# MATH 4322 Final Project Group 9

## Introduction

## Neural Network Model

In this report, we employ a neural network to investigate the correlation between lifestyle factors—such as diet, exercise, and smoking—and the risk of cardiovascular disease. Using inputs ranging from physiological data to behavioral patterns, we aim to predict the occurrence of cardiovascular conditions. This analysis is designed to highlight key lifestyle influences on heart health and inform preventative strategies.

We choose a Neural network model as we wanted to see if the dataset contained more complex data that possibly not linear, as well as investigating all possible interactions between the input variables and 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**

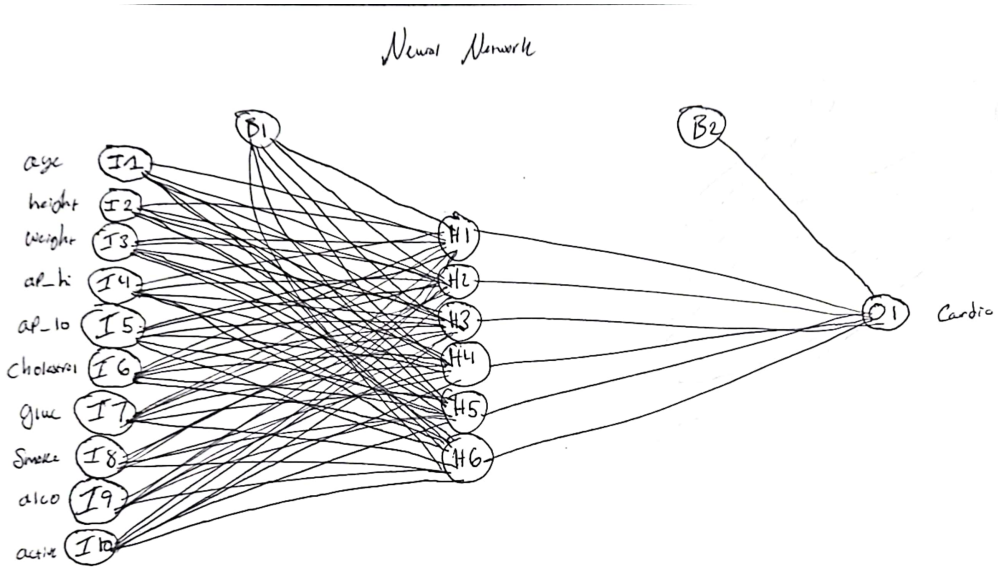



Figure 1: Neural Network Concept

## Methods

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

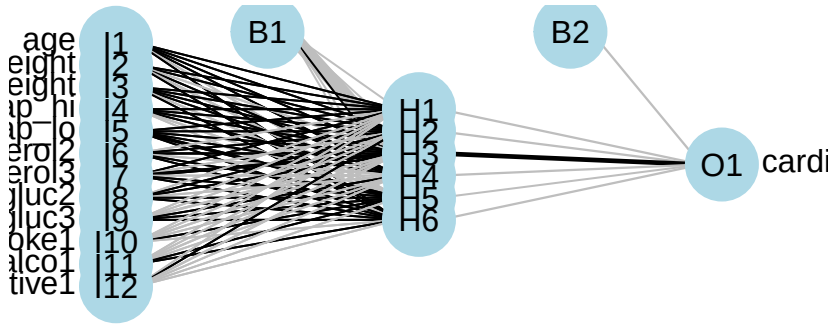
```
# Separate the data into 80% training and 20% testing.
sample = sample(1 : nrow(cardio.data), floor(nrow(cardio.data) * 0.8))
cardio.train = cardio.data[sample, ]
cardio.test = cardio.data[-sample, ]

# 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
```

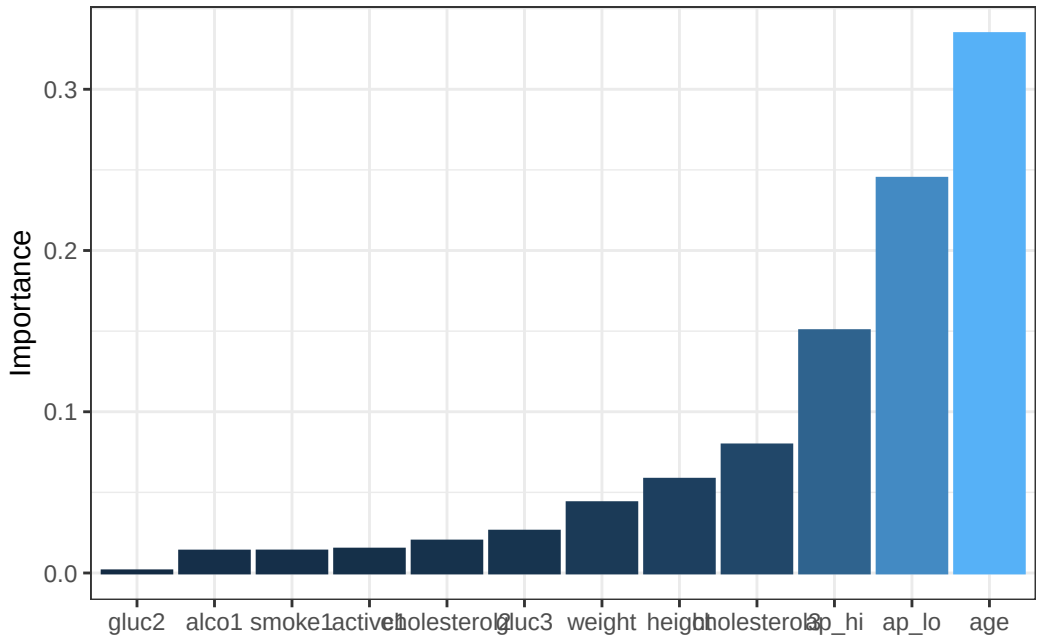
```
# Print the model's information, plot, summary, and garson plot.
print(cardio.model)
```

```
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
```

```
plotnet(cardio.model)
```




```
# summary(cardio.model)
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 inference 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.

## Results

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

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.

## 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. Thus displaying 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.

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


Include the inputs and the outputs of the data.

- Input
  - 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.

Include the question you are wanting to answer and the response variable.

- Output (response variable)
    - cardio
        * Indicates the presence or absence of cardiovascular disease.

Question we are trying to answer:

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