Correlation between Risk Factors and Coronary Heart Disease (CHD)

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

1. Abstract

The topic for investigation is the correlation between risk factors and coronary heart disease (CHD), where we look into the association of CHD and aspects ranging from daily life practices like doing sports, to other further complications like stroke. For the results, we have concluded that there are a few risk factors that are significantly related to the prevalence rate of CHD; the correlation of each risk factor is visualized by a scatter plot. Stroke and diabetes were also investigated as complications of CHD, where they also share similar association patterns with certain risk factors as that of CHD. The extent of correlation is represented by a correlation matrix of those risk factors and CHD, stroke, and diabetes. The dataset we adopted is a survey with around 250,000 responses, which has broad coverage. Accuracy prediction of how comprehensively the results predict the prevalence rate of CHD is at 90.67%.

2. Introduction

Chronic heart diseases have been a prominent issue in society and CHD in particular, is responsible for around 37,000 deaths in 2021 (HealthHK, 2021). It is also the third most common cause of death in Hong Kong (Statista, 2020). Lifestyle diseases share risk factors similar to prolonged exposure to modifiable lifestyle behaviours like smoking, unhealthy diet, and physical inactivity, and they result in the development of chronic diseases, specifically heart disease. The course of the disease in patients with chronic heart disease is different. Complications can occur to different extents that may lead to worsening of the disease and even death. Even an experienced specialist can not always foresee the development of these complications (Golovenkin, S.E. et al., 2020).

Leading questions surrounding the main topic:

(1) How different risk factors affect the prevalence rate of CHD?
(2) Which factors are positively or negatively correlated with CHD?
(3) Do other common complications share any similarities or differences with CHD in the correlations between different risk factors?

3. Methodology

Dataset from the Behavioural Risk Factor Surveillance System (BRFSS), which is an annual survey conducted by the Centers for Disease Control and Prevention (CDC), is used for our investigation. The survey results were uploaded onto the public accessible website Kaggle by author Alex Teboul (Teboul, A., 2022). The survey collects responses related to risk factors such as personal health status, living habits and anamnesis. JupyterLab 3.5.3 is used to perform all data processing in our project. Data preprocessing and cleaning, including dropping missing values, filtering unwanted outliers, and scaling numeric features is first done in order to lead to a more conclusive and accurate decision. We excluded any survey responses which are incomplete and dropped risk factors that are insignificant in correlation representation.

To preliminarily view the relationship between CHD and different risk factors, several graphs of the prevalence rate of CHD against different age levels were plotted (see Appendix A). Each graph involves a change in one risk factor, and the correlation changes due to the variation in the risk factors can be observed by comparing the difference in the trend of the colour dots. The factors we investigated include body mass index (BMI), walking ability, blood pressure, blood cholesterol level, smoking habit, physical activity level, vegetable and fruit consumption. Afterwards, the correlation significance between CHD and different risk factors can be visualized by the graphs.

In view of the problem that the scatter plots mentioned above may be affected by the correlation among the risk factors themselves, pair plots were made to further investigate the strength of correlations between CHD and a particular risk factor when other factors also vary (see Appendix B). The pair plots show the general correlation between any two risk factors, and the difference of these correlations when CHD exists or does not exist, represented by different colours of dots. The strength of the correlation between CHD and one risk factor can be observed by viewing the separation of the colour dots across all the pair plots related to the target risk factor.

Correlation analysis is derived to investigate whether other common complications, such as stroke and diabetes, share similarity or difference with CHD in the correlation with different risk factors. Due to the nonlinearity of the relationships, Spearman's Correlation is used to seek for the interdependence of the risk factors and the diseases. Correlation matrices of different diseases were generated using the calculated Spearman's Correlations, and the similarity and difference can be obtained by comparing the three matrices (see Appendix C).

Machine learning algorithms were also applied to make predictions. The goal is to predict heart disease or attack based on the observation or features from the dataset. We used python, pandas and scikit-learn libraries to perform accuracy prediction. Several machine learning algorithms such as Logistic Regression, K-Nearest Neighbors, Random Forest, XGBoost, AdaBoost, Gradient Boosting were used to predict CHD (see Appendix D). Each algorithm is trained and evaluated by Log Loss and F1 Score.

4. Results

The scatter plots (see Appendix A) obtained prove that all the risk factors show significant correlation with CHD prevalence. The pair plots (see Appendix B) indicate that walking difficulty, physical health status, smoking habit and amount of vegetable consumption have more obvious associations with CHD compared to other risk factors. The correlation matrices (see Appendix C) show that walking ability, physical health status and consumption of vegetables are the risk factors that are significantly and simultaneously correlated to all of these diseases, whereas blood pressure level and smoking habit affect CHD more when compared to the other two complications. The Gradient Boosting Classifier achieved the highest accuracy of 90.76% (see Appendix D), which implies that it is the most effective algorithm for predicting heart disease in this dataset.

In conclusion, different risk factors do have an effect on the prevalence rate of coronary heart disease, with walking difficulty being most positively correlated and physical health being the most negatively associated. Furthermore, stroke and diabetes share a similar correlation as CHD for a few of the risk factors, namely difficulty in walking, physical health and vegetable consumption. Machine learning algorithms applied to the dataset were able to return accurate predictions. The project can be expanded in the future by incorporating additional medical attributes, exploring different algorithms, or applying them to different datasets.

5. Author Contribution

(1) Au Yat Sin Candice 20860834: Desktop research, data cleaning and preprocessing
(2) Chan Chung Yin 20978198: Desktop research, data visualisation (correlation matrix and machine learning prediction)
(3) Chui Yuen Tsun 20864165: Desktop research, data visualization (scatter plot and pairplot)

6. References

Golovenkin, S.E. et al. (2020). 'Myocardial infarction complications Data Set', *UCL Machine Learning Repository,* 9 December. Available at: https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications

HealthyHK. (2021). *Coronary Heart Diseases.* Available at: https://www.healthyhk.gov.hk/phisweb/en/chart_detail/24/

Statista. (2020). *Number of registered deaths in Hong Kong in 2020, by cause of death.* Available at: https://www.statista.com/statistics/1286021/hong-kong-number-of-registered-deaths-by-cause/

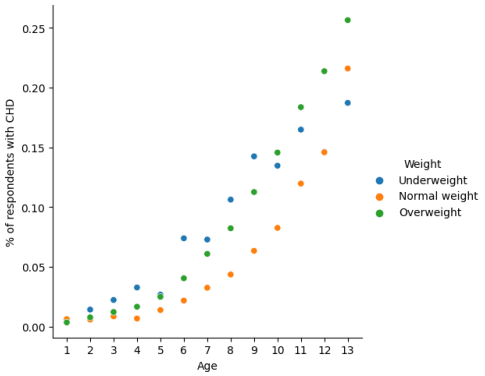
Teboul, A. (2022). 'Heart Disease Health Indicators Dataset', *Kaggle.* Available at: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/versions/3?resource=download

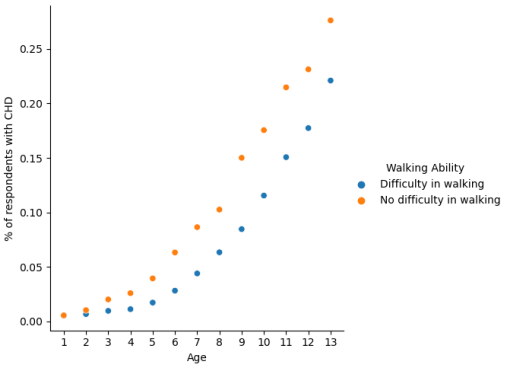
Wu, Daniel. (2019). 'Inspecting a 15 year CDC Chronic Disease Dataset', *Towards Data Science,* 3 March. Available at: https://towardsdatascience.com/inspecting-a-cdc-chronic-disease-dataset-e1685a6b525a

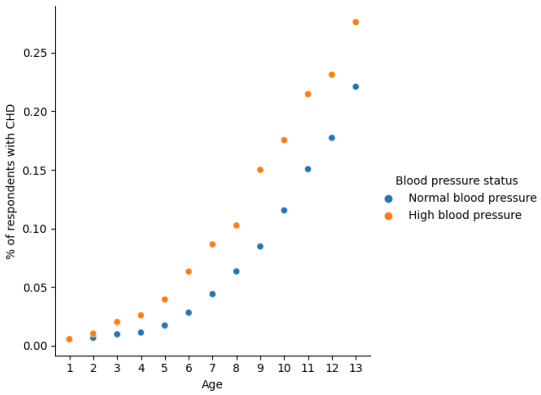
7. Appendixes

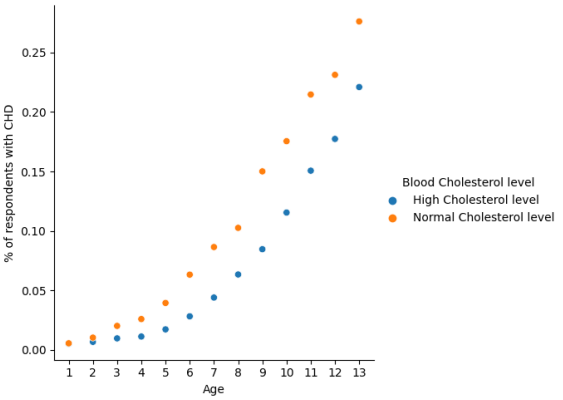
Appendix A

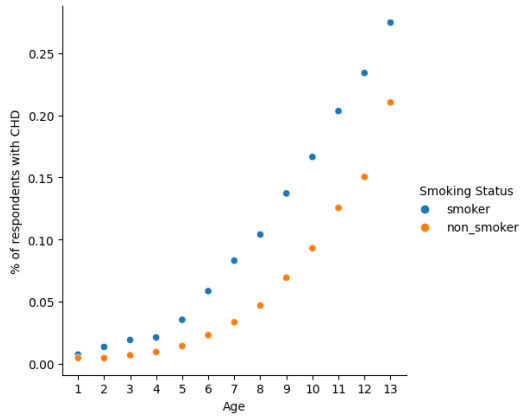
Scatter plots of prevalence rate of CHD against different age levels for eight risk factors.

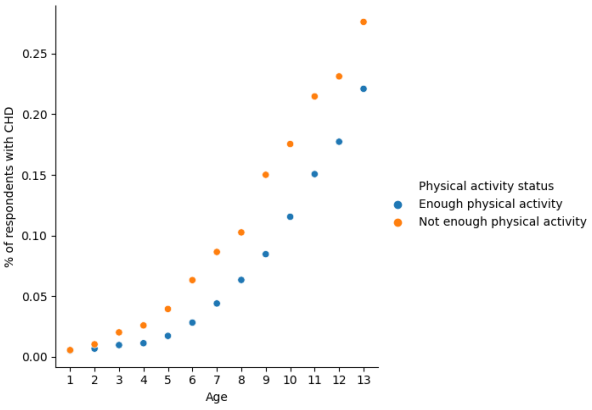


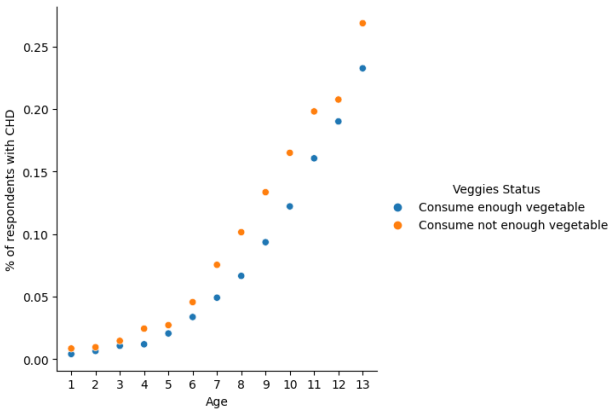


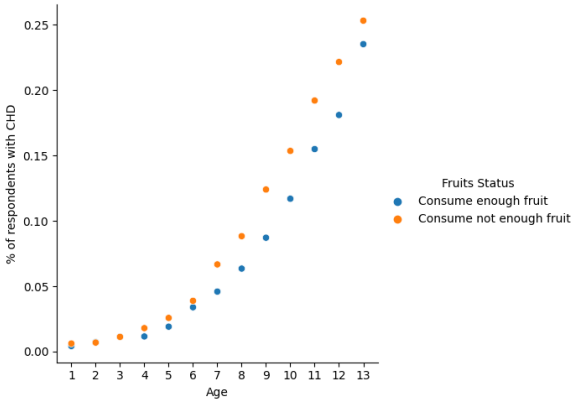










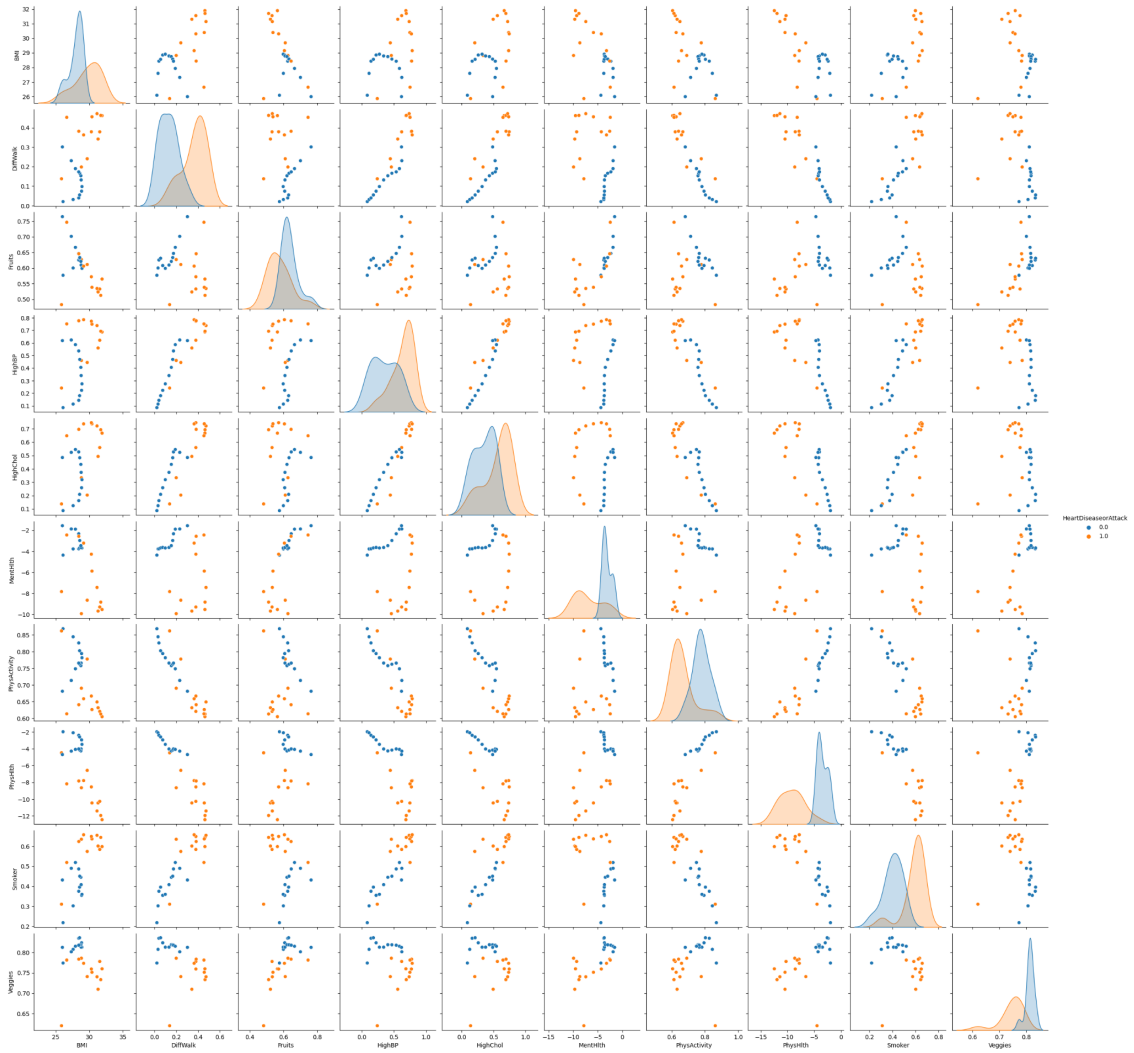


Age levels and their respective age groups:

Age Level	Real Age
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 or older

Appendix B

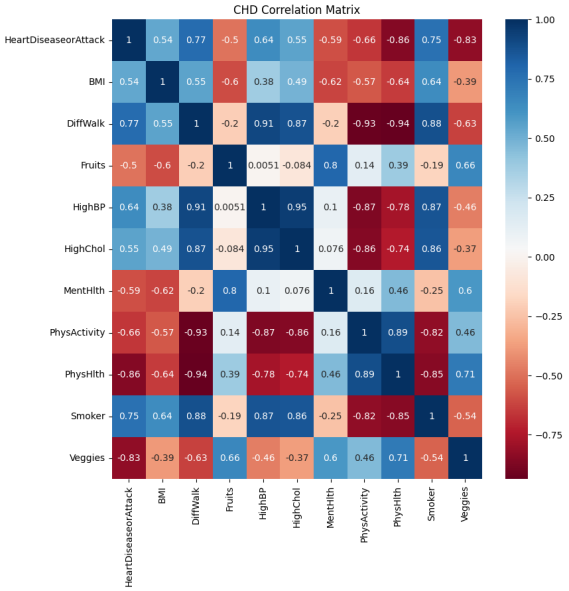
Orange dots represent those with heart disease or attack, whereas blue dots represent those without CHD. The strength of the correlation between CHD and one risk factor can be observed by viewing the separation of the colour dots across all the pair plots related to the target risk factor. For example, there is a more obvious association between difficulty in walking and having CHD, as the two groups of dots of different colours are more separated, which implies that having difficulty in walking indicates a significantly high chance in prevalenting CHD, whereas the correlation between fruit consumption and CHD is more subtle, which represents that consuming more fruit does not necessarily indicate a lower chance in prevalenting CHD.

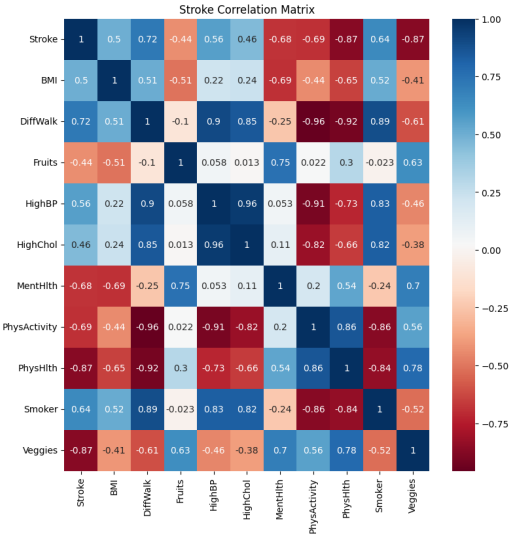


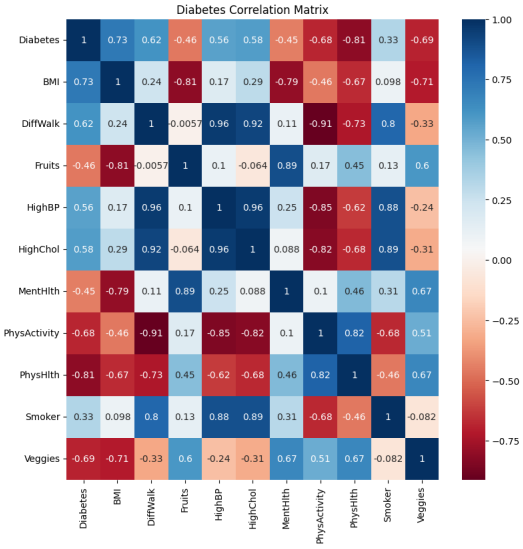
Appendix C

The numbers on the matrix are the Spearman's Correlation Coefficients calculated.

Blue colour indicates a positive correlation and red colour indicates a negative correlation. The larger the magnitude of the number, the stronger the risk factor is positively or negatively correlated with heart disease or attack.







Appendix D

In this section, we describe the machine learning algorithms used in our analysis and present the prediction results, including evaluation metrics used to assess the performance of machine learning predictions.

Data preprocessing:

Before making predictions, we performed some data preprocessing and cleaning, including dropping missing values, filtering unwanted outliers, and scaling numeric features.

Machine Learning Algorithms:

To perform our analysis, we utilised Python, Pandas, and Scikit-learn libraries. Our approach was to employ various machine learning algorithms to predict coronary heart disease. We selected six well-established models namely Logistic Regression, K-Nearest Neighbors, Random Forest, XGBoost, AdaBoost and Gradient Boosting.

Machine Learning techniques allowed us to accurately predict the occurrence of CHD based on the selected risk factors with high accuracy and precision. We believe that our approach has the potential to significantly deepen our understanding of the relationship between the risk factors and CHD. By leveraging the power of the machine learning algorithms, we were able to identify complicated patterns and correlations that may have been missed by traditional statistical methods.

Machine Learning Prediction Results:

Algorithm	Accuracy	Log loss	F1 score
Logistic Regression	90.6457%	3.371633	0.8782
KNeighborsClassifier	89.5676%	3.760230	0.8701
RandomForestClassifier	90.1845%	3.537870	0.8748
XGBClassifier	90.6871%	3.356714	0.8780
AdaBoostClassifier	90.6851%	3.357424	0.8815
GradientBoostingClassifier	90.7640%	3.329008	0.8782

Steps of how the accuracy prediction was performed:

```
algorithm = [
    LogisticRegression(),
    KNeighborsClassifier(),
    RandomForestClassifier(),
    XGBClassifier(),
    AdaBoostClassifier(),
    GradientBoostingClassifier(),
]
```

```
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)
    print('****Results****')

    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))


    f1score = f1_score(Y_test, train_predictions, average='weighted')
    print("F1 score: {}".format(f1score))

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

print("=" * 30)
```

```
==============================
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
```