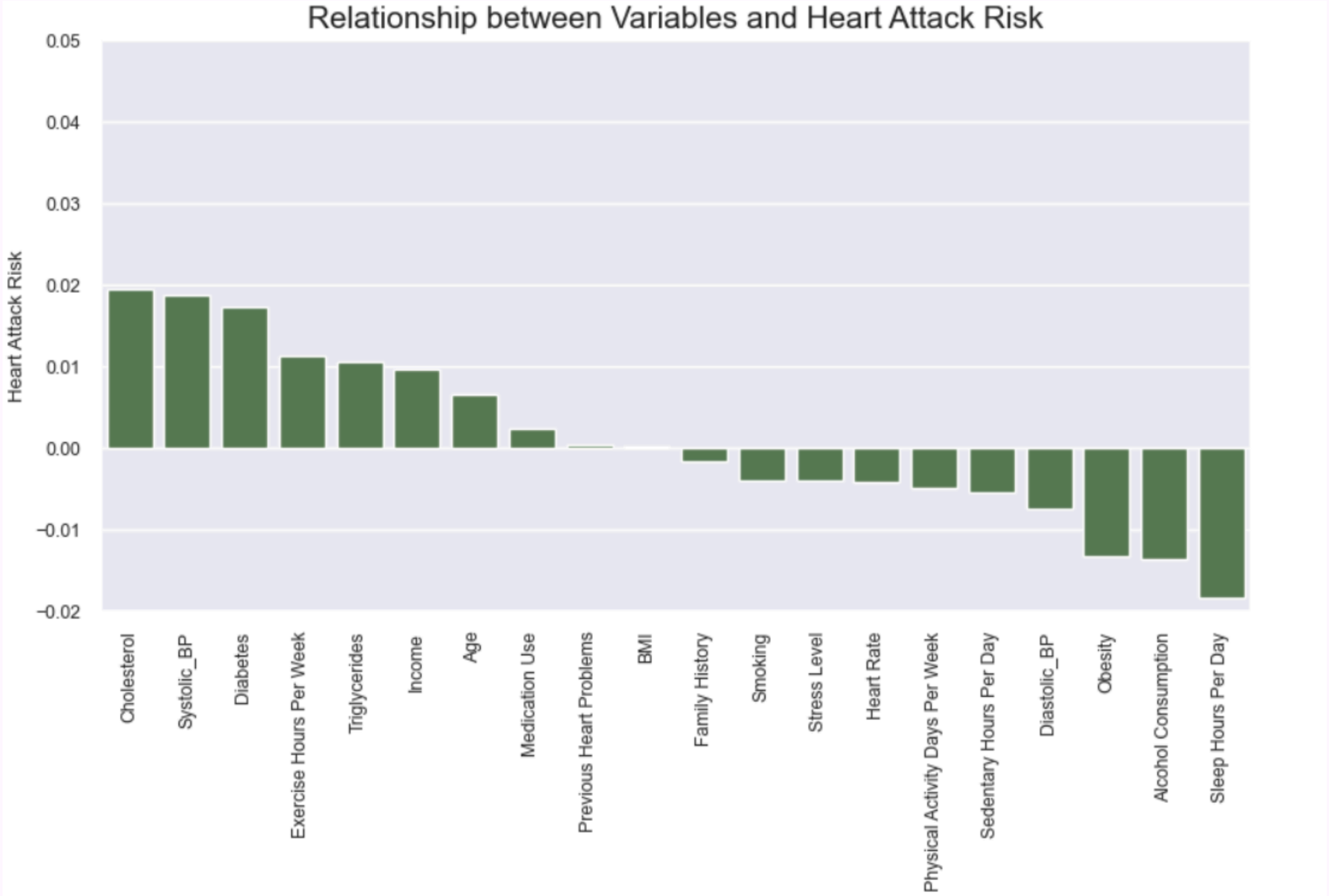
```
Non-Null Count
--------------
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
8763 non-null
```

Removing Outliers

```
Q1s = numeric_values.quantile(0.25)
Q3s = numeric_values.quantile(0.75)
IQRs = Q3s - Q1s
outliers = numeric_values[(numeric_values < (Q1s - threshold * IQRs)) | (numeric_values > (Q3s + threshold * IQRs))]
```

```
Number of outliers for Age : 0
Number of outliers for Cholesterol : 0
Number of outliers for Heart Rate : 0
Number of outliers for Exercise Hours Per Week : 0
Number of outliers for Stress Level : 0
Number of outliers for Sedentary Hours Per Day : 0
Number of outliers for Income : 0
Number of outliers for BMI : 0
Number of outliers for Triglycerides : 0
Number of outliers for Physical Activity Days Per Week : 0
Number of outliers for Sleep Hours Per Day : 0
Number of outliers for Systolic_BP : 0
Number of outliers for Diastolic_BP : 0
```

Exploratory Data Analysis



Data Categorisation

Demographics

```
demographics = a[['Age', 'Sex', 'Income','Heart Attack Risk']]
```

Health Status

```
health_status = a[['BMI', 'Heart Rate','Diastolic_BP','Systolic_BP','Heart Attack Risk']]
```

Risk Factors

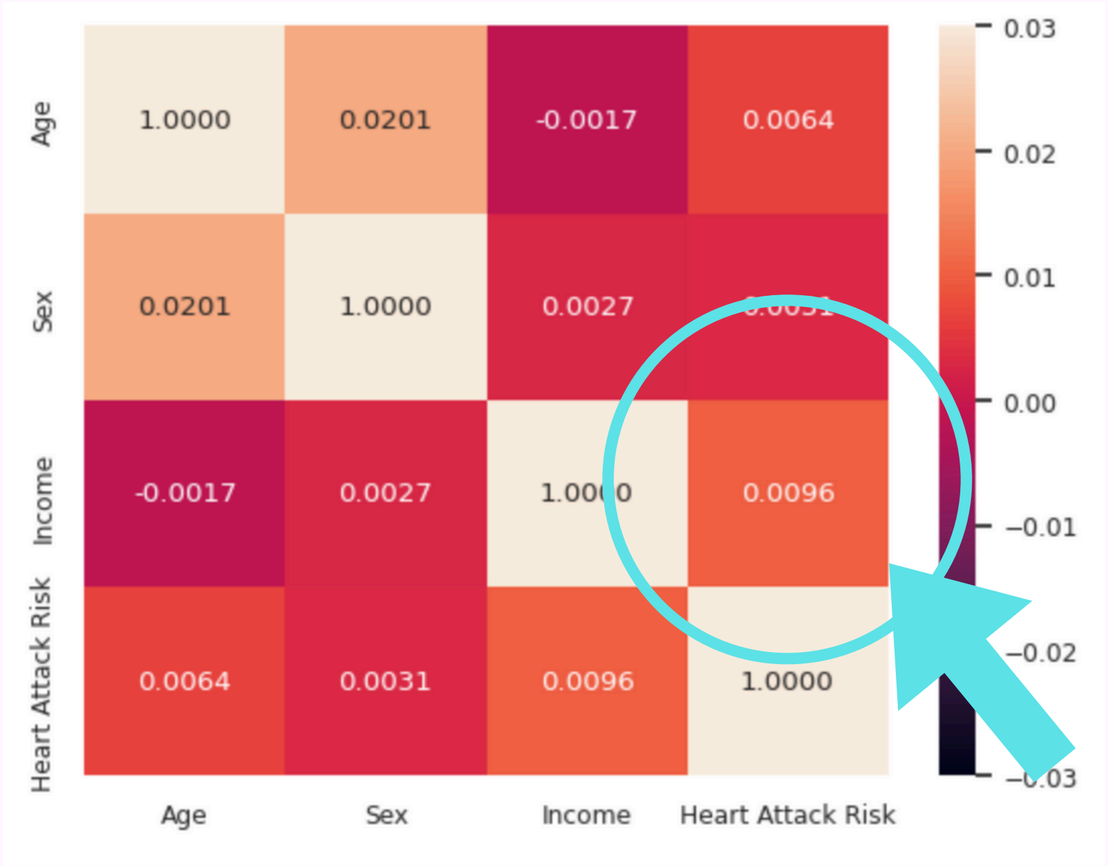
```
risk_factors = a[['Cholesterol','Diabetes', 'Obesity', 'Triglycerides', 'Heart Attack Risk']]
```

Medical History

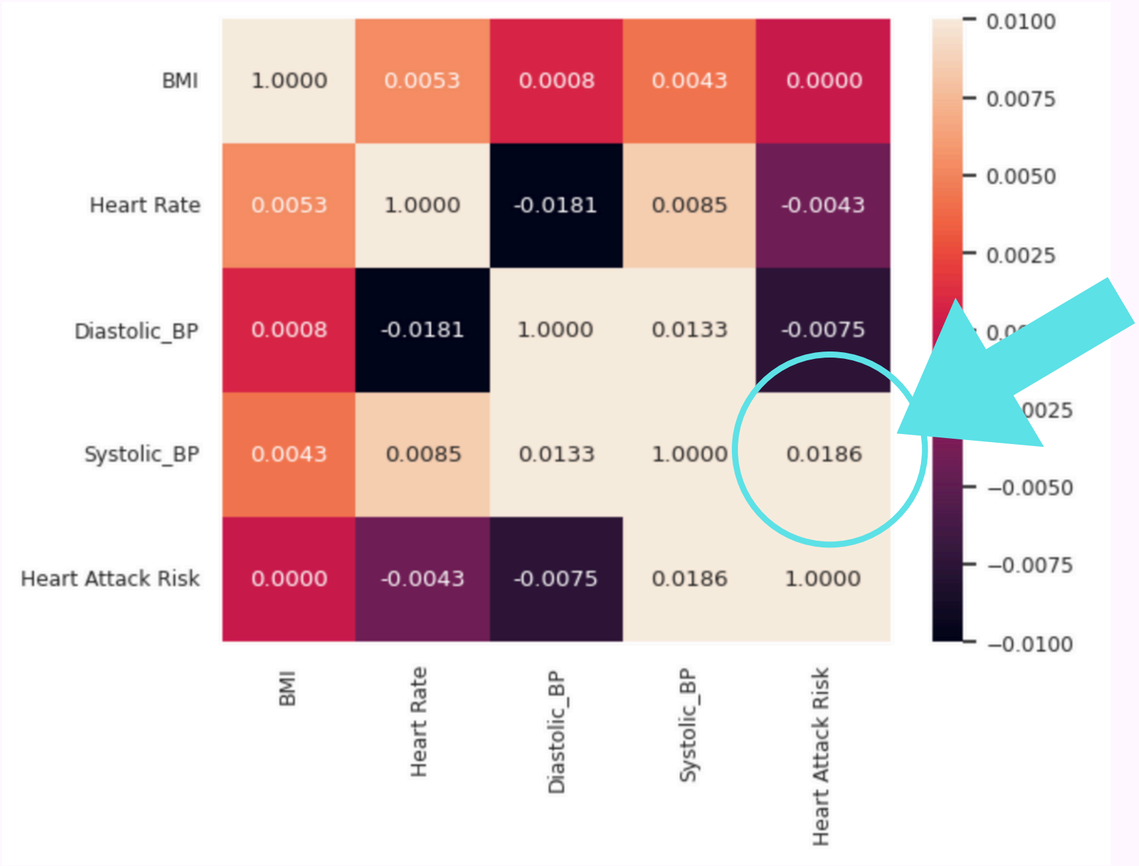
```
medical_history = a[['Previous Heart Problems', 'Medication Use','Heart Attack Risk']]
```

Heat Maps

Demographics

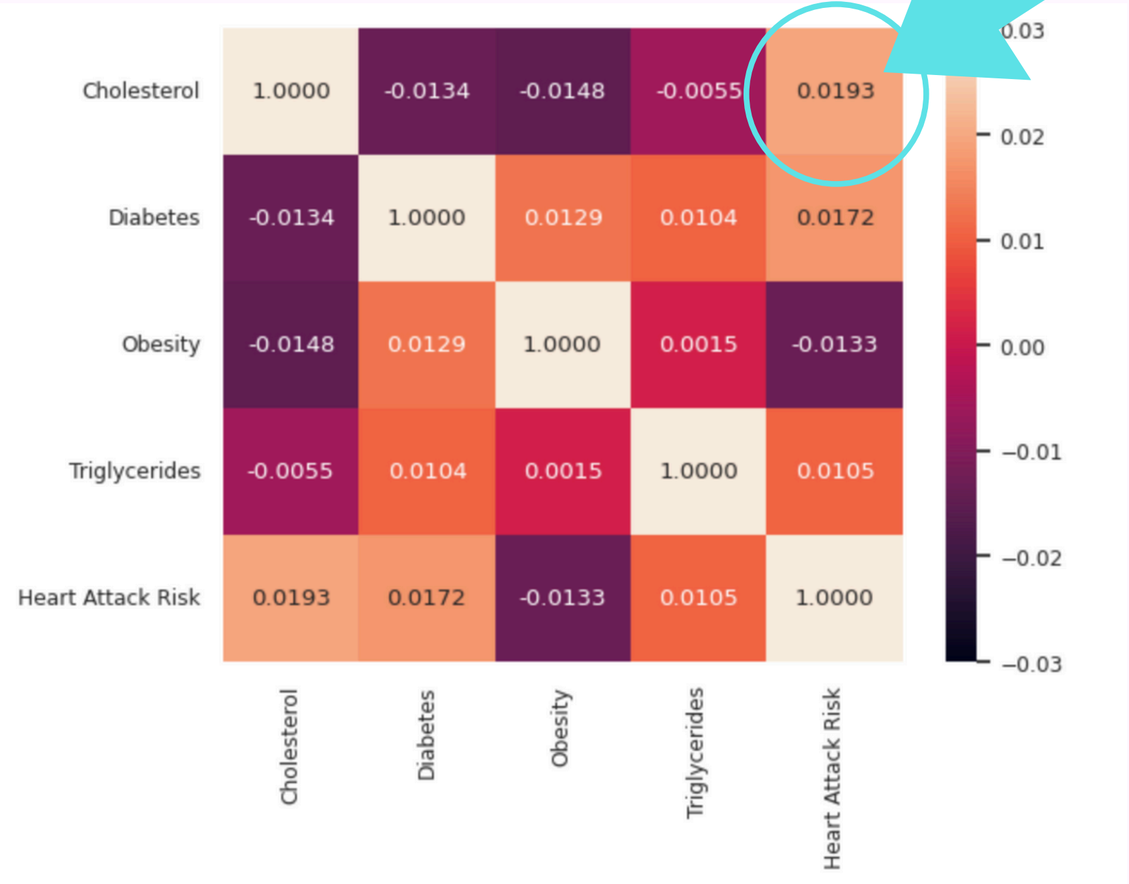


Health Status

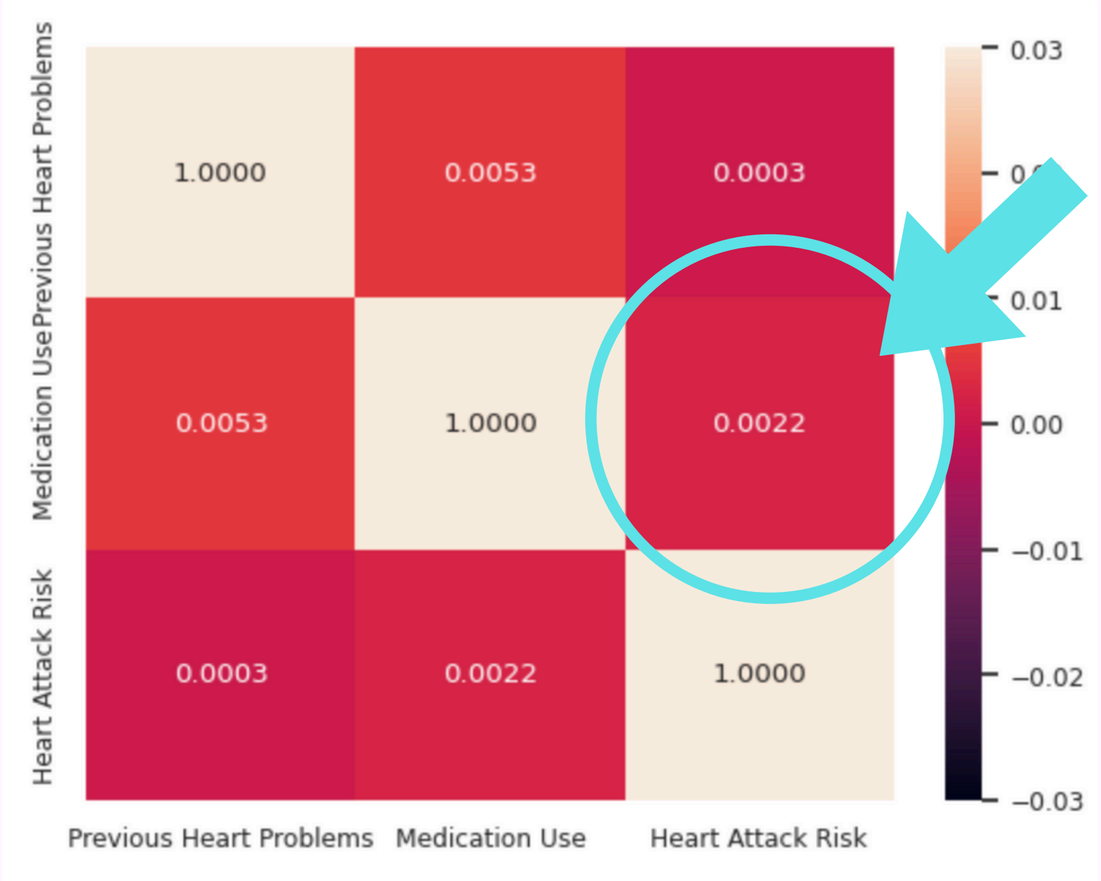


Heat Maps

Risk Factors



Medical History

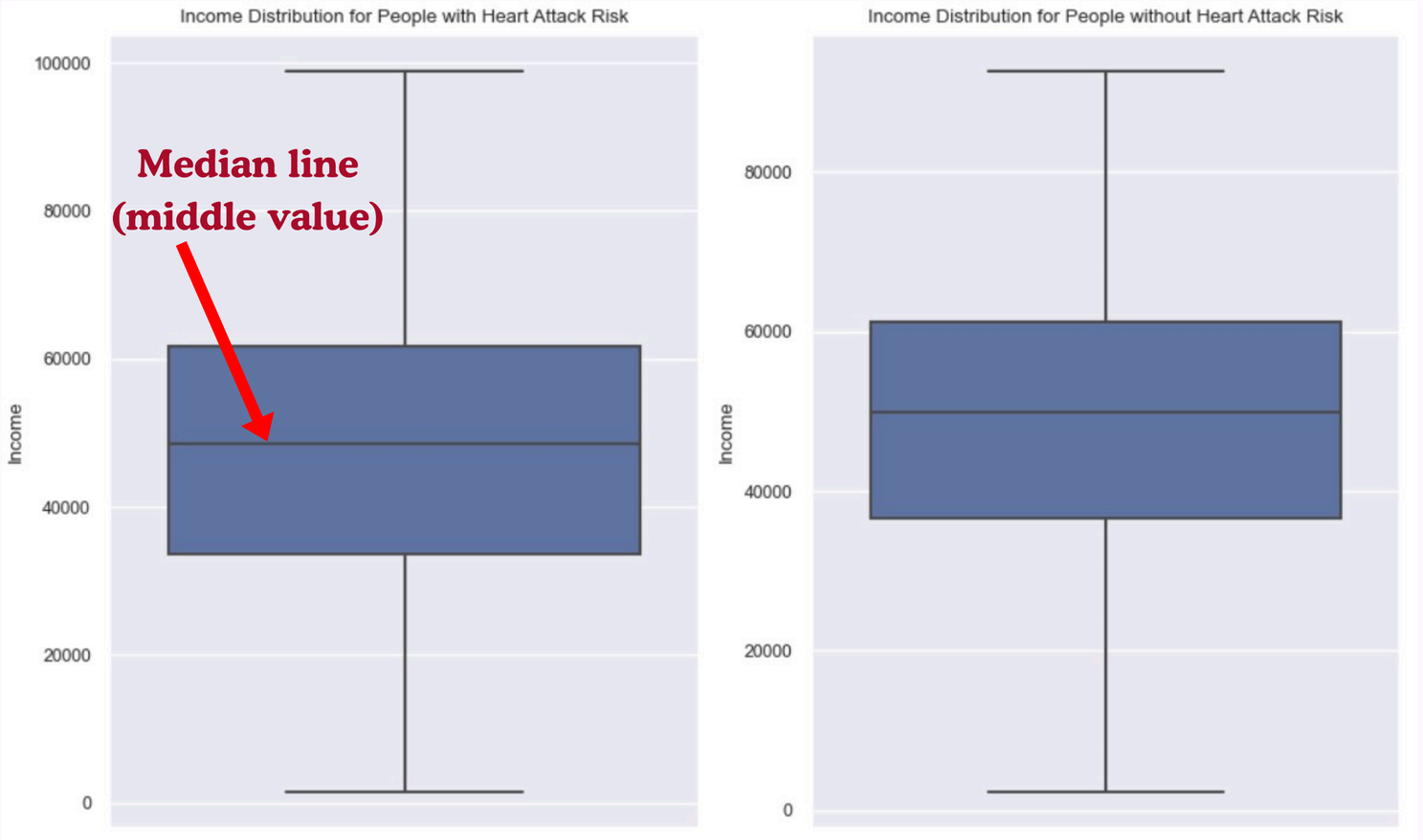


Variables chosen

Demographics	→	Income
Health Status	→	Systolic Blood Pressure
Risk Factors	→	Cholesterol
Medical History	→	Medication Use

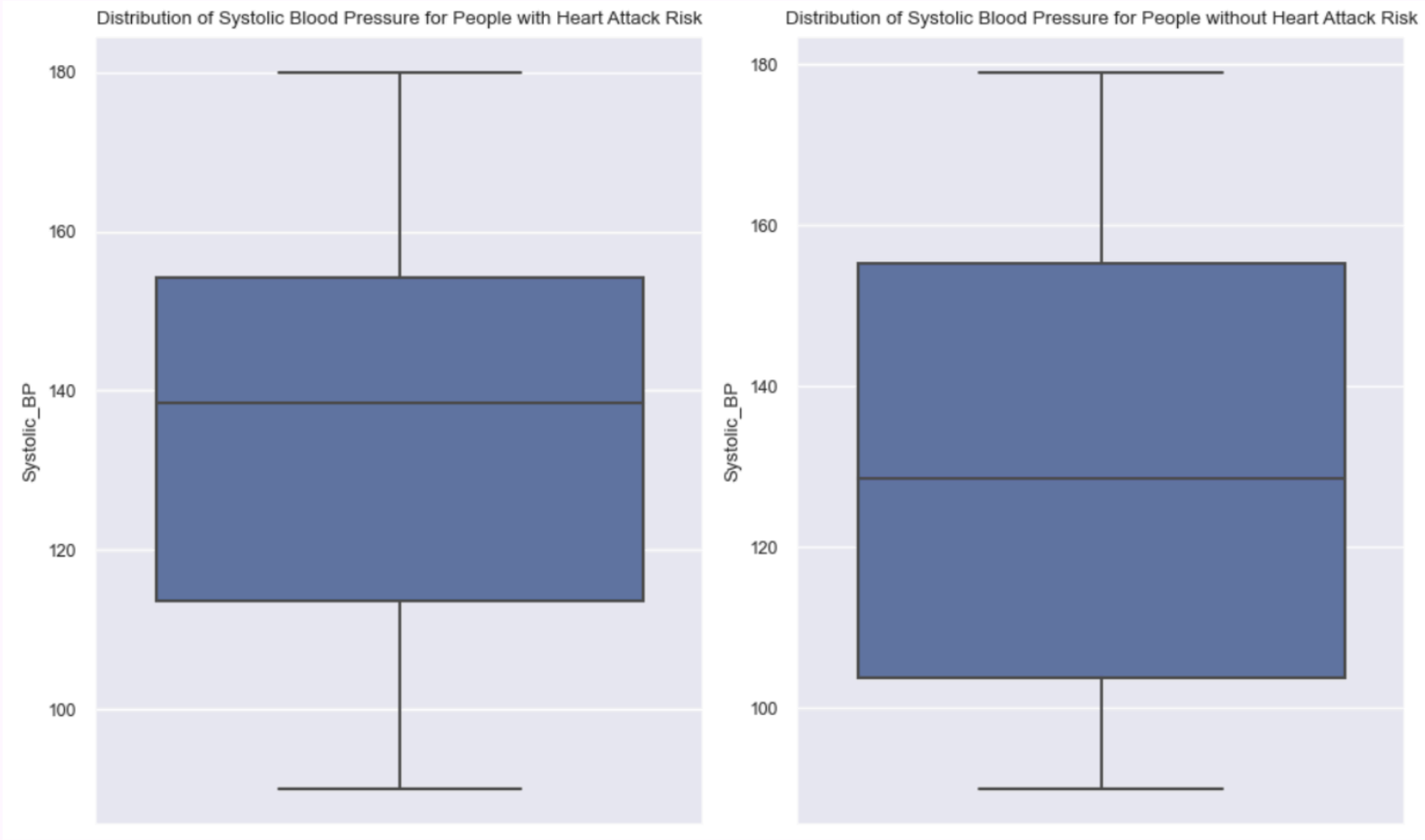
Income

Finding #1:
People with **higher Income Level** tend to have a **lower Heart Attack Risk**



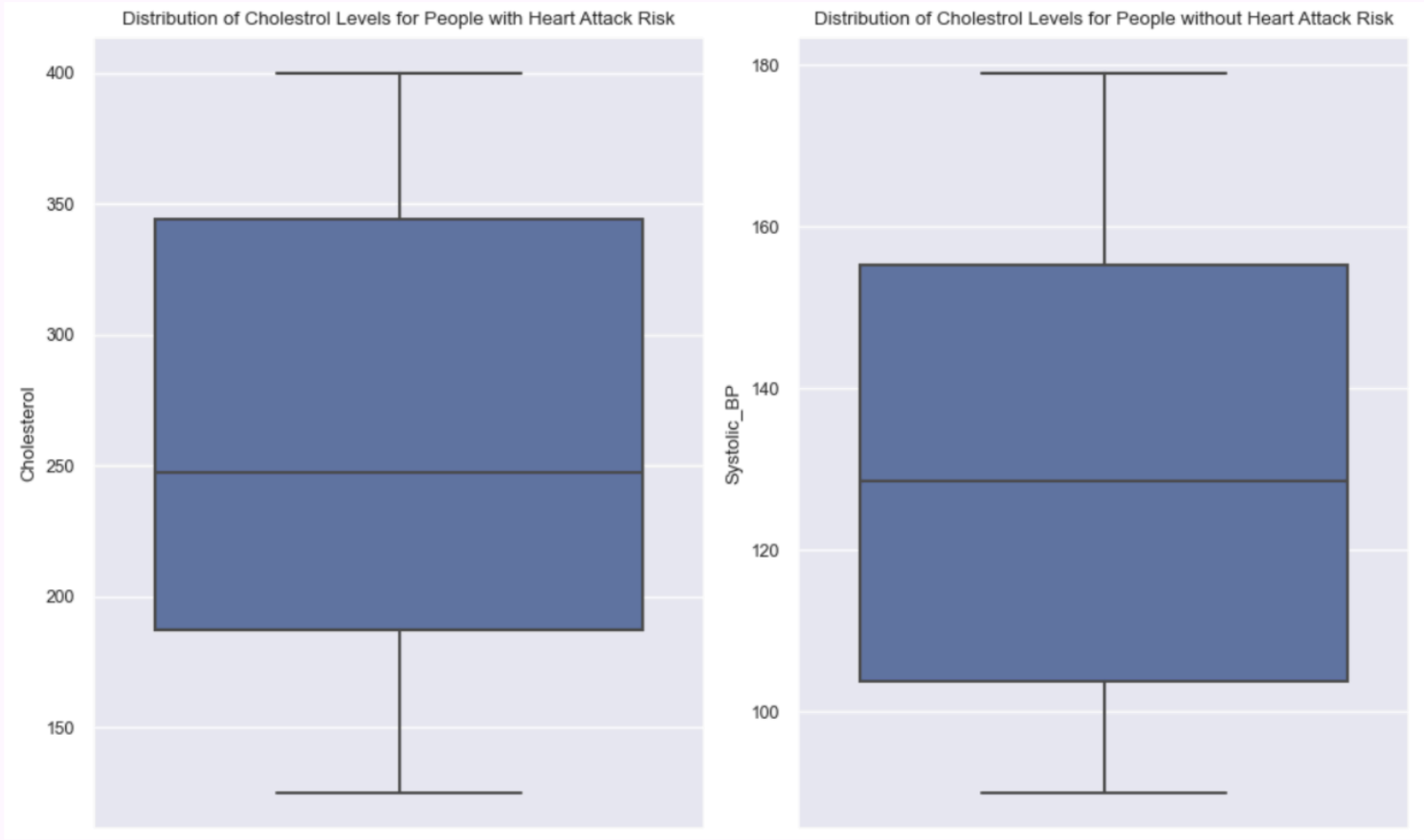
Systolic Blood Pressure

Finding #2:
Higher Systolic Blood Pressure points to a **higher Heart Attack Risk**



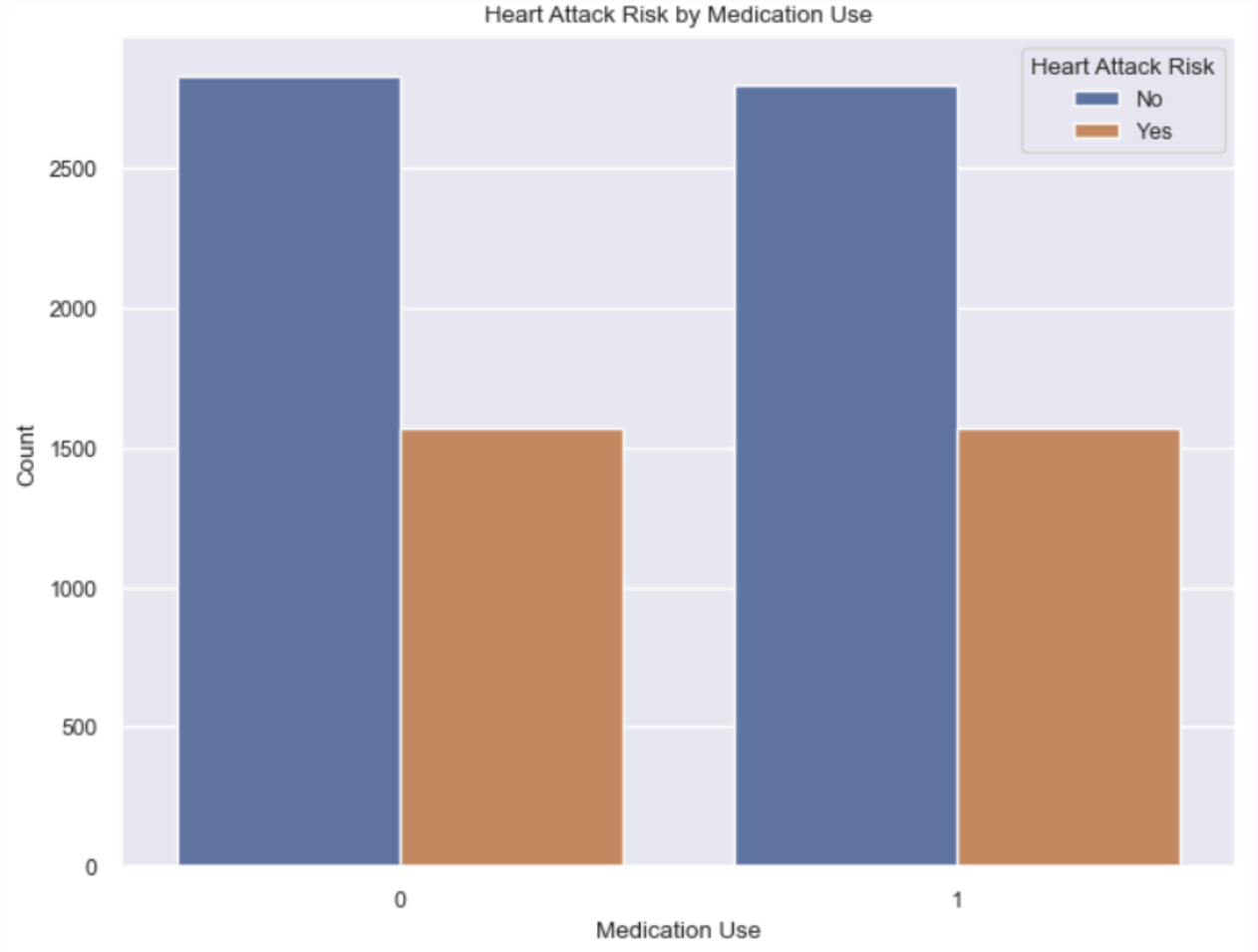
Cholesterol

Finding #3:
Higher cholesterol tends to lead to a **higher Heart Attack Risk**



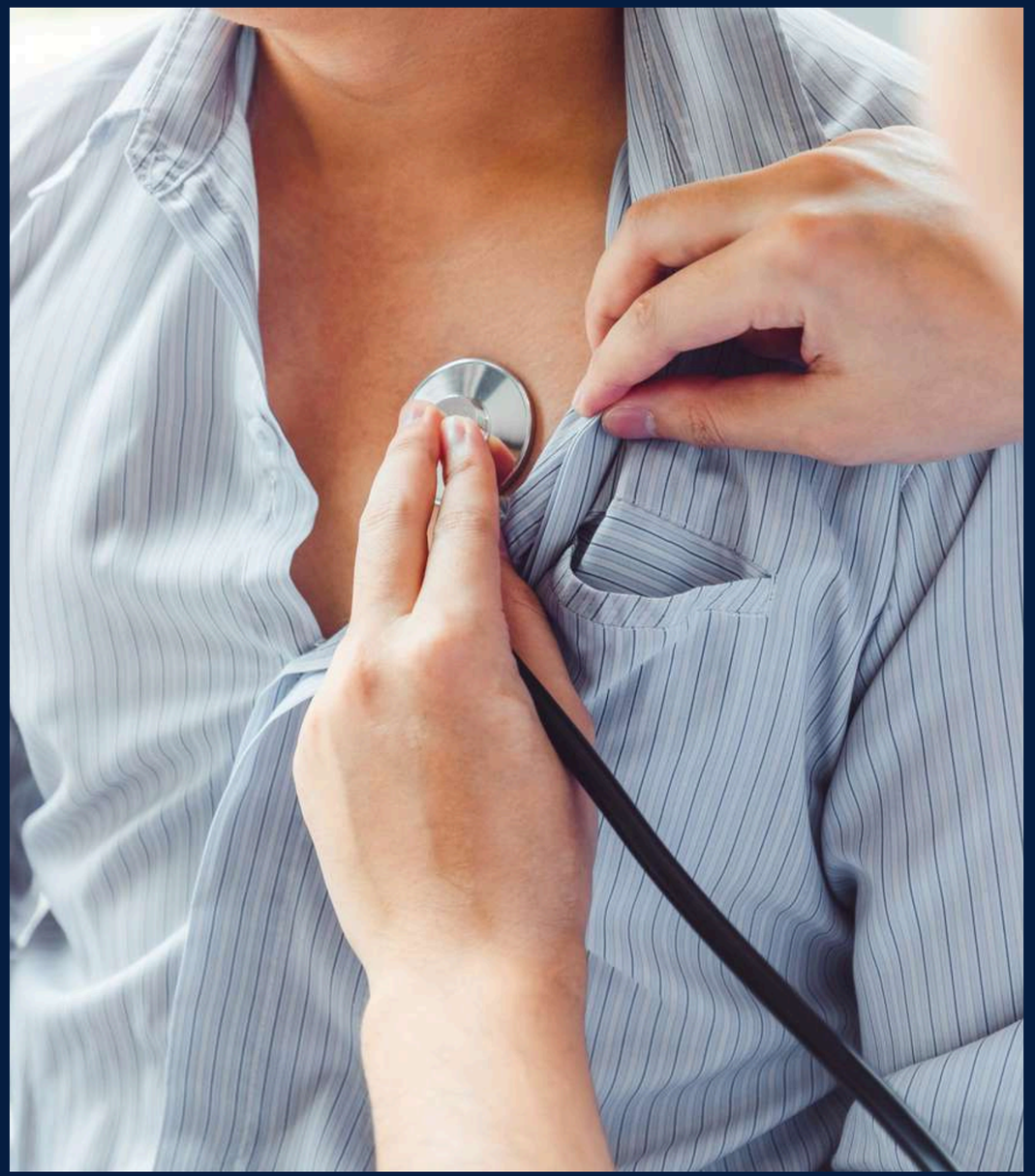
Medication Use

Finding #4:
Despite having the highest correlation value under Medical History sub-categorical group,
Medication Use does not affect Heart Attack Risk

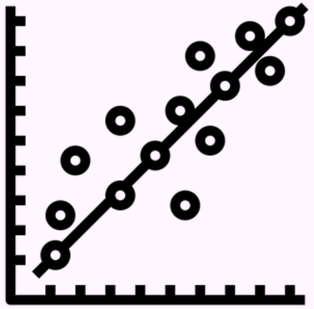


Our Task

To accurately classify the Heart Attack Risk (0 or 1) of an individual based on these factors using Machine Learning

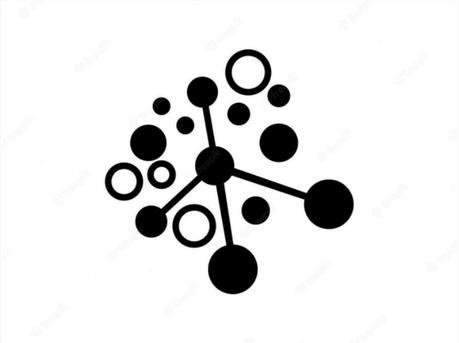


Machine Learning Models

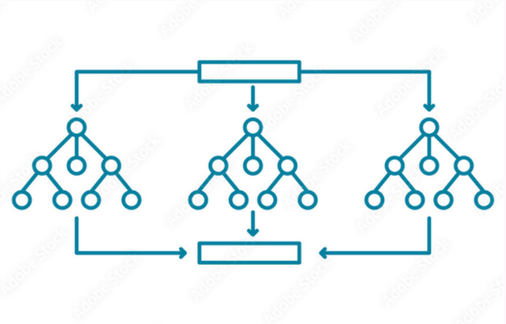


Linear Regression

Logistic Regression



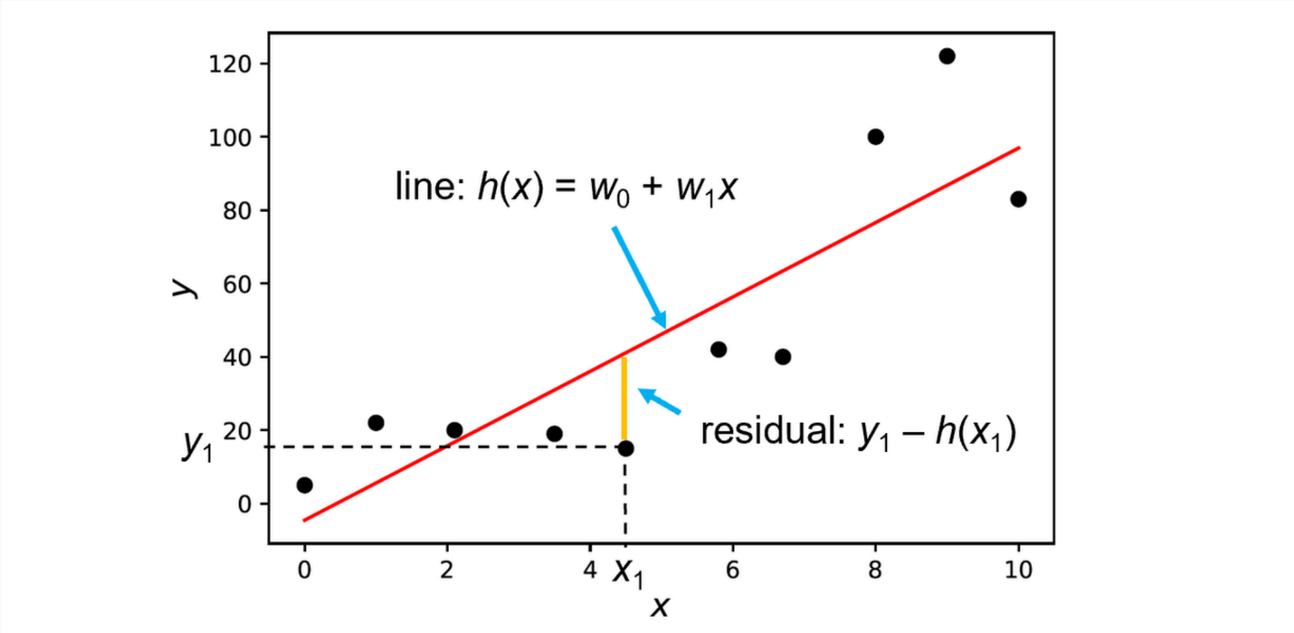
Nearest Neighbour



Random Forest

Linear Regression

Describes the relationship between a dependent variable (y) and one or more independent variables (x).The goal is to find the **best-fitting linear line** that **minimises the sum of squared differences** between the **observed** and **predicted** values.



Uni-Variate Linear Regression (Goodness of Fit)

Split Data Set Randomly into Train (70%) and Test (30%) (Train,Test)	Income	Systolic Blood Pressure	Cholesterol	Medication Use
Explained Variance	-1x10^-2	-9.3	0.007	-0.0005
Mean Squared Error	(0.23)	(0.23)	(0.229)	(0.23)
Classification Accuracy (Test Data)	0.54	0.63	0.619	0.635

Linear Regression

We found out that linear regression is **NOT** an accurate model to represent Heart Attack Risk.

Heart Attack Risk (dependent variable) takes a **binary value(0 or 1)**.

Linear Regression assumes a linear relationship between the independent variables and the dependent variable.

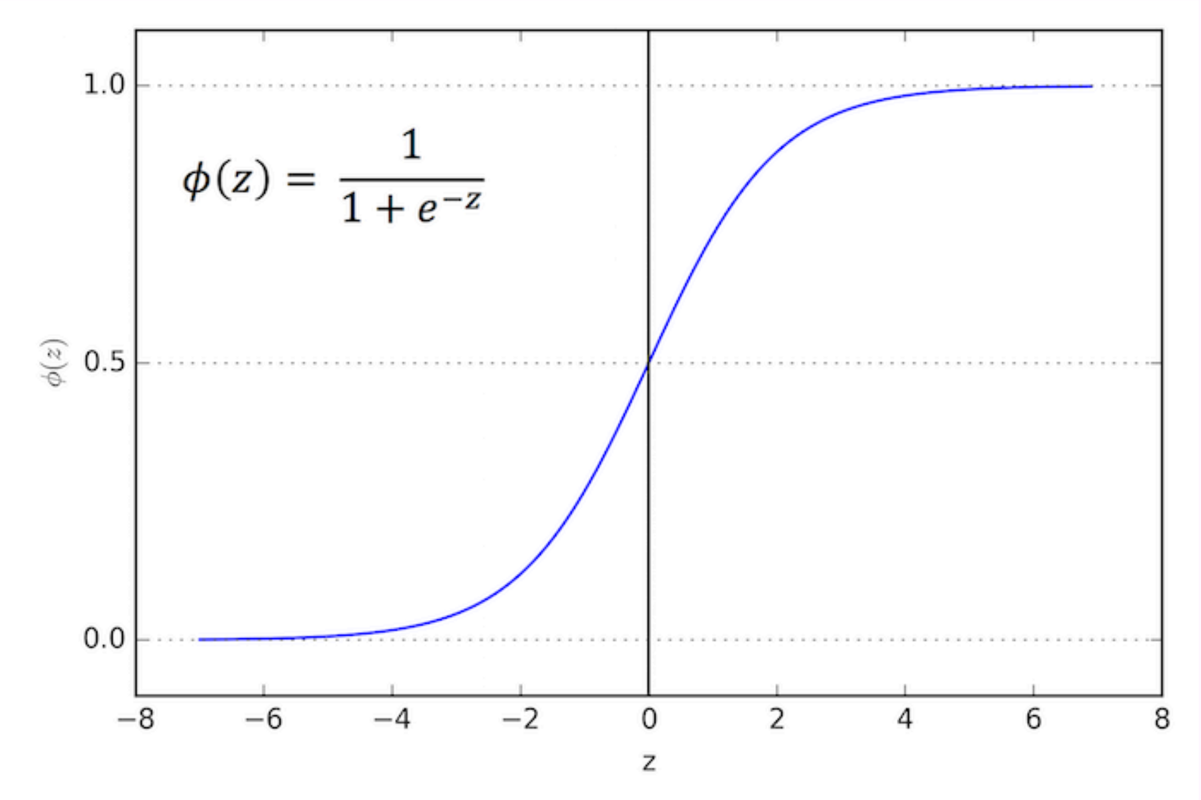
- Unable to capture the non-linear relationships between predictors and binary outcomes.

Invalid assumption leads to error in coefficients of the dependent variables.

Logistic Regression

A logistic function is used to model the relationships between predictor variables and the binary response variable.

Goal : Estimate the coefficients of the logistic regression model that maximise the probability of the observed data



Uni-Variate Logistic Regression

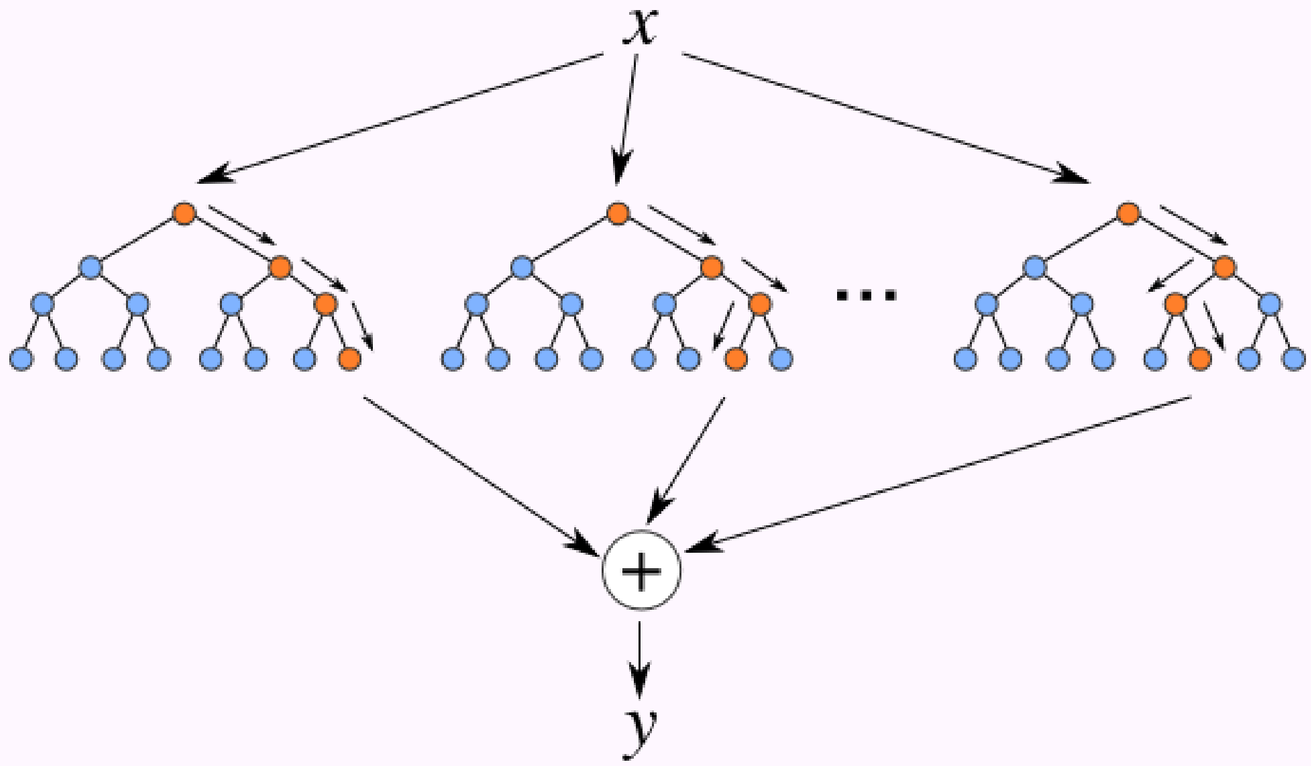
Split data set randomly into Train (70%) and Test (30%)	Income	Systolic Blood Pressure	Cholesterol	Medication Use
Classification Accuracy (Test Data)	0.6428	0.6386	0.6477	0.6314

Multi-Variate Logistic Regression

Split data set randomly into Train (70%) and Test(30%)	Income + Systolic Blood Pressure + Cholesterol +Medication Use
Classification Accuracy (Test Data)	0.6432

Random Forest

A Random Forest is like a group decision-making team in machine learning. It combines the opinions of many “trees”. It then takes these many decision trees and combines them to avoid overfitting and to produce more accurate predictions.



Random Forest Machine Learning

1st Iteration

TPR	0.3742
TNR	0.6422
FPR	0.3578
FNR	0.6258

Accuracy
0.6040

2nd Iteration

TPR	0.0379
TNR	0.9527
FPR	0.0473
FNR	0.9621

Accuracy
0.6386

Random forest, with balanced data

```
from imblearn.over_sampling import SMOTE

smote = SMOTE(random_state = 42)

X_smote, y_smote = smote.fit_resample(X, y)
```

Final Iteration Accuracy:0.6930

TPR	0.6905
TNR	0.6954
FPR	0.3046
FNR	0.3095

Attempted GridSeach CV to find Best Model

Accuracy
0.6386

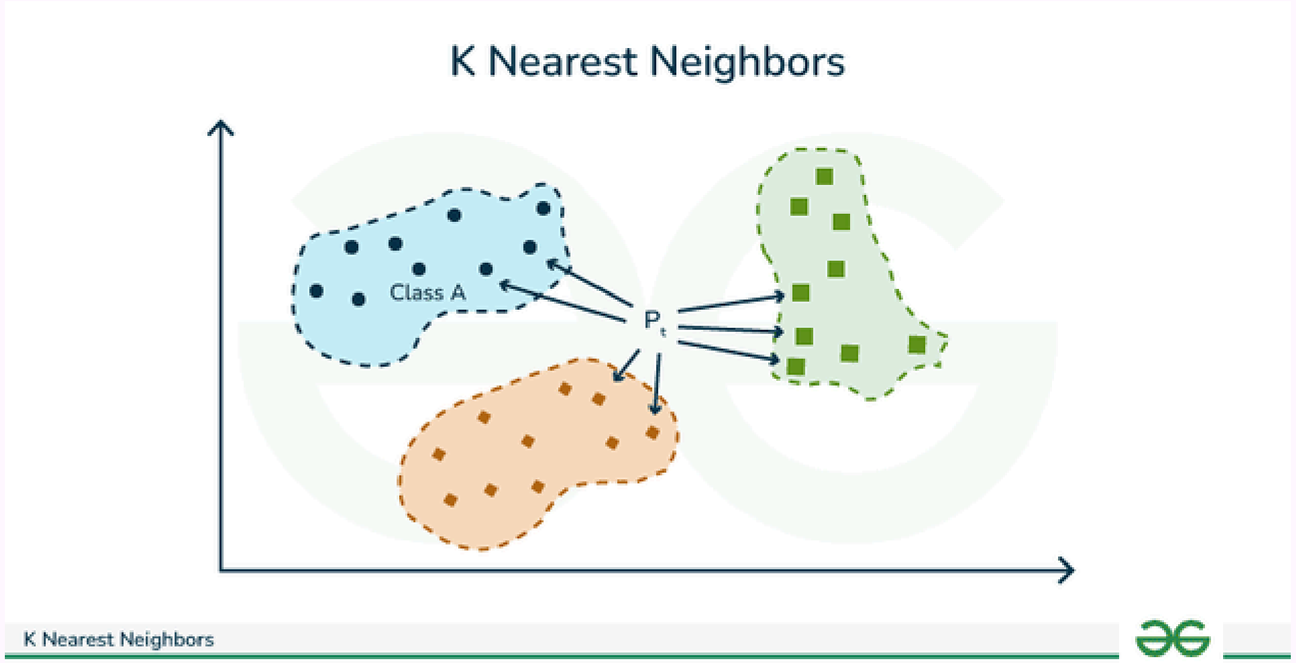
```
TPR Test :      0.0
TNR Test :      1.0

FPR Test :      0.0
FNR Test :      1.0
```

K-Nearest Neighbour

A non-parametric, distance-based supervised learning classifier that works based on the principle that data points with similar characteristics tend to belong to the same class or have similar values

Goal: Locate all closest neighbours of a data point to figure out which class it belongs to

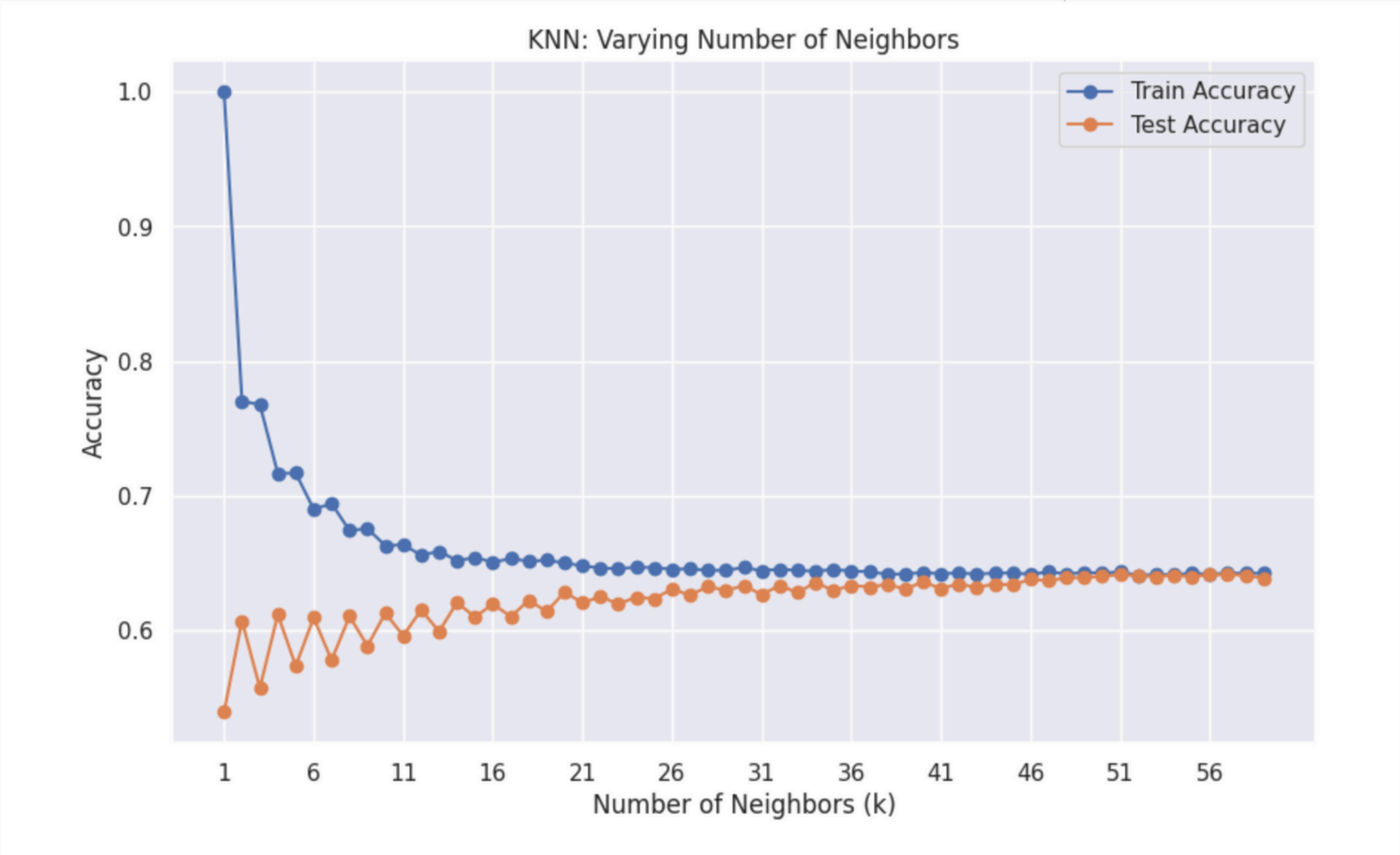


K-Nearest Neighbour

Split data set randomly into Train (70%) and Test(30%) (Train,Test)	Income + Systolic Blood Pressure + Cholesterol +Medication Use
Classification Accuracy (Test Data)	0.6071

TPR	0.1205
TNR	0.8770
FPR	0.1230
FNR	0.8795

K-Nearest Neighbour



Results

```
Best Model Test Accuracy: 0.6378851274248764
Best Model Parameters: {'n_neighbors': 46}
```

K-Nearest Neighbour

Split data set randomly into Train (70%) and Test(30%) (Train,Test)	Income + Systolic Blood Pressure + Cholesterol +Medication Use
Classification Accuracy (Test Data)	0.6379

TPR	0.0160
TNR	0.9830
FPR	0.0171
FNR	0.9840

Conclusion

- After trying out 4 machine learning models, Random Forest is the best at predicting Heart Attack Risk among individuals
- Highest Accuracy of 0.693 and F1-Score of 0.70.

However, these values are not considered high enough. It shows that the current set of data and variables used are not able to accurately predict Heart Attack Risk. However, these are proven to be the variables from each category with the highest correlation to Heart Attack Risk.

Conclusion

Complexity

It reveals that the complexity of factors contributing to heart attack risk and the variables included in our analysis cannot fully capture it.
For Example:

- BMI/Obesity is not an accurate representation (Body Builders)
- One might have been living a healthy lifestyle and only just started living an unhealthy lifestyle. (Past lifestyle choices not considered)

Proper medical advice and tests from professionals would be a more accurate way to determine one's heart risk. As such, we believe that further research incorporating advanced modelling techniques and a broader range of variables may be necessary to gain deeper insights into the complex nature of heart attack risk and improve the prediction accuracy. Exploring additional variables, collecting more comprehensive datasets, or experimenting with different feature engineering techniques could potentially improve predictive performance.

Thank You

References

- iStock.(2018).Heart Check Up.https://www.gettyimages.com/detail/illustration/heart-checkup-royalty-free-illustration/484003828?adppopup=true
- Singapore Heart Foundation.(n.d.).Heart Attack. https://www.myheart.org.sg/health/heart-conditions/heart-attack/
- Pinngam, S.(n.d). EyeEm. Getty Images.
- Becris.(n.d.). Linear Regression. https://www.flaticon.com/free-icon/linear-regression_2103601
- Flaticon.(n.d). Logistic Regression. https://www.flaticon.com/free-icon/logistic-regression_1998661
- Vector, T.(n.d). Random Forest Line Icon. https://stock.adobe.com/sg/images/random-forest-line-icon-decision-trees-symbol-machine-learning-technique-that-s-used-to-solve-regression-and-classification-problems-complex-problems-solution-vector-illustration-flat-clip-art/474661732
- Vitality Learning.(Sep 29,2022).Nearest neighbor with TensorFlow. https://vitalitylearning.medium.com/nearest-neighbor-with-tensorflow-a3875eaaa9a3
- Yehoshua,R.(April 19,2023). Linear Regression In Depth (Part 1). https://towardsdatascience.com/linear-regression-in-depth-part-1-485f997fd611
- Chaya. (June 9, 2020). Random Forest Regression. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
- GeeksforGeeks. (14 December,2023). How to Find The Optimal Value of K in KNN. https://www.geeksforgeeks.org/how-to-find-the-optimal-value-of-k-in-knn/