Cardiovascular Disease Report

Valerio Martinez, Cruz Salas, Ryan Nguyen, Simrat Clair, Alan Johnson

2024-04-28

Introduction

Our project involves analyzing the Cardiovascular Disease dataset found on Kaggle to determine which factors significantly contribute to the development of cardiovascular disease in an individual. We were inspired by lecture 9, which introduced logistic regression using the BreastCancer dataset to analyze attributes of cells that imply whether it is benign or malignant. The dataset contains exactly 70,000 observations, in which each observation is associated with 13 variables.

Input Descriptions:

- age: Shown in days.
- height: Shown in cm.
- weight: Shown in float kg.
- gender: Shown as a categorical code.
- ap_hi: Systolic blood pressure.
- ap_lo: Diastolic blood pressure.
- cholesterol: Shown in three different ways: 1: normal cholesterol, 2: above normal cholesterol, and 3: well above cholesterol.
- gluc: Shown in three different ways: gluc1: normal glucose, gluc2: above normal glucose, and gluc3: well above normal glucose.
- smoke: Binary classification variable: smoke0: non-smoker, smoke1: smoker.
- alco: Binary classification variable: alco0: low alcohol intake, alco1: high alcohol intake.
- active: Binary classification variable: active0: low physical activity, active1: regular physical activity.

Output Description:

- cardio: Indicates the presence or absence of cardiovascular disease.

Research Question

How do lifestyle factors correlate with the risk of developing cardiovascular disease?

Methods

Logistic Model (Ryan Nguyen, Alan Johnson)

We chose a logistic model because our response variable, cardio, is categorial (0 or 1). Also, the logistic model can handle large datasets efficiently, making it suitable for our dataset which contains 70,000 observations.

Some disadvantages of the logistic model is that it can be sensitive to outliers, which may disproportionately influence the model's coefficients and predictions. Also, if the relationship

between predictors and the log-odds of the outcome is highly non-linear or complex, logistic model may not capture it effectively.

Model Formula

$$P(\text{cardio} = 1|X) = \frac{\begin{array}{l}\exp(\beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{height} + \beta_3 \times \text{weight} + \beta_4 \times \text{gender} + \beta_5 \times \text{ap\_hi} \\ + \beta_6 \times \text{ap\_lo} + \beta_7 \times \text{cholesterol} + \beta_8 \times \text{gluc} + \beta_9 \times \text{smoke} + \beta_{10} \times \text{alco} \\ + \beta_{11} \times \text{active})\end{array}}{\begin{array}{l}1 + \exp(\beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{height} + \beta_3 \times \text{weight} + \beta_4 \times \text{gender} + \beta_5 \times \text{ap\_hi} \\ + \beta_6 \times \text{ap\_lo} + \beta_7 \times \text{cholesterol} + \beta_8 \times \text{gluc} + \beta_9 \times \text{smoke} + \beta_{10} \times \text{alco} \\ + \beta_{11} \times \text{active})\end{array}} \tag{1}$$

Model 1 - Include all predictors

```
library(readr)
cardio.data <- read_delim("cardio_train.csv",
                          delim = ";", escape_double = FALSE, trim_ws = TRUE)

cardio.data$gender = as.factor(cardio.data$gender)
cardio.data$cholesterol = as.factor(cardio.data$cholesterol)
cardio.data$gluc = as.factor(cardio.data$gluc)
cardio.data$smoke = as.factor(cardio.data$smoke)
cardio.data$alco = as.factor(cardio.data$alco)
cardio.data$active = as.factor(cardio.data$active)
cardio.data$cardio = as.factor(cardio.data$cardio)

heart.logistic1 = glm(cardio ~ . - id, family = "binomial",
                      data = cardio.data)
```

```
summary(heart.logistic1)
```

```
Call:
glm(formula = cardio ~ . - id, family = "binomial", data = cardio.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.9635  -0.0980   0.9907   4.6621
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.084e+00  2.213e-01 -36.535  < 2e-16 ***
age           1.485e-04  3.557e-06  41.735  < 2e-16 ***
gender2       1.430e-02  2.107e-02   0.679    0.497
height       -5.626e-03  1.232e-03  -4.567 4.95e-06 ***
weight        1.521e-02  6.607e-04  23.023  < 2e-16 ***
ap_hi         3.951e-02  6.057e-04  65.235  < 2e-16 ***
ap_lo         3.004e-04  6.735e-05   4.460 8.18e-06 ***
cholesterol2  4.222e-01  2.593e-02  16.285  < 2e-16 ***
cholesterol3  1.134e+00  3.444e-02  32.929  < 2e-16 ***
gluc2         3.011e-02  3.438e-02   0.876    0.381
gluc3        -3.387e-01  3.809e-02  -8.894  < 2e-16 ***
smoke1       -1.314e-01  3.320e-02  -3.958 7.57e-05 ***
alco1        -1.695e-01  4.026e-02  -4.211 2.54e-05 ***
active1      -2.101e-01  2.105e-02  -9.981  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 97041  on 69999  degrees of freedom
Residual deviance: 80883  on 69986  degrees of freedom
AIC: 80911

Number of Fisher Scoring iterations: 25
```

Predictors with p-values less than 0.05 are considered significant. For example, age, height, weight, blood pressure, cholesterol levels, smoking status, alcohol consumption, and physical activity are all significant predictors. Gender, glucose level don't appear to be significant predictors as their p-values exceed chosen significance level.

Model 2 - Only include statistically significant predictors

We will use the significant predictors found from heart.logistic1 to to create Model 2. This means all predictors except for id, gender, and glucose will be included.

```
heart.logistic2 = glm(cardio ~ age+height+weight+ap_hi+ap_lo+cholesterol+smoke+alco+act
                      , family = "binomial",
                    data = cardio.data)
```

We next applied the step() function on the initial model (heart.logistic1) to perform stepwise regression based on the Akaike Information Criterion (AIC). The stepwise process sequentially evaluates each predictor's contribution to the model and removes predictors that do not significantly improve the model fit, based on AIC.

```
step(heart.logistic1)
```

```
Start:  AIC=80910.62
cardio ~ (id + age + gender + height + weight + ap_hi + ap_lo +
    cholesterol + gluc + smoke + alco + active) - id

              Df Deviance   AIC
- gender       1    80882 80908
<none>              80883 80911
- smoke        1    80900 80926
- alco         1    80902 80928
- ap_lo        1    80902 80928
- height       1    80917 80943
- gluc         2    80960 80984
- active       1    80987 81013
- weight       1    81248 81274
- cholesterol  2    82098 82122
- age          1    82467 82493
- ap_hi        1    87965 87991

Step:  AIC=80907.87
cardio ~ age + height + weight + ap_hi + ap_lo + cholesterol +
    gluc + smoke + alco + active

              Df Deviance   AIC
<none>              80882 80908
- ap_lo        1    80901 80925
- smoke        1    80901 80925
- alco         1    80901 80925
- height       1    80921 80945
- gluc         2    80984 81006
- active       1    80986 81010
- weight       1    81247 81271
- cholesterol  2    82099 82121
- age          1    82466 82490
- ap_hi        1    87982 88006
```

```
Call:  glm(formula = cardio ~ age + height + weight + ap_hi + ap_lo +
    cholesterol + gluc + smoke + alco + active, family = "binomial",
    data = cardio.data)

Coefficients:
 (Intercept)           age        height        weight         ap_hi
  -8.1440960     0.0001485    -0.0052551     0.0152121     0.0395279
       ap_lo  cholesterol2  cholesterol3         gluc2         gluc3
   0.0003007     0.4217513     1.1337397     0.0300489    -0.3389234
      smoke1         alco1       active1
  -0.1256165    -0.1681051    -0.2100651

Degrees of Freedom: 69999 Total (i.e. Null);  69987 Residual
Null Deviance:      97040
Residual Deviance: 80880    AIC: 80910
```

Model 3 - Using predictors from stepwise regression

After the stepwise regression process, the final selected predictors were age, height, weight, ap_hi (systolic blood pressure), ap_lo (diastolic blood pressure), cholesterol, gluc, smoke, alco (alcohol consumption), and active (physical activity). The coefficients for these predictors are provided in the summary output of the final model (heart.logistic3).

The coefficients in the final model represent the log odds of the outcome (cardiovascular disease) associated with each unit change in the predictor, assuming other predictors are constant. For example, a positive coefficient for a predictor indicates an increase in the probability of cardiovascular disease with an increase in that predictor, while a negative coefficient indicates a decrease in the probability.

```
heart.logistic3 = glm(formula = cardio ~ age + height + weight + ap_hi + ap_lo +
    cholesterol + gluc + smoke + alco + active, family = "binomial",
    data = cardio.data)
```

Determining Best Model

```
            Model Null_Deviance Residual_Deviance R_Squared      AIC      BIC
1 heart.logistic1      97040.58          80882.62 0.1665072 80910.62 81038.81
2 heart.logistic2      97040.58          80983.71 0.1654656 81005.71 81106.43
3 heart.logistic3      97040.58          80881.87 0.1665149 80907.87 81026.91
```

It is evident that Model 3 is the best model out of all the models because the AIC and BIC is the lowest. The AIC of Model 3 is 80907.87 while the AIC of Model 1 is 80910.62 and the AIC

of Model 2 is 81005.71. We will further test Model 3 (Best Model) on training and validation tests.

Final Equation for Logistic Model

$$P(\text{cardio} = 1|X) = \frac{\begin{array}{l}\exp(-8.188 + 0.0001485 \times \text{age} + -0.005255 \times \text{height} + 0.01521 \times \text{weight} \\ + 0.03953 \times \text{ap\_hi} + 0.0003007 \times \text{ap\_lo} + 0.4218 \times \text{cholesterol2} \\ + 1.134 \times \text{cholesterol3} + 0.03005 \times \text{gluc2} + -0.3389 \times \text{gluc3} \\ + -0.1256 \times \text{smoke1} + -0.1681 \times \text{alco1} + -0.2101 \times \text{active1})\end{array}}{\begin{array}{l}1 + \exp(-8.188 + 0.0001485 \times \text{age} + -0.005255 \times \text{height} + 0.01521 \times \text{weight} \\ + 0.03953 \times \text{ap\_hi} + 0.0003007 \times \text{ap\_lo} + 0.4218 \times \text{cholesterol2} \\ + 1.134 \times \text{cholesterol3} + 0.03005 \times \text{gluc2} + -0.3389 \times \text{gluc3} \\ + -0.1256 \times \text{smoke1} + -0.1681 \times \text{alco1} + -0.2101 \times \text{active1})\end{array}} \tag{2}$$

Training/ Validation

The dataset is randomly split into training and validation sets using an 80-20 split. The logistic model uses predictors from heart.logistic3 which are age, height, weight, blood pressure (ap_hi and ap_lo), cholesterol level, glucose level, smoking status, alcohol consumption, and physical activity.

The trained model is then used to predict the probability of cardiovascular disease for the validation set using the predict.glm() function.

Predicted probabilities are converted to binary predictions using a threshold of 0.5, where probabilities greater than 0.5 are classified as 1 (indicating presence of cardiovascular disease) and probabilities less than or equal to 0.5 are classified as 0 (indicating absence of cardiovascular disease).

The test error rate is calculated by creating a confusion matrix and dividing the sum of the missclassified observations (false positives and false negatives) by all the observations. Mean Test Error Rate Calculation: The mean error rate is calculated by averaging the test error rates obtained from 10 iterations of the validation process. In each iteration, the logistic model is trained on a randomly selected 80% of the data, and the mean error rate is calculated based on predictions made on the remaining 20% of the data. The test error rate provides an estimate of the model's performance in predicting cardiovascular disease risk.

The mean error rate obtained from the validation process is approximately 27.74%. This indicates that, on average, the logistic model misclassifies cardiovascular disease status in the validation set for 27.74% of observations.

Results

```r
summary(heart.logistic3)
```

```
Call:
glm(formula = cardio ~ age + height + weight + ap_hi + ap_lo +
    cholesterol + gluc + smoke + alco + active, family = "binomial",
    data = cardio.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.9638  -0.0979   0.9908   4.6623

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.144e+00  2.030e-01 -40.123  < 2e-16 ***
age           1.485e-04  3.556e-06  41.760  < 2e-16 ***
height       -5.255e-03  1.104e-03  -4.760 1.94e-06 ***
weight        1.521e-02  6.607e-04  23.024  < 2e-16 ***
ap_hi         3.953e-02  6.051e-04  65.330  < 2e-16 ***
ap_lo         3.007e-04  6.736e-05   4.464 8.04e-06 ***
cholesterol2  4.218e-01  2.592e-02  16.273  < 2e-16 ***
cholesterol3  1.134e+00  3.444e-02  32.921  < 2e-16 ***
gluc2         3.005e-02  3.438e-02   0.874    0.382
gluc3        -3.389e-01  3.809e-02  -8.898  < 2e-16 ***
smoke1       -1.256e-01  3.209e-02  -3.914 9.07e-05 ***
alco1        -1.681e-01  4.021e-02  -4.181 2.90e-05 ***
active1      -2.101e-01  2.105e-02  -9.980  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 97041  on 69999  degrees of freedom
Residual deviance: 80882  on 69987  degrees of freedom
AIC: 80908

Number of Fisher Scoring iterations: 13
```

The inclusion of statistically significant predictors, such as age, blood pressure, cholesterol levels, smoking status, alcohol consumption, and physical activity, in the final models under-

scores their crucial roles in assessing cardiovascular risk. These findings align with established medical knowledge regarding the impact of these factors on heart health. Additionally, the error rate obtained from the validation process indicates that the logistic models are performing well in predicting cardiovascular disease status, with an average error rate of only around 27.74%.

Overall, the results suggest that the logistic model, when incorporating significant predictors, can effectively predict cardiovascular disease risk. By leveraging key demographic, clinical, and lifestyle factors, healthcare practitioners can utilize these models to identify individuals at higher risk of cardiovascular disease and tailor preventive strategies and interventions accordingly. These findings contribute to a better understanding of cardiovascular risk assessment and highlight the potential of logistic models in supporting proactive heart health management strategies.

Neural Network Model (Valerio Martinez, Simrat Clair, Cruz Salas)

We chose a Neural network model as we wanted to see if the dataset contained more complex relationships that are not linear, as well as investigating all possible interactions between the input variables having an effect on the response variable.

Advantages of using a neural network model is that it can handle complex data that is not linear and adapt to changing input, also neural networks can detect all possible interactions between predictor variables, and it can also be developed using multiple different training algorithms.

While some of the disadvantages can be the "black-box" nature and have limited ability to identify possible casual relationships, and also limited to the computational power that we have access to, as having additional data or adding more complexity could affect the time to compute a model.

Neural Network Sketch

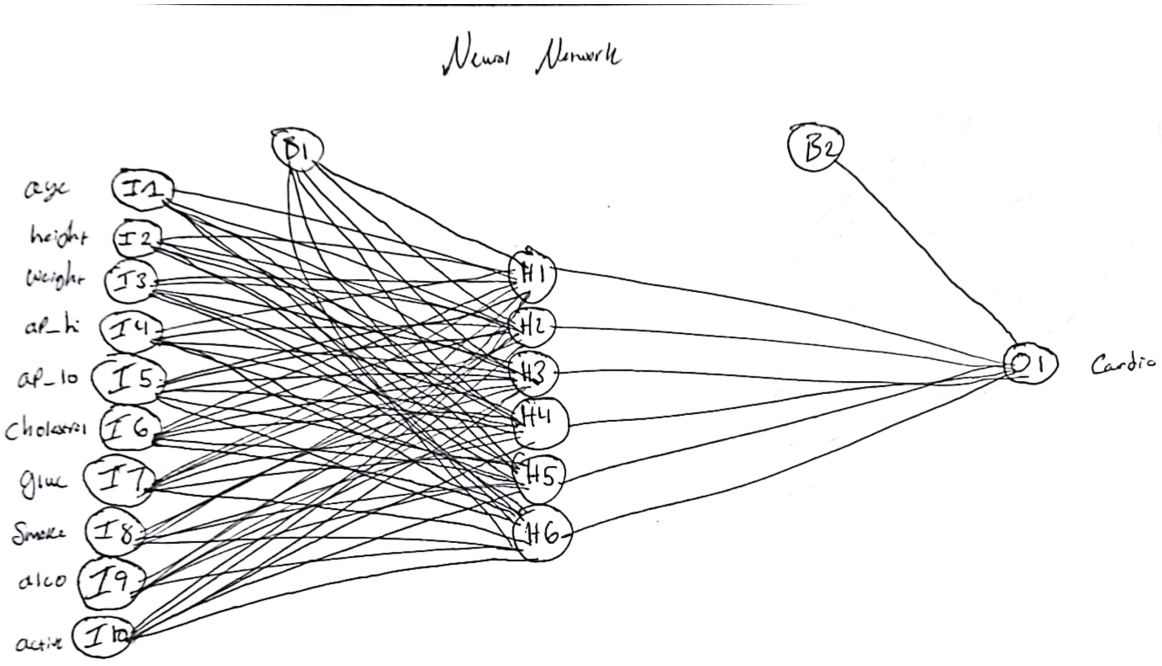


Figure 1: Neural Network Concept

Thought Process

We decided to use a single hidden layer as our data did not deviate that much from a linear relationship and did not have as much noise thus the data not being as complex, we decided a single hidden layer would be optimal.

We decided that using 6 nodes in the hidden layer would be a good starting point for our model as it was half of our initial input nodes which was 12. However, we found out it's essential to validate this choice through testing and adjust based on the specific needs and outcomes of your model's performance on validation datasets and could be implemented in further iterations of the model.

Regarding the nnet() function in R, "entropy" refers to the cross-entropy loss function, not an activation function. Cross-entropy is used to optimize classification accuracy by penalizing the difference between predicted probabilities and actual class labels. It complements the logistic and softmax activation functions used in nnet() for binary and multi-class classification, respectively. Thus, the term "entropy" describes the loss function used for training the model efficiently.

The initial weights in a neural network play a crucial part in the learning process, as it affect the speed of convergence in a neural network. The reason why weights can't generally be set to 0 is because the neurons in the network would receive the same signal and, therefore, will create inefficiency in training. This is why we decided to use 0.5 to make sure that the neurons are small enough to ensure that there's no numerical instability—allowing the learning to be normal, rather than too slow or diverging.

Training/ Validation

```
# Separate the data into 80% training and 20% testing.
sample = sample(1 : nrow(cardio.data), floor(nrow(cardio.data) * 0.8))
cardio.train = cardio.data[sample, ]
cardio.test = cardio.data[-sample, ]

# Build the neural network.
cardio.model = nnet(cardio ~ . - id - gender, data = cardio.train,
              size = 6, rang = 0.5, decay = 5e-2, maxit = 5000)
```

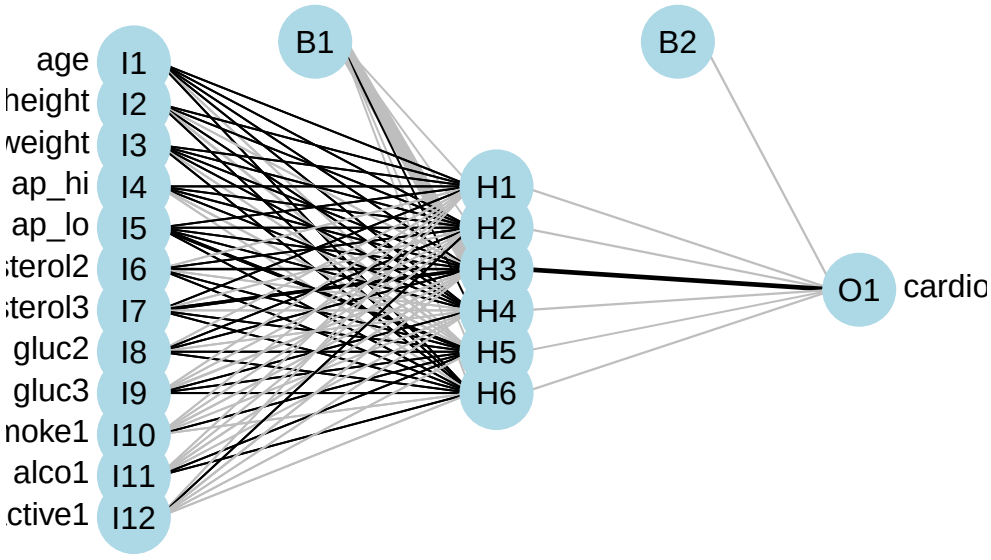
```
# weights:  85
initial  value 39022.927542
iter  10 value 38814.263045
iter  20 value 38643.556537
iter  30 value 38619.193201
```

```
iter  40 value 38599.346136
iter  50 value 36518.868686
iter  60 value 33905.073765
iter  70 value 33195.135531
iter  80 value 32690.991492
iter  90 value 31656.974514
iter 100 value 31019.534509
iter 110 value 30949.784523
final  value 30949.524137
converged
```

```r
# Print the model's information, plot, and summary
print(cardio.model)
```

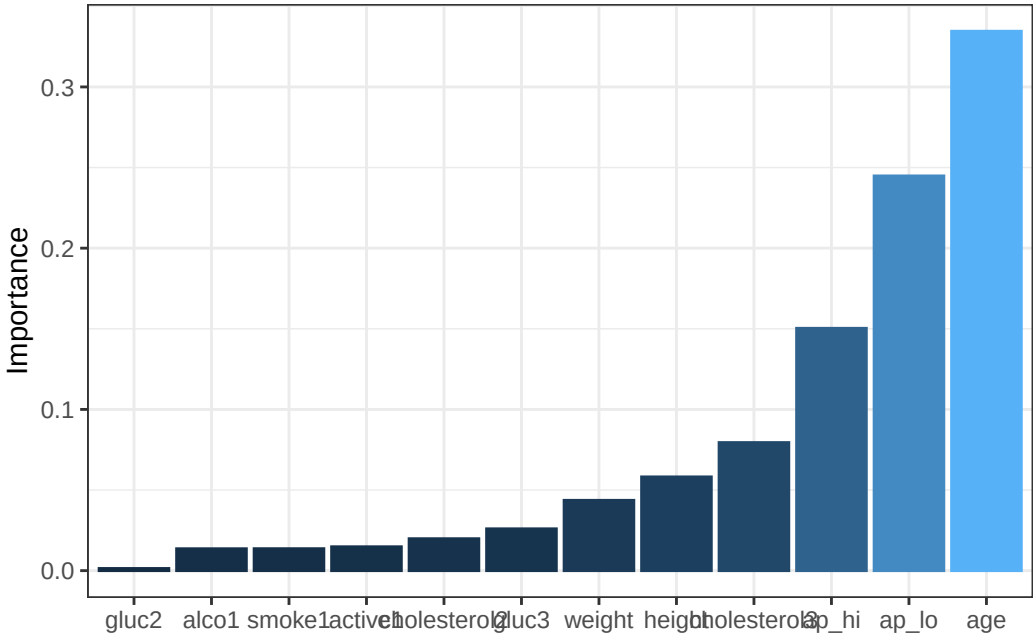
```
a 12-6-1 network with 85 weights
inputs: age height weight ap_hi ap_lo cholesterol2 cholesterol3 gluc2 gluc3 smoke1 alco1 acti
output(s): cardio
options were - entropy fitting  decay=0.05
```

```r
par(mfrow = c(1, 1), mar = c(1, 1, 1, 1))
plotnet(cardio.model)
```



Results

```
garson(cardio.model)
```



The Garson plot is used to interpret the relative importance of input features in a neural network. It decomposes the neural network weights into contributions by each input variable toward the output. In the Garson plot that we produced, it can be inferred that the variables **cholesterol, gluc, and smoke** were the variables that had the most significant factors in the cardio dataset, as these factors had the strongest correlation to increased risk of heart disease.

By analyzing the weights and influence of different inputs in the network, you can identify which factors are most predictive of cardiovascular diseases. This insight can help in understanding risk factors better and potentially guiding preventive measures as people who have significant predictors as lifestyle factors can be aware of the at-risk factors. The neural network model, trained on various patient health metrics, can predict the presence of cardiovascular disease with a reasonable level of accuracy. The model identifies key predictors and their influence, offering insights into the underlying patterns and risk factors associated with cardiovascular conditions.

```
                  cardio.test.predict
cardio.test.values    0    1
                 0 5459 1548
                 1 2213 4780
```

The Neural Network had a testing error rate of 26.38% as seen by the confusion matrix. The model helped answer our research question by indicating factors such as cholesterol, gluc, and smoke increase the rizk of developing cardiovascular disease.

Conclusion

The neural network model had a slightly lower test error rate than the logistic model with a 26.38% testing error rate compared to a 27.74% testing error rate of the logistic model. This shows that the neural network model performed better on the testing data set than the logistic model. The neural network model also better indicated which predictors are more statistically significant in causing heart disease, which answers our research question.

We can potentially improve the logistic model by improving our method of filtering variables to identify more statistically significant predictors, as stepwise regression eliminated only one variable. Some methods to improve the accuracy of the neural network model include training the model on more testing Data, and increasing model complexity (more hidden layers).

Bibliography

- https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
- "An Introduction to Statistical Learning with Applications in R" by Gareth M. James, Daniela Witten, Trevor Hastie, Robert Tibshirani