**Predicting the outcome of 10-year risk of coronary heart disease from Heart Disease Study**

**Background**: As infectious diseases gradually be eradicated from the U.S., chronic diseases become the top killers. According to CDC, coronary heart disease (CHD) is the leading cause of death for any gender and any race/ethnic group (Centers for Disease Control and Prevention, 2022). How to use certain medical parameters to predict the probability that a person might develop coronary heart disease becomes important. Chronic disease like heart disease will impose huge burden on the medical system and the patient himself/herself have to suffer from life-long treatment and medication. The best way is to predict the probability and advise potential patients to change their life habits to prevent the heart disease from even occurring. Therefore, a model to predict the 10-year risk of coronary heart disease is developed for this project. The data is from a public database which includes more than 6000 records and 15 medical parameters, such as total cholesterol level, age, heart rate, etc. In this study, there are 644 out of 4240 participants who are diagnosed as having CHD for 10 years, which accounts for 15.2% of the study population. Figure 1 shows that project overview.

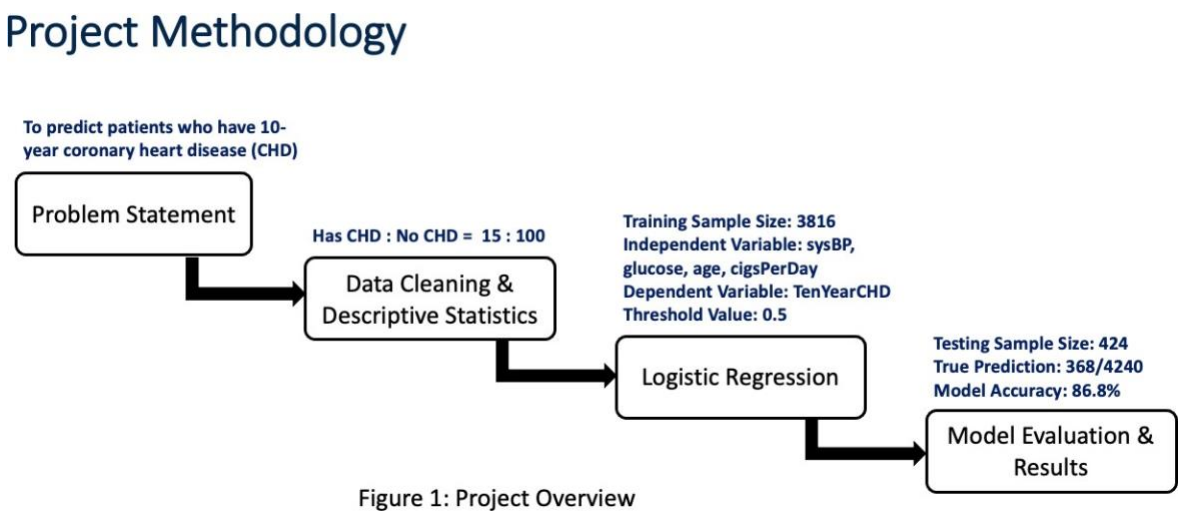

Figure 1: Project Overview

**Methods**: I used Visual Studio Code to develop my model. Packages such as pandas, numpy, sklearn, statsmodel are imported. The dataset was already cleaned with no missing values as part of the class material that we could download from Canvas.

I started off the project with using Exploratory Data Analysis (EDA) techniques to have a better understanding of the data. The standard deviation, mean, min and max values are calculated to see which variable has more spread-out data points. Correlations were performed to see if any two variables have high collinearity. If they had a correlation coefficient greater than 0.75, then only of them could be included in the model because they might represent the same thing.

Because the dependent variable 10-year risk of CHD is a categorical variable, 1 means 'yes' and 0 means 'no', the binomial logistics regression is used to develop the model. By combining research on what factors increase the risk of having CHD and the prediction accuracy of different models from binomial logistics regression analyses, I chose variables systolic blood pressure, glucose, age, and cigarettes per day. For high systolic blood pressure, it will thicken the arteries and narrow the channel that blood can flow through. High blood pressure also increases the risk of stroke, which is a strong predictor of having heart disease (Gosmanova et al., 2016). If a person has high glucose in the blood, it may damage the blood vessels and nerves and therefore lead to increased risk of CHD (Fuller et al., 1980). For cigarettes, they lead to the formation of plaque in the blood vessels which narrow the blood vessels (Willett et al., 1987). Last but not the least, studies have been showing that people who are 65 or older will have a higher probability to have CHD, so age is included as an independent risk factor (Rodgers et al., 2019). Then I used machine learning to split the whole dataset to two subsets: train and test, 90% and 10%, respectively. Training sample included 3816 records, and testing sample included 424 records.

My first model tested all variables one by one so I can have a basic understanding of which variables better predict 10-year risk of CHD and which do not. The systolic blood pressure variable – sysBP and glucose had the highest prediction rate, so they were selected. Then variables age, sex, current smoker or not, and cigarettes per day are tested with the combination of sysBP and glucose. I chose my third variable cigarettes per day. Next, variables age and sex are tested with the combination of sysBP, glucose, and cigarettes per day. Sex was excluded because it decreased the prediction rate of the whole model, and the other variables are kept. The results of these regressions can be found in Tables 1-4 in Appendix A.

**Results/Analysis**: The first model of the relationship between 'sysBP' and 'glucose' and 'TenYearCHD' has a prediction accuracy 85.8%. The final model that examines the relationship between 'sysBP', 'glucose', 'age', and 'cigsPerDay'and 'TenYearCHD' has a prediction accuracy 86.8%. This is expected because only 15.2% of the study population have 10-year risk CHD, and there are not enough data of people who have 10-year risk of CHD. For the 4 variables chosen, their p values are all <.05, which means they are all statistically significant in predicting the risk of 10-year CHD. I also calculated true prediction, false prediction, and prediction accuracy for all three models. For the first model, the true prediction is 364; the second model has 365 true predictions; the third model has 368 true predictions. However, most of the predictions are about people who do not have 10-year risk CHD. For people who do have 10-year risk CHD, model 1 successfully predicted 2 records (see code line 11). After adding variables, model 3 successfully predicted 6 records out of 644 in total (see code 12).

**Conclusion**: In conclusion, the model that uses systolic blood pressure, glucose, age, and cigarettes can achieve 86.8% prediction accuracy. Among the records that it successfully predicts, most of them are patients who do not have 10-year risk CHD. It is mainly because only 15.2% of the study population have 10-year risk CHD. To achieve a better prediction accuracy, at least 50% of the study population should be people who have 10-year risk of CHD, and there should be a larger dataset. This study is important because even though we know many factors that are associated with coronary heart disease, we are still not clear which are more important and which are less important. This model allows us to understand crucial medical parameters that better predict CHD, which could help prevent life threats like heart strokes (Vijayalakshmi & Muruganand, 2021).

## Appendix A

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | TenYearCHD | No. Observations: | 3472 |
| Model: | GLM | Df Residuals: | 3469 |
| Model Family: | Binomial | Df Model: | 2 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -1405.7 |
| Date: | Thu, 24 Feb 2022 | Deviance: | 2811.4 |
| Time: | 23:39:21 | Pearson chi2: | 3.46e+03 |
| No. Iterations: | 5 | Pseudo R-squ. (CS): | 0.05039 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -5.5895 | 0.309 | -18.093 | 0.000 | -6.195 | -4.984 |
| sysBP | 0.0237 | 0.002 | 11.910 | 0.000 | 0.020 | 0.028 |
| glucose | 0.0078 | 0.002 | 4.196 | 0.000 | 0.004 | 0.011 |

**Table 1.** Crude Logistic Regression model predicting the outcome of 10-year CHD by sysBP and glucose

Generalized Linear Model Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | TenYearCHD | No. Observations: | 3449 |
| Model: | GLM | Df Residuals: | 3445 |
| Model Family: | Binomial | Df Model: | 3 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -1387.6 |
| Date: | Thu, 24 Feb 2022 | Deviance: | 2775.1 |
| Time: | 23:39:21 | Pearson chi2: | 3.43e+03 |
| No. Iterations: | 5 | Pseudo R-squ. (CS): | 0.05668 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -5.9468 | 0.323 | -18.432 | 0.000 | -6.579 | -5.314 |
| sysBP | 0.0247 | 0.002 | 12.215 | 0.000 | 0.021 | 0.029 |
| glucose | 0.0084 | 0.002 | 4.472 | 0.000 | 0.005 | 0.012 |
| cigsPerDay | 0.0187 | 0.004 | 4.859 | 0.000 | 0.011 | 0.026 |

**Table 2.** Crude Logistic Regression model predicting the outcome of 10-year CHD by sysBP glucose, and cigsPerDay

| Generalized Linear Model Regression Results | | | |
|---|---|---|---|
| Dep. Variable: | TenYearCHD | No. Observations: | 3449 |
| Model: | GLM | Df Residuals: | 3444 |
| Model Family: | Binomial | Df Model: | 4 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -1327.3 |
| Date: | Thu, 24 Feb 2022 | Deviance: | 2654.6 |
| Time: | 23:39:22 | Pearson chi2: | 3.39e+03 |
| No. Iterations: | 5 | Pseudo R-squ. (CS): | 0.08907 |
| Covariance Type: | nonrobust | | |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | -8.5080 | 0.425 | -20.010 | 0.000 | -9.341 | -7.675 |
| sysBP | 0.0165 | 0.002 | 7.681 | 0.000 | 0.012 | 0.021 |
| glucose | 0.0079 | 0.002 | 4.218 | 0.000 | 0.004 | 0.012 |
| age | 0.0702 | 0.007 | 10.723 | 0.000 | 0.057 | 0.083 |
| cigsPerDay | 0.0281 | 0.004 | 6.970 | 0.000 | 0.020 | 0.036 |

**Table 3.** Crude Logistic Regression model predicting the outcome of 10-year CHD by sysBP glucose, cigsPerDay, and age

## References


Centers for Disease Control and Prevention. (2022, February 7). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved February 25, 2022, from https://www.cdc.gov/heartdisease/facts.htm

Fuller, J. H., Shipley, M. J., Rose, G., Jarrett, R. J., & Keen, H. (1980). Coronary-heart-disease risk and impaired glucose tolerance the whitehall study. *The Lancet*, *315*(8183), 1373–1376. https://doi.org/10.1016/s0140-6736(80)92651-3

Gosmanova, E. O., Mikkelsen, M. K., Molnar, M. Z., Lu, J. L., Yessayan, L. T., Kalantar-Zadeh, K., & Kovesdy, C. P. (2016). Association of systolic blood pressure variability with mortality, coronary heart disease, stroke, and renal disease. Journal of the American College of Cardiology, 68(13), 1375–1386. https://doi.org/10.1016/j.jacc.2016.06.054

Rodgers, J. L., Jones, J., Bolleddu, S. I., Vanthenapalli, S., Rodgers, L. E., Shah, K., Karia, K., & Panguluri, S. K. (2019). Cardiovascular risks associated with gender and aging. *Journal of Cardiovascular Development and Disease*, *6*(2), 19. https://doi.org/10.3390/jcdd6020019

Vijayalakshmi, S. R., & Muruganand, S. (2021). Medical internet of health things (miot). *Securing IoT in Industry 4.0 Applications with Blockchain*, 81–113. https://doi.org/10.1201/9781003175872-4

Willett, W. C., Green, A., Stampfer, M. J., Speizer, F. E., Colditz, G. A., Rosner, B., Monson, R. R., Stason, W., & Hennekens, C. H. (1987). Relative and absolute excess risks of coronary heart disease among women who smoke cigarettes. New England Journal of Medicine, 317(21), 1303–1309. https://doi.org/10.1056/nejm198711193172102