Correlation between Risk Factors and Coronary Heart Disease (CHD)

-Group N3-

Au Yat Sin Candice 20860834
Chan Chung Yin 20978198
Chui Yuen Tsun 20864165

Introduction

01

~37,000 deaths
were related to coronary heart diseases in 2021

02

3rd most common
cause of death in Hong Kong

03

Risk factors
like smoking, drinking, unhealthy diet, physical inactivity

04

Unpredictable Complications
even experienced specialists may not always foresee the development of these complications

Questions we would like to answer

(1) How different risk factors affect the prevalence rate of CHD?

(2) Which factors are positively/negatively correlated with heart disease or attack?

(3) Do other common complications share any similarities with CHD in the correlations between different risk factors?

Data collection and processing

Data collection

1. Dataset derived from a survey conducted from the Behavioural Risk Factor Surveillance System:

    - Respondents were asked about:
        - Health status & living habit
        - Have CHD or not

BRFSS™
Behavioral Risk Factor Surveillance System

Data cleaning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 20 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   HeartDiseaseorAttack  253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   BMI                   253680 non-null  float64
 4   Smoker                253680 non-null  float64
 5   Stroke                253680 non-null  float64
 6   Diabetes              253680 non-null  float64
 7   PhysActivity          253680 non-null  float64
 8   Fruits                253680 non-null  float64
 9   Veggies               253680 non-null  float64
 10  HvyAlcoholConsump     253680 non-null  float64
 11  AnyHealthcare         253680 non-null  float64
 12  GenHlth               253680 non-null  float64
 13  MentHlth              253680 non-null  float64
 14  PhysHlth              253680 non-null  float64
 15  DiffWalk              253680 non-null  float64
 16  Sex                   253680 non-null  float64
 17  Age                   253680 non-null  float64
 18  Education             253680 non-null  float64
 19  Income                253680 non-null  float64
dtypes: float64(20)
memory usage: 38.7 MB
```

Data processing

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0

	Diabetes	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare
0	0.0	0.0	0.0	1.0	0.0	1.0
1	0.0	1.0	0.0	0.0	0.0	0.0
2	0.0	0.0	1.0	0.0	0.0	1.0
3	0.0	1.0	1.0	1.0	0.0	1.0
4	0.0	1.0	1.0	1.0	0.0	1.0

	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education
0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0
1	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0
2	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0
3	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0
4	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0

	Income
0	3.0
1	1.0
2	8.0
3	6.0
4	4.0

	HeartDiseaseorAttack	HighBP	HighChol	BMI	Smoker	Stroke	Diabetes
0	0.0	1.0	1.0	40.0	1.0	0.0	0.0
1	0.0	0.0	0.0	25.0	1.0	0.0	0.0
2	0.0	1.0	1.0	28.0	0.0	0.0	0.0
3	0.0	1.0	0.0	27.0	0.0	0.0	0.0
4	0.0	1.0	1.0	24.0	0.0	0.0	0.0

	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	GenHlth
0	0.0	0.0	1.0	0.0	1.0	5.0
1	1.0	0.0	0.0	0.0	0.0	3.0
2	0.0	1.0	0.0	0.0	1.0	5.0
3	1.0	1.0	1.0	0.0	1.0	2.0
4	1.0	1.0	1.0	0.0	1.0	2.0

	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	-18.0	-15.0	1.0	0.0	9.0	4.0	3.0
1	-0.0	-0.0	0.0	0.0	7.0	6.0	1.0
2	-30.0	-30.0	1.0	0.0	9.0	4.0	8.0
3	-0.0	-0.0	0.0	0.0	11.0	3.0	6.0
4	-3.0	-0.0	0.0	0.0	11.0	5.0	4.0

Methodology

2. Multivariate dataset from the UCI Machine Learning Repository

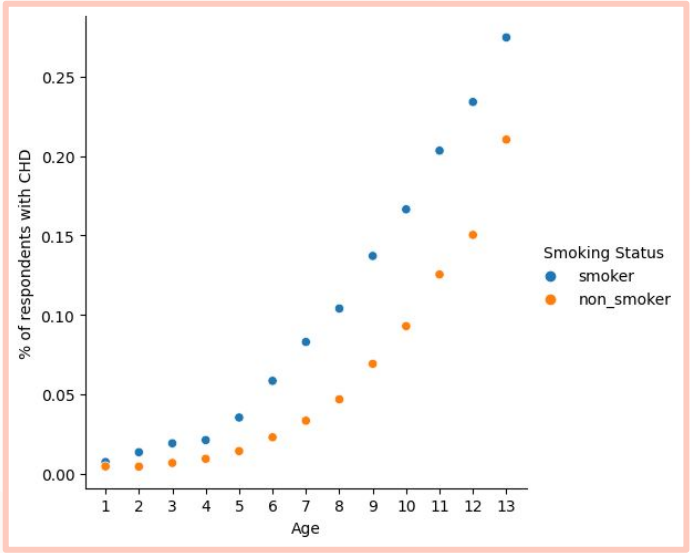
- survey where all respondents have a certain degree of chronic heart disease and related symptoms
- list of complications, outcomes and any other disease that may result from coronary heart diseases

Data analysis and visualization

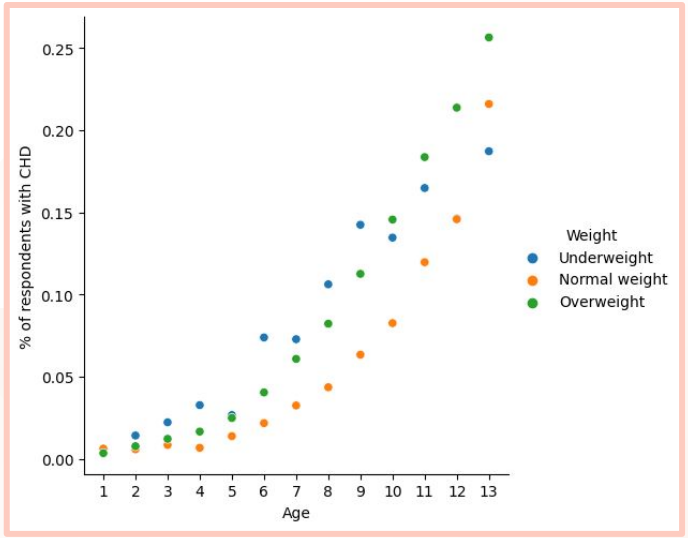
Methodology

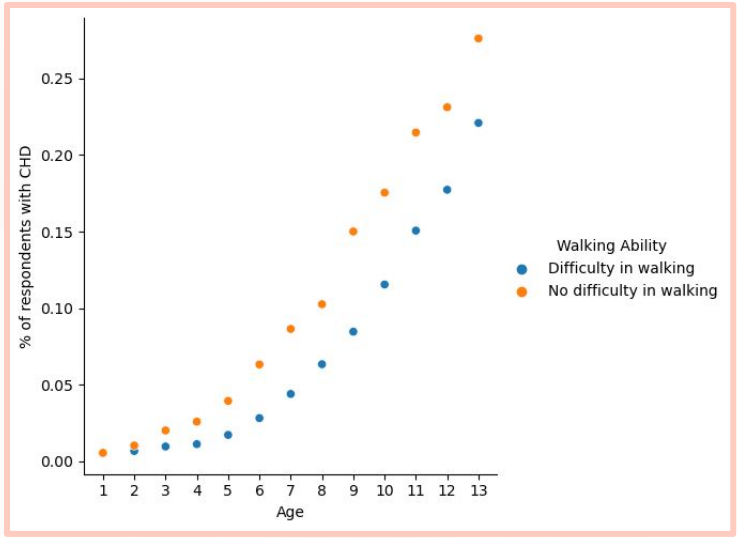
- Plot graphs of prevalence rate of CHD against different age levels
- Each graphs show variation in one risk factor
- Risk factors:
  - Smoking habit
  - Blood pressure level
  - Blood cholesterol level
  - Body Mass Index (BMI)
  - Walking difficulty
  - Physical activity
  - Fruit and Vegetable consumption

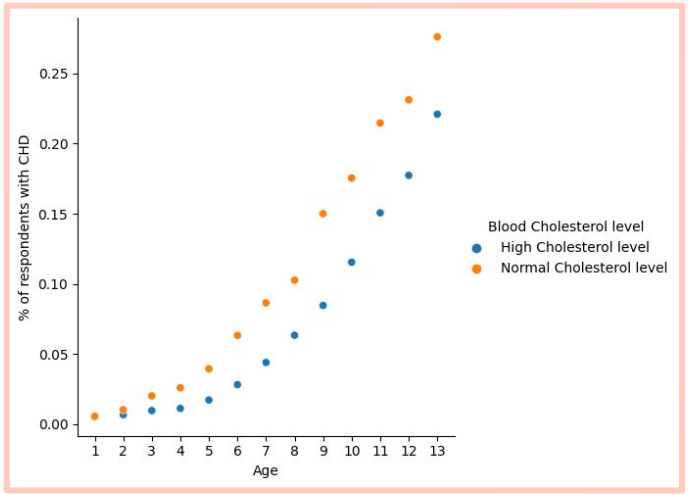
E.g. Variation in Smoking habits

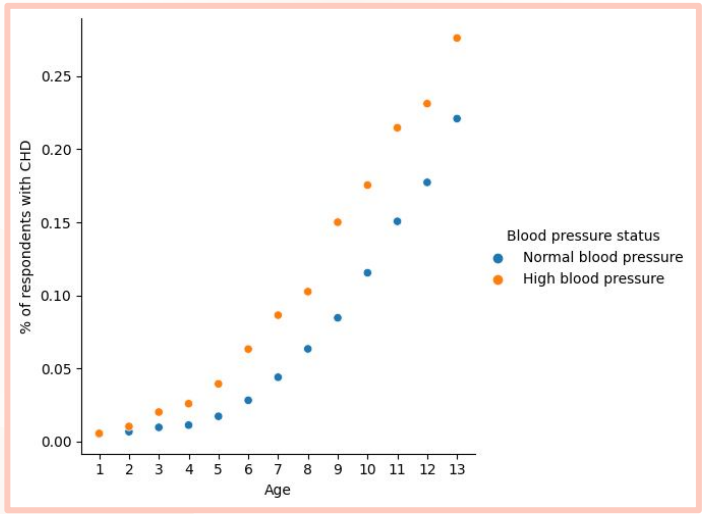


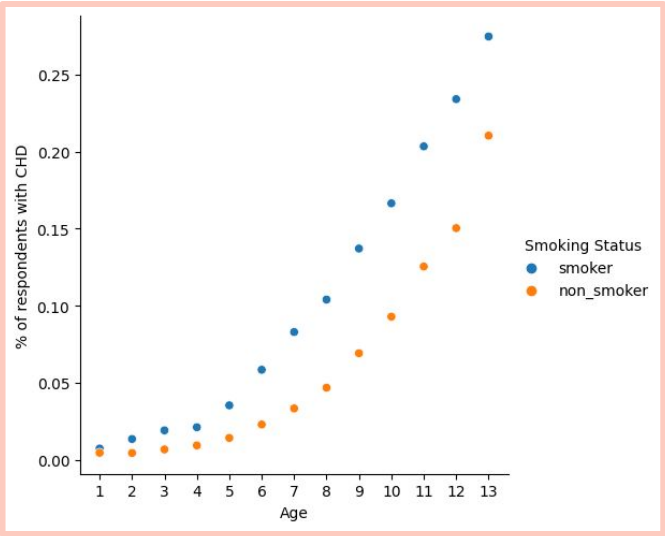
Age Level:
1: 18-24
2: 25-29
3: 30-34
4: 35-39
5: 40-44
6: 45-49
7: 50-54
8: 55-59
9: 60-64
10: 65-69
11: 70-74
12: 75-79
13: >= 80

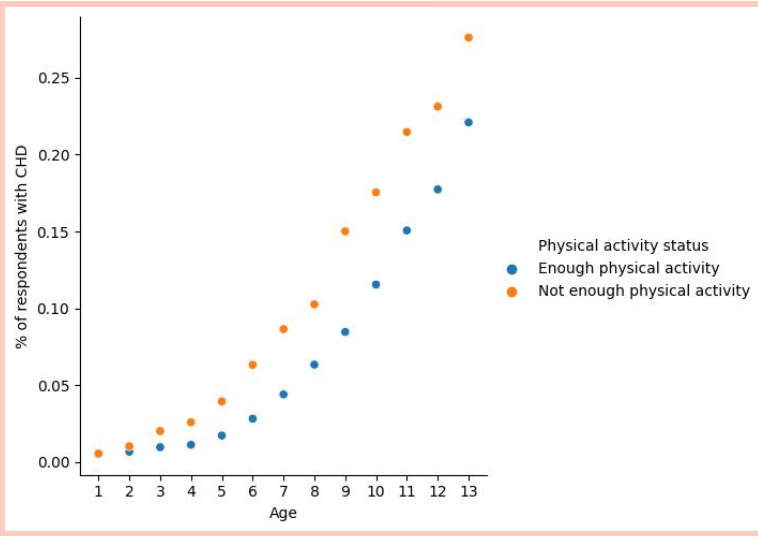

% of respondents with CHD
Age
Weight
Underweight
Normal weight
Overweight

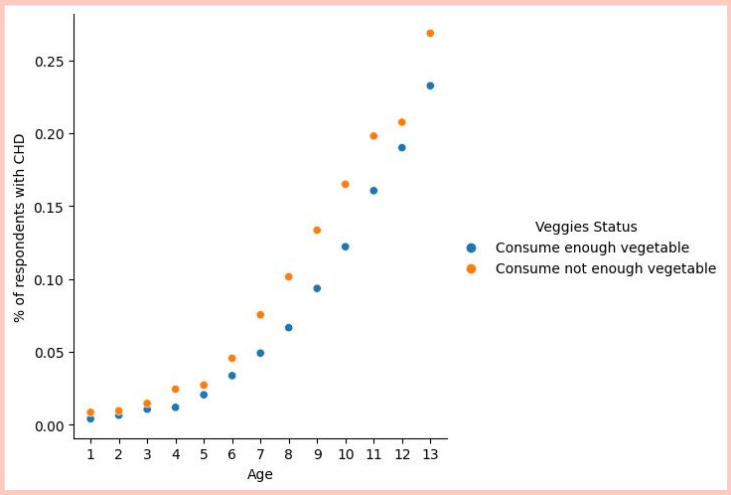

% of respondents with CHD
Age
Walking Ability
Difficulty in walking
No difficulty in walking

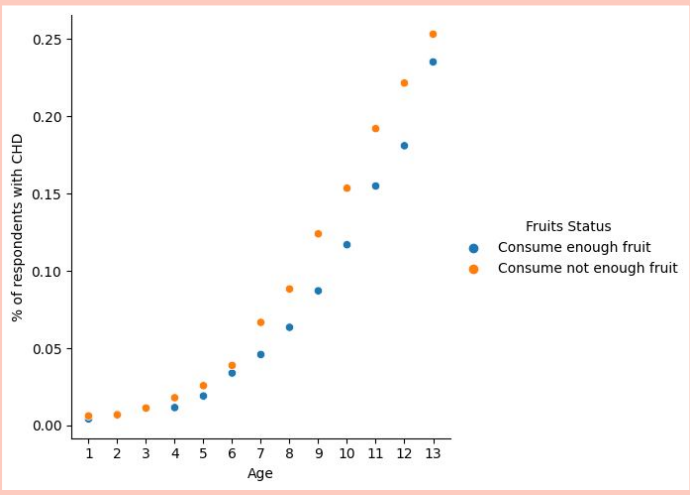

% of respondents with CHD
Age
Blood Cholesterol level
High Cholesterol level
Normal Cholesterol level


% of respondents with CHD
Age
Blood pressure status
Normal blood pressure
High blood pressure


% of respondents with CHD
Age
Smoking Status
smoker
non_smoker


% of respondents with CHD
Age
Physical activity status
Enough physical activity
Not enough physical activity


% of respondents with CHD
Age
Veggies Status
Consume enough vegetable
Consume not enough vegetable


% of respondents with CHD
Age
Fruits Status
Consume enough fruit
Consume not enough fruit

BMI

Difficulty in Walking

Amount of Fruit Consumption

Blood Pressure Level

Blood Cholestrol Level

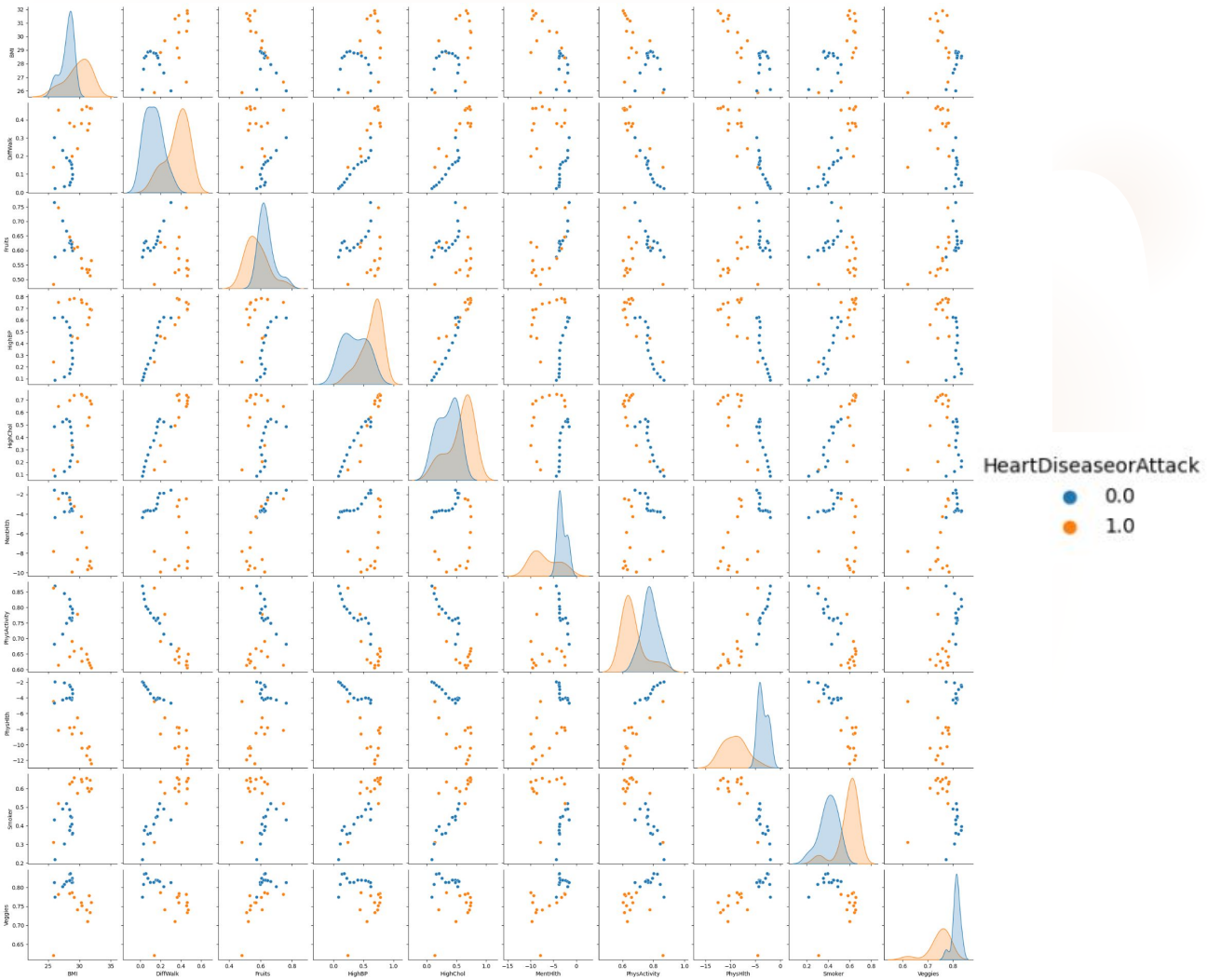
Mental Health Status

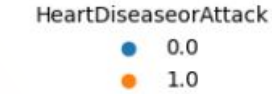
Physical Activity Level

Physical Health Status

Smoking habits

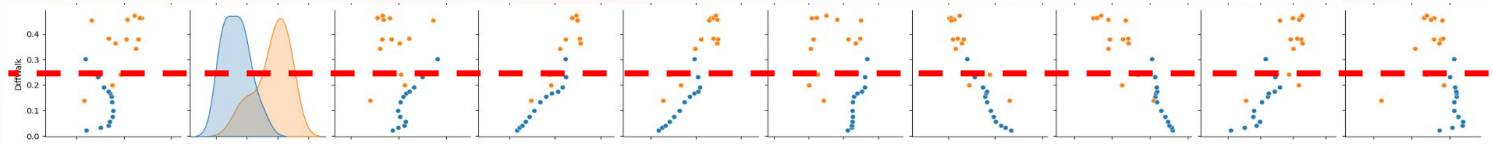
Amount of Vegetable Consumption





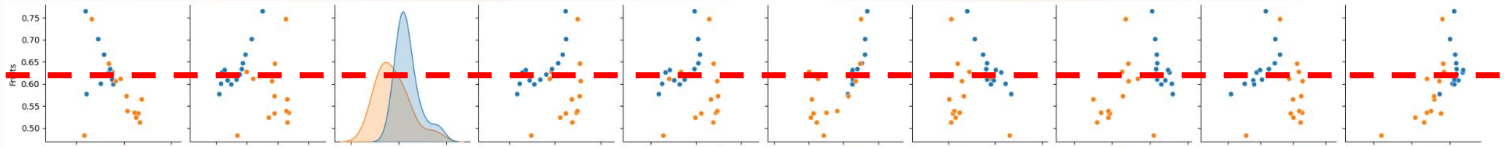
More obvious correlation with CHD

Difficulty in Walking

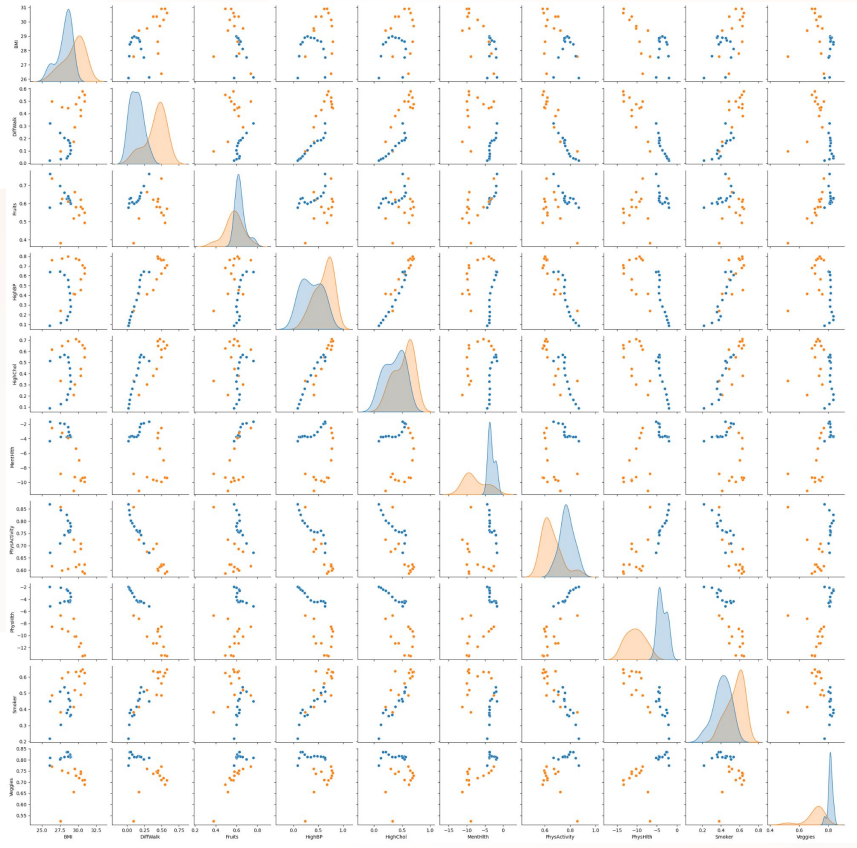


Less obvious correlation with CHD

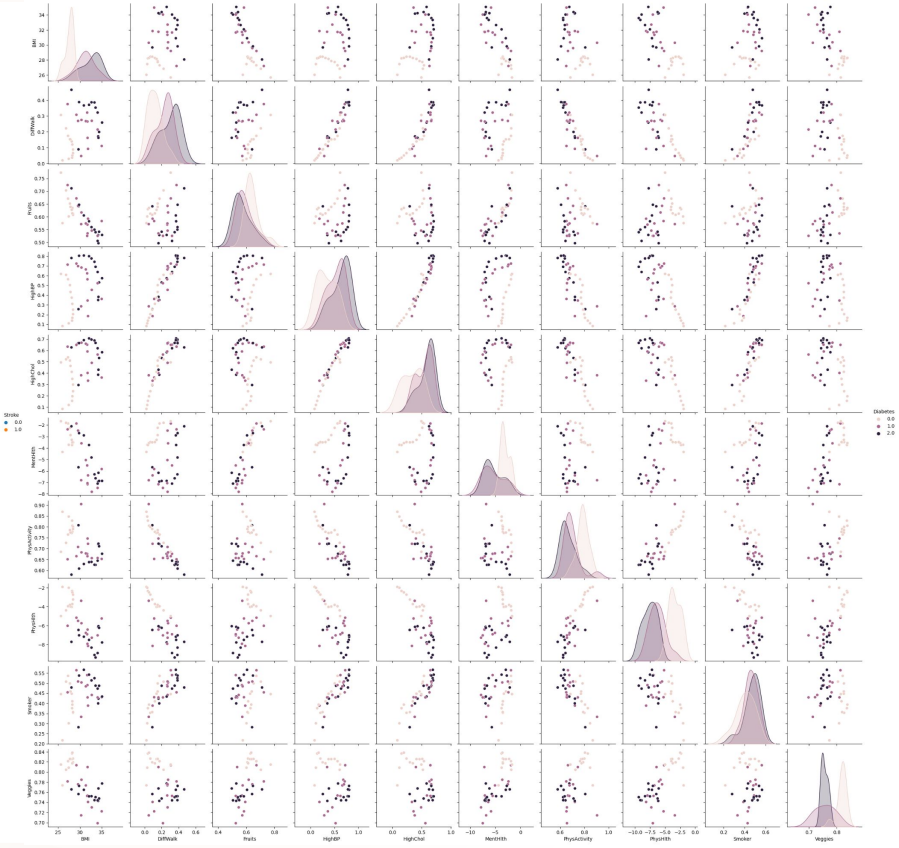
Amount of Fruit Consumption



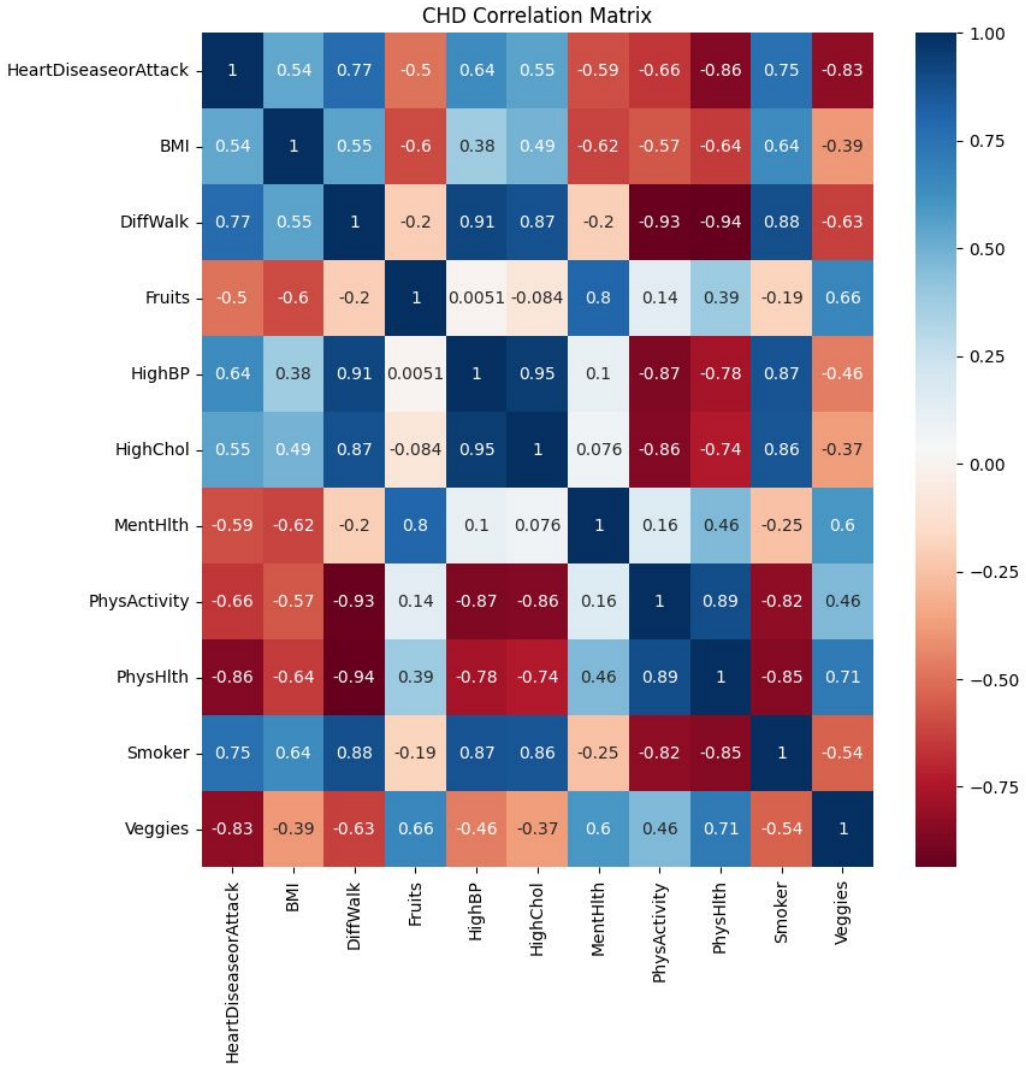
Stroke

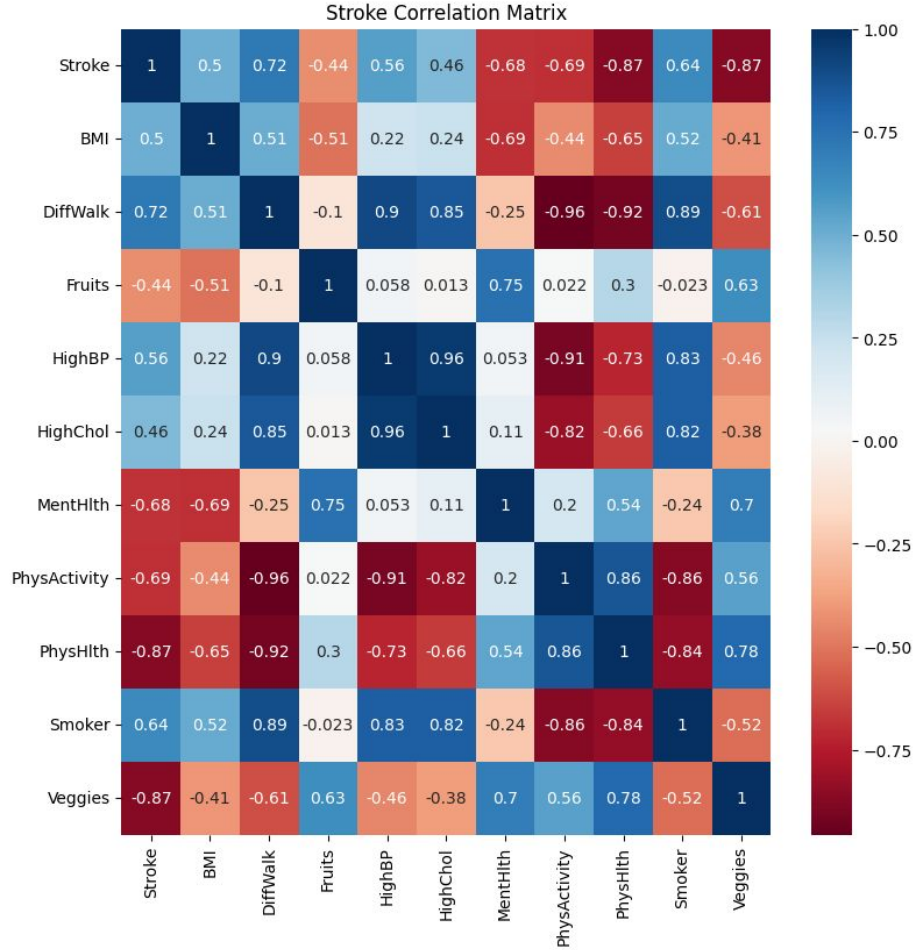


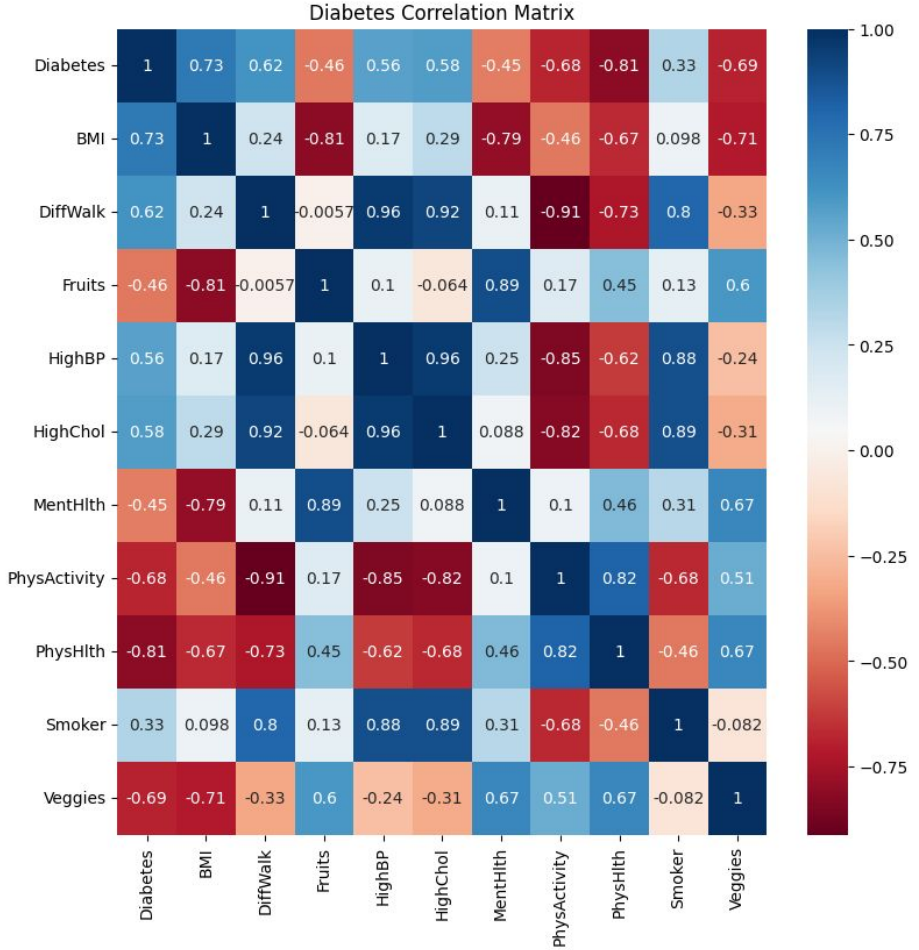
Diabetes

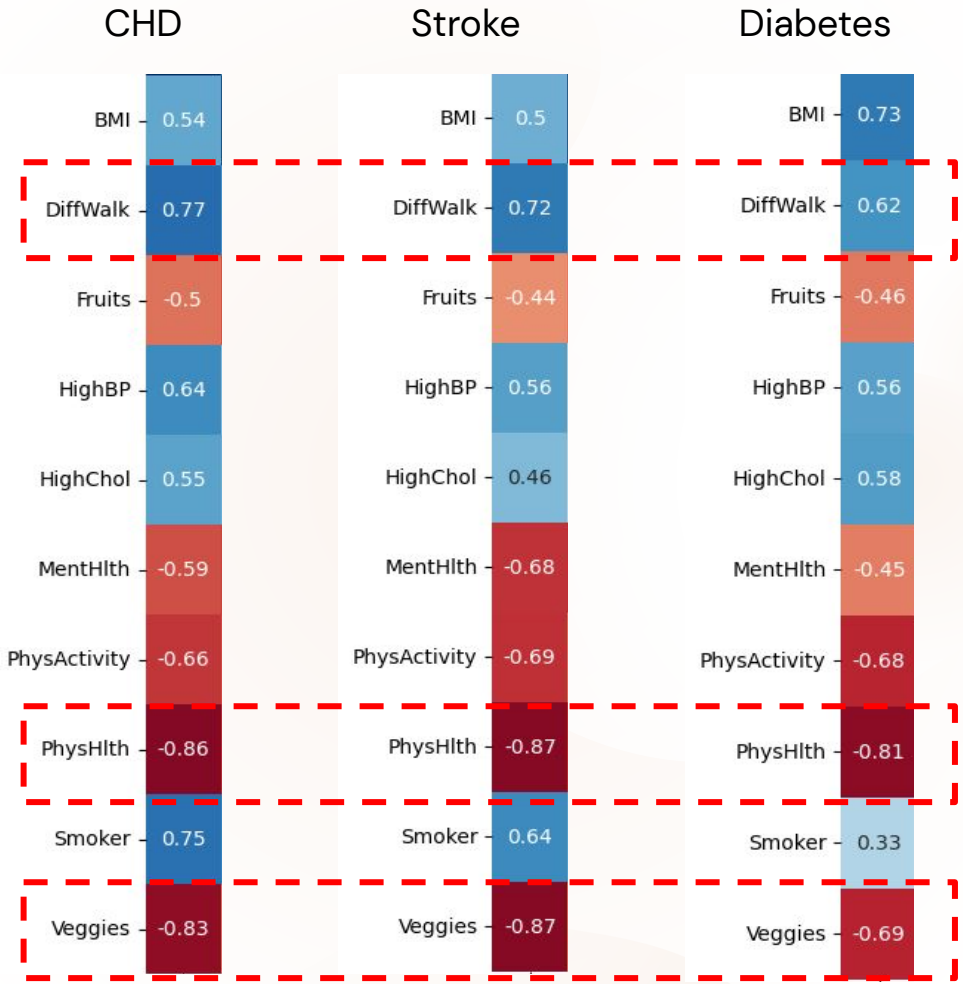


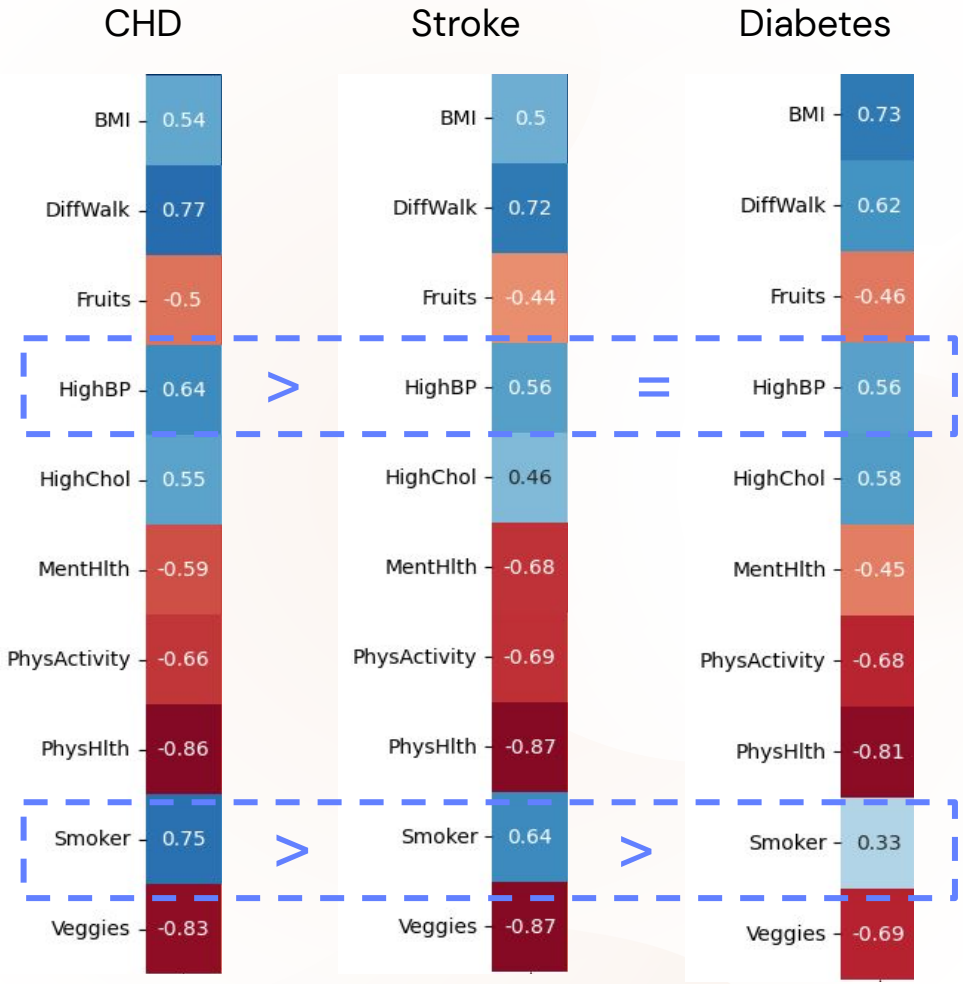
Correlation Analysis


CHD Correlation Matrix
HeartDiseaseorAttack 1 0.54 0.77 -0.5 0.64 0.55 -0.59 -0.66 -0.86 0.75 -0.83
BMI 0.54 1 0.55 -0.6 0.38 0.49 -0.62 -0.57 -0.64 0.64 -0.39
DiffWalk 0.77 0.55 1 -0.2 0.91 0.87 -0.2 -0.93 -0.94 0.88 -0.63
Fruits -0.5 -0.6 -0.2 1 0.0051 -0.084 0.8 0.14 0.39 -0.19 0.66
HighBP 0.64 0.38 0.91 0.0051 1 0.95 0.1 -0.87 -0.78 0.87 -0.46
HighChol 0.55 0.49 0.87 -0.084 0.95 1 0.076 -0.86 -0.74 0.86 -0.37
MentHlth -0.59 -0.62 -0.2 0.8 0.1 0.076 1 0.16 0.46 -0.25 0.6
PhysActivity -0.66 -0.57 -0.93 0.14 -0.87 -0.86 0.16 1 0.89 -0.82 0.46
PhysHlth -0.86 -0.64 -0.94 0.39 -0.78 -0.74 0.46 0.89 1 -0.85 0.71
Smoker 0.75 0.64 0.88 -0.19 0.87 0.86 -0.25 -0.82 -0.85 1 -0.54
Veggies -0.83 -0.39 -0.63 0.66 -0.46 -0.37 0.6 0.46 0.71 -0.54 1
HeartDiseaseorAttack
BMI
DiffWalk
Fruits
HighBP
HighChol
MentHlth
PhysActivity
PhysHlth
Smoker
Veggies
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Stroke Correlation Matrix
Stroke 1 0.5 0.72 -0.44 0.56 0.46 -0.68 -0.69 -0.87 0.64 -0.87
BMI 0.5 1 0.51 -0.51 0.22 0.24 -0.69 -0.44 -0.65 0.52 -0.41
DiffWalk 0.72 0.51 1 -0.1 0.9 0.85 -0.25 -0.96 -0.92 0.89 -0.61
Fruits -0.44 -0.51 -0.1 1 0.058 0.013 0.75 0.022 0.3 -0.023 0.63
HighBP 0.56 0.22 0.9 0.058 1 0.96 0.053 -0.91 -0.73 0.83 -0.46
HighChol 0.46 0.24 0.85 0.013 0.96 1 0.11 -0.82 -0.66 0.82 -0.38
MentHlth -0.68 -0.69 -0.25 0.75 0.053 0.11 1 0.2 0.54 -0.24 0.7
PhysActivity -0.69 -0.44 -0.96 0.022 -0.91 -0.82 0.2 1 0.86 -0.86 0.56
PhysHlth -0.87 -0.65 -0.92 0.3 -0.73 -0.66 0.54 0.86 1 -0.84 0.78
Smoker 0.64 0.52 0.89 -0.023 0.83 0.82 -0.24 -0.86 -0.84 1 -0.52
Veggies -0.87 -0.41 -0.61 0.63 -0.46 -0.38 0.7 0.56 0.78 -0.52 1
Stroke BMI DiffWalk Fruits HighBP HighChol MentHlth PhysActivity PhysHlth Smoker Veggies
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75


Diabetes Correlation Matrix
Diabetes 1 0.73 0.62 -0.46 0.56 0.58 -0.45 -0.68 -0.81 0.33 -0.69
BMI 0.73 1 0.24 -0.81 0.17 0.29 -0.79 -0.46 -0.67 0.098 -0.71
DiffWalk 0.62 0.24 1 -0.0057 0.96 0.92 0.11 -0.91 -0.73 0.8 -0.33
Fruits -0.46 -0.81 -0.0057 1 0.1 -0.064 0.89 0.17 0.45 0.13 0.6
HighBP 0.56 0.17 0.96 0.1 1 0.96 0.25 -0.85 -0.62 0.88 -0.24
HighChol 0.58 0.29 0.92 -0.064 0.96 1 0.088 -0.82 -0.68 0.89 -0.31
MentHlth -0.45 -0.79 0.11 0.89 0.25 0.088 1 0.1 0.46 0.31 0.67
PhysActivity -0.68 -0.46 -0.91 0.17 -0.85 -0.82 0.1 1 0.82 -0.68 0.51
PhysHlth -0.81 -0.67 -0.73 0.45 -0.62 -0.68 0.46 0.82 1 -0.46 0.67
Smoker 0.33 0.098 0.8 0.13 0.88 0.89 0.31 -0.68 -0.46 1 -0.082
Veggies -0.69 -0.71 -0.33 0.6 -0.24 -0.31 0.67 0.51 0.67 -0.082 1
Diabetes BMI DiffWalk Fruits HighBP HighChol MentHlth PhysActivity PhysHlth Smoker Veggies
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75


CHD
BMI 0.54
DiffWalk 0.77
Fruits -0.5
HighBP 0.64
HighChol 0.55
MentHlth -0.59
PhysActivity -0.66
PhysHlth -0.86
Smoker 0.75
Veggies -0.83
Stroke
BMI 0.5
DiffWalk 0.72
Fruits -0.44
HighBP 0.56
HighChol 0.46
MentHlth -0.68
PhysActivity -0.69
PhysHlth -0.87
Smoker 0.64
Veggies -0.87
Diabetes
BMI 0.73
DiffWalk 0.62
Fruits -0.46
HighBP 0.56
HighChol 0.58
MentHlth -0.45
PhysActivity -0.68
PhysHlth -0.81
Smoker 0.33
Veggies -0.69


CHD
Stroke
Diabetes
BMI 0.54
DiffWalk 0.77
Fruits -0.5
HighBP 0.64
HighChol 0.55
MentHlth -0.59
PhysActivity -0.66
PhysHlth -0.86
Smoker 0.75
Veggies -0.83
>
BMI 0.5
DiffWalk 0.72
Fruits -0.44
HighBP 0.56
HighChol 0.46
MentHlth -0.68
PhysActivity -0.69
PhysHlth -0.87
Smoker 0.64
Veggies -0.87
=
BMI 0.73
DiffWalk 0.62
Fruits -0.46
HighBP 0.56
HighChol 0.58
MentHlth -0.45
PhysActivity -0.68
PhysHlth -0.81
Smoker 0.33
Veggies -0.69
>
>
=

Heart Disease Accuracy Prediction

```
log_cols=["Classifier", "Accuracy", "Log Loss"]
log = pd.DataFrame(columns=log_cols)

for cla in algorithm:
    cla.fit(X_train, Y_train)
    name = cla.__class__.__name__
    print("=" * 30)
    print(name)

    train_predictions = cla.predict(X_test)
    acc = accuracy_score(Y_test, train_predictions)
    print("Accuracy: {:.4%}".format(acc))

    train_predictions = cla.predict(X_test)
    ll = log_loss(Y_test, train_predictions)
    print("Log Loss: {}".format(ll))

    log_entry = pd.DataFrame([[name, acc * 100, ll]], columns = log_cols)
    log = log.append(log_entry)
```

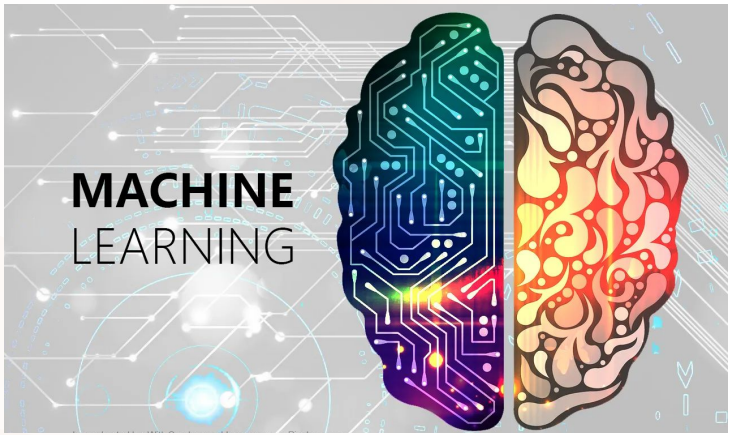
```
==============================
LogisticRegression
****Results****
Accuracy: 90.6398%
Log Loss: 3.373764387119941
==============================
KNeighborsClassifier
****Results****
Accuracy: 89.5676%
Log Loss: 3.760230554016814
==============================
RandomForestClassifier
****Results****
Accuracy: 90.2318%
Log Loss: 3.5208204469501854
==============================
XGBClassifier
****Results****
Accuracy: 90.6871%
Log Loss: 3.356714409168609
==============================
AdaBoostClassifier
****Results****
Accuracy: 90.6851%
Log Loss: 3.3574248249165812
==============================
GradientBoostingClassifier
****Results****
Accuracy: 90.7640%
Log Loss: 3.329008194997694
```

Conclusion

Different Risk Factors

Prevalence Rate of CHD

Positively/Negatively correlated



Reference List

Teboul, A. (2022). Heart Disease Health Indicators Dataset. Kaggle. Retrieved 30 April, 2023, from https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

Pixabey - Wittyspark (n.d.). Retrieved 30 April, 2023, from https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08

Member contributions

- Au Yat Sin Candice: Desktop research, data cleaning and processing
- Chan Chung Yin: Desktop research, data visualization (Correlation Matrix) and Machine Learning Prediction
- Chui Yuen Tsun: Desktop research, data visualization (Scatter plot & Pairplot)


Thank you~