Baked Cauliflower (Kayleigh Kubit, Eddieb Sadat)
EM 489 - Data-Mining & Risk Assessment
Dr. Zigh
December 19, 2022

We pledge our honor that we have abided by the Stevens Honor System.

# Final Project: Cardiovascular Disease Causation Analysis

## Using a Neural Network to Determine Root Causes for Cardiovascular Disease

## Table of Contents

## 1. Introduction

As society and technology have grown, the emphasis on better healthcare and medical capabilities have grown exponentially with the hopes of improving people's lives. This has come in the form of new treatment solutions, increased knowledge in human physiology, and new technologies to aid in recovery. These solutions, though beneficial to medicine, require extensive knowledge in the medical field, as they will be directly used by patients or caregivers. The result is a limit on growth of the medical field caused by knowledge gaps.

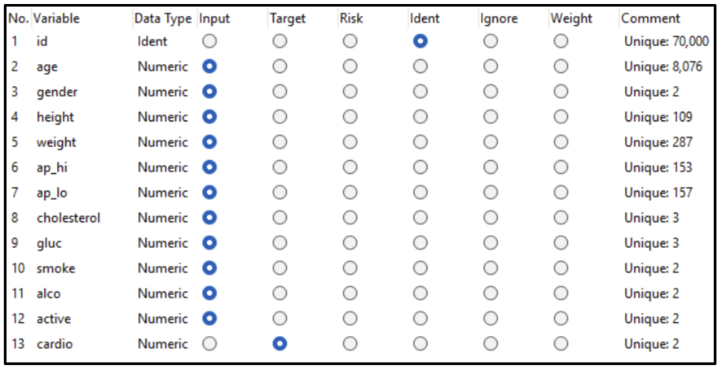
Another exponentially growing field is data science and analysis. Data science, which utilizes statistics, analysis, and visualization, has allowed people to better understand the world around them. What normally would look like a massive pile of irrelevant data to a person could actually contain invaluable information about trends, patterns, and relationships between different variables. With proper data analysis, this information can be identified and utilized to improve society (e.g. optimizing a car for less emissions).

One particular data science technique is to use existing data to generate predictive models. In a predictive model are input variables and a target variable. The model utilizes the information from input variables to predict what the target variable will be. Artificial Neural Networks (ANNs), models that simulate the human neural network, are capable of producing robust and accurate predictive models. Thus, by developing an ANN using medical data, it is possible to predict different medical outcomes, which is important for disease-prevention or early-detection. Furthermore, because only the data is being analyzed, extensive medical knowledge is not needed, albeit useful.

In the following report, an aggregate dataset of patient information relating to cardiovascular disease is used to generate an ANN. It will showcase the process of obtaining, analyzing, and preparing the data, the iterative process to develop the best possible ANN, and finally demonstrate the capabilities of data science in conjunction with the medical industry.

## 2. Data Understanding

To begin, the data was imported into Rattle. On closer inspection, we found that the dataset was not comma separated, but semicolon separated. To correctly import the data, the separator in Rattle was set as a semicolon.



| No. | Variable | Data Type | Input | Target | Risk | Ident | Ignore | Weight | Comment |
|---|---|---|---|---|---|---|---|---|---|
| 1 | id | Ident | ○ | ○ | ○ | ● | ○ | ○ | Unique: 70,000 |
| 2 | age | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 8,076 |
| 3 | gender | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 2 |
| 4 | height | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 109 |
| 5 | weight | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 287 |
| 6 | ap_hi | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 153 |
| 7 | ap_lo | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 157 |
| 8 | cholesterol | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 3 |
| 9 | gluc | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 3 |
| 10 | smoke | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 2 |
| 11 | alco | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 2 |
| 12 | active | Numeric | ● | ○ | ○ | ○ | ○ | ○ | Unique: 2 |
| 13 | cardio | Numeric | ○ | ● | ○ | ○ | ○ | ○ | Unique: 2 |

*Figure 1: Initial configuration for data once imported to Rattle*

The dataset was found to contain 13 columns and 70,000 rows and no missing values [Fig. 1]. The last variable, cardio, is a binary variable which indicates whether or not there is a presence of cardiovascular disease. We considered this to be a good target variable to use for a predictive model for cardiovascular disease detection because the ANN we intended to create would output binary values. (The dataset can be found at https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.)

## 3. Data Preparation

### 3.1 Erroneous Data

We then cleaned the data, transforming certain variables and removing erroneous values from others. We began with transforming the Age variable in two ways. First, we converted it from days to years by dividing the given value in days by 365.25 and rounding to four decimal places. We then used the results from the age in years to create an additional variable with Age in bins by splitting age-in-years into four groups: <40, 40 - 50, 50 - 60, and 60+. This simplification of data was intended to allow the ANN to produce more effective results, as it tends to favor more simple data with less unique variables. Using a 70/15/15 partition, a bar chart was generated to determine the distribution of values across these bins [Fig. 2].

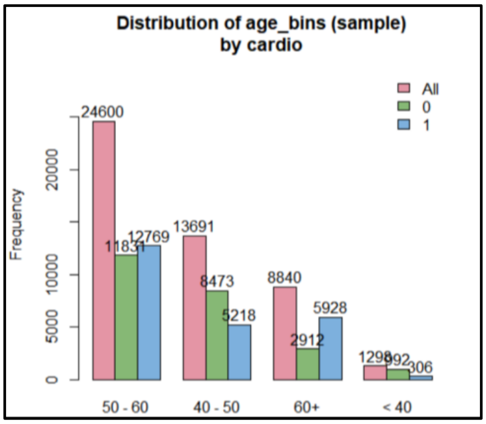
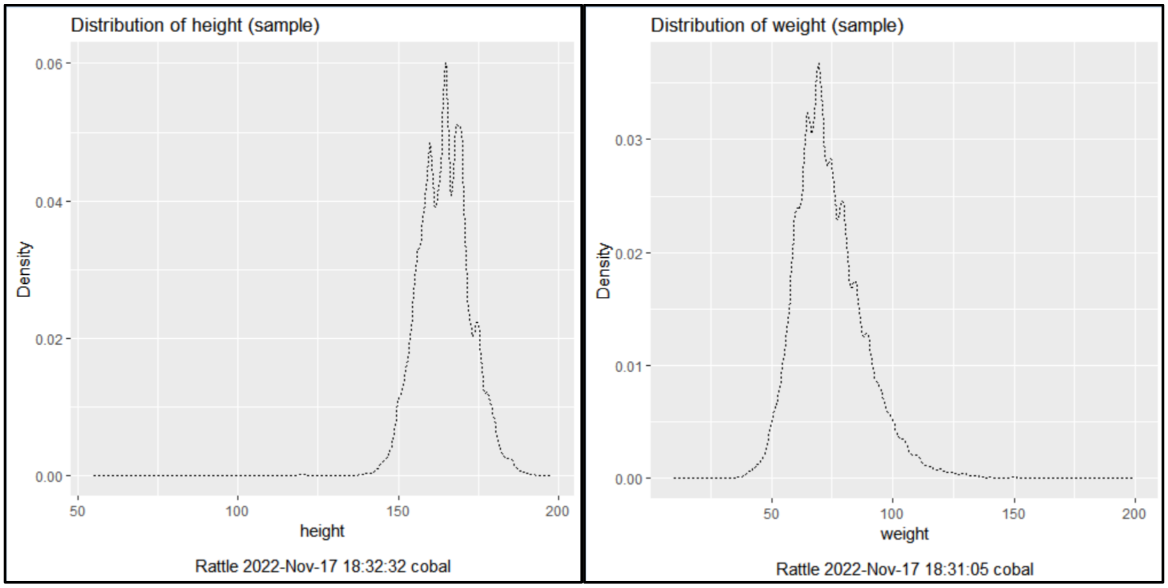


*Figure 2: Distribution of age_bins variable (1 = Cardio. disease present, 0 = absent)*

Since cardiovascular disease occurs in naturally higher frequencies in older populations, the skew in the data did not concern us. As shown in Figure 2, as age increases, cardiovascular disease (represented by "1") increases proportionally to within each group.

The Gender categorical column was transformed to binary (1 - Female, 0 - Male) from its previous configuration (1 - Female, 2 - Male) in order to make the data easier for the ANN to interpret. Furthermore, height values under 140cm removed or above 207cm were removed. Anything outside this range is considered to be an error in data input as they are generally unrealistic heights for humans. Similarly, weights under 40kg removed, as an adult weight under

40kg is highly unlikely. Figures 3 and 4 show the distribution of these variables; our decisions are justified because the vast majority of data points fall within our retained range.



*Figures 3 and 4: Histograms of height and weight, respectively*

Next, we adjusted the blood pressure variables. Systolic blood pressure (ap_hi) values outside the range of human viability by +/- 15 (under 75 mmHG or over 195) were removed, as these are unrealistic values which would lead to immediate hospitalization or even death. Similarly, diastolic blood pressure (ap_lo) values under 45 mmHG or over 135 mmHG were removed.

Initially, 1637 data points were set to be excluded from the set based on the above criteria, where 79 of these were data points that had errors in multiple columns. Diastolic blood pressure (ap_lo) accounted for the vast majority of data points removed, with 1090 data points having errors in this column. We considered whether there might be some systematic entry error for values in this column, such as incorrect units of measurement. However, this is not the case as blood pressure is always recorded in mmHg, and there were varying errors in the data themselves (some were too large, yet not easily explainable by a mistyped key, such as “2088”; some were too small and similarly not explainable, such as “0” or “40”; one was negative, which is clearly impossible, but had an error in another column that meant the data point would be removed anyway). We did notice a large number of “1000,” “1100,” and “1200” values, which could be explained by an accidental zero being input. When we replaced these values with “100,” “110,” and “120,” the number of removed data points from this column dropped to 247. Further adjustments were not completed, but given the reasonability of this assumption, we excluded only the 247 data points that still appeared erroneous. Due to the complexity of comparing height and weight to determine feasibility of the values, we assumed that values that were not excluded by the above bounds were accurate. It is important to note that adults who are missing certain limbs can result in ‘abnormally’ less weight and height compared to the ‘average’ person. However, these are very specific circumstances that add unwanted complexity to the scope of this project. Thus, we stand by the decision to remove the extreme outlying values from the dataset, although it is possible that some inaccurate data was included.

With these adjustments completed, the total of data points removed was 815, with 59 having errors in multiple columns. As our dataset contains 70000 entries, the removed data points account for only ~1.2% of the dataset; this is an acceptable percentage to remove to maintain data integrity.

### 3.2 Preliminary Analysis

The graph was generated to visualize the distribution of the data by gender [Fig. 5]. According to the graph, the mean is approximately 0.65, indicating that there is a higher proportion of females to males in the dataset. By manual analysis of the data, it was determined there are 24175 included male entries and 45010 female entries.

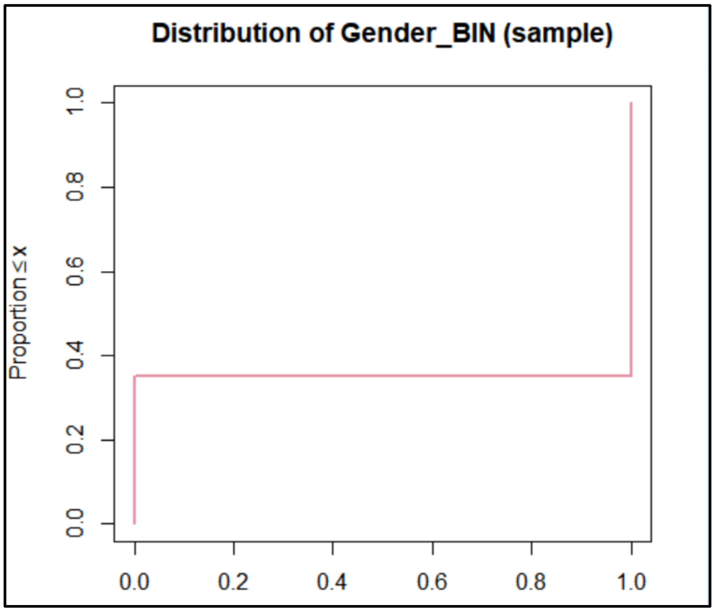


*Figure 5: Distribution of Gender*

Boxplots were used to quickly and effectively analyze variables. They show distributions for each category in a variable, and basic statistics such as mean and median. This information is a quick method to better understand each variable and determine any potential areas for concern.

A boxplot of age-in-years (age_year) by the binned ages (age_bins) was generated [Fig. 6]. Save for a few outliers in the youngest group, the means of each bin sit fairly squarely in the middle of the range (e.g. the 40 - 50 bin has a mean of ~45). This reaffirms that the binning categories were reasonable. The tails of each bin also do not exceed their age range, indicating that all data is in their appropriate categories. No areas of concern are identified.

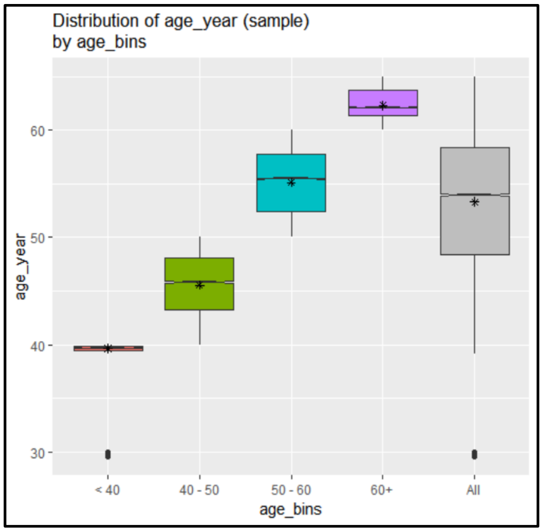


*Figure 6: Boxplot of age_years to age_bins*

Boxplots for height and weight were also each generated by age_bin [Fig. 7 & 8]. Height and weight are similarly distributed evenly by age; this leads us to think that age will not be a confounding variable (causing significant unexpected differences in other variables that may create error in the model).

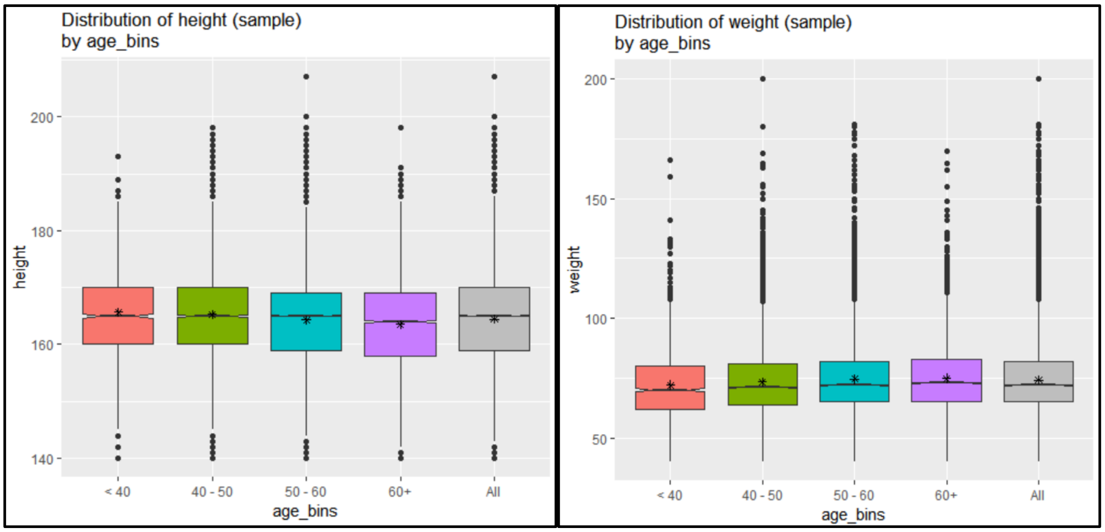


*Figure 7 & 8: Boxplots of height and weight to age_bins*

Boxplots for systolic (ap_hi) and diastolic (ap_lo) were also each generated by age_bin [Fig. 9 & 10]. Blood pressure readings for both systolic and diastolic tend to rise with age, as shown by the gradual increase in blood pressure means per increased age group. Though these rises seem small, they can be incredibly deadly. As adaptable as human bodies are, the cardiovascular system is a fine-tuned system that has thresholds. Surprisingly still, the distributions of blood pressures per age group also have a very large range, albeit a vast majority are within a small range.

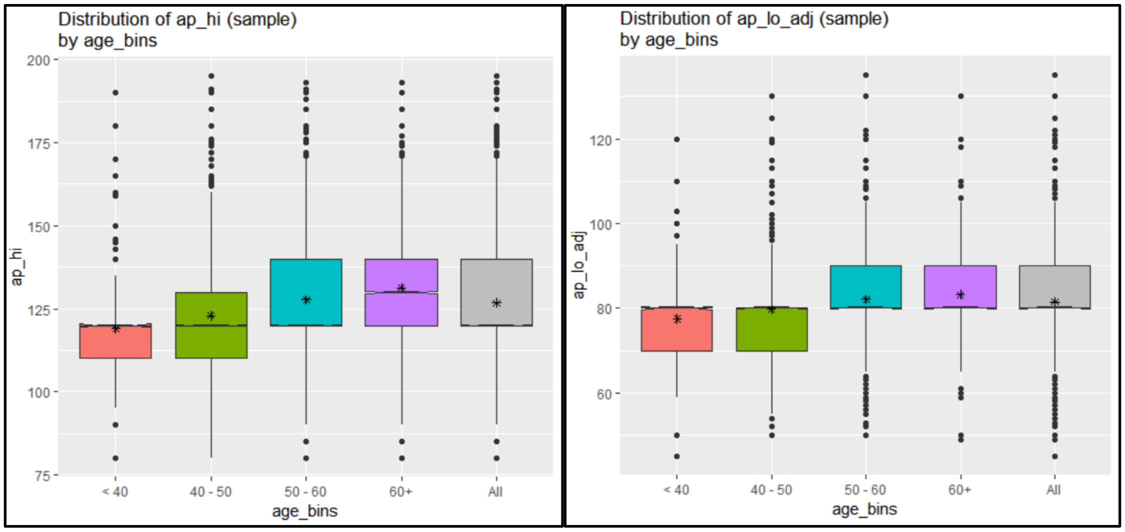


*Figure 9 & 10: Boxplots of ap_hi and ap_lo to age_bins*

Next, boxplots for systolic (ap_hi) and diastolic (ap_lo) were each generated by cardio [Fig. 11 & 12]. In both cases, people who were positive for cardiovascular disease (cardio = 1) showed consistently higher blood pressures than those who were negative. The means also appear to be very skewed, indicating that some blood pressures are extremely more common than other values.

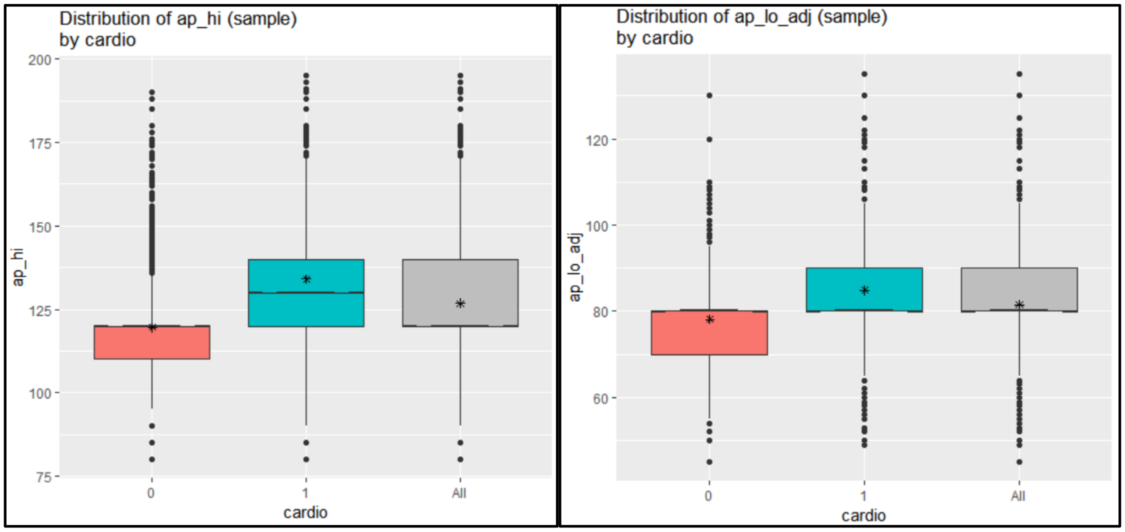


*Figure 11 & 12: Boxplots of ap_hi and ap_lo to cardio*

Finally, a boxplot for age in years (age_year) and cardio was generated [Fig. 13]. The mean age of users who were positive for cardiovascular disease was larger than the mean for those who were negative. This indicates that onset of cardiovascular disease occurs more frequently in older age. Interestingly, the distributions of each category look even, indicating a normal distribution of age to cardio, with a slight skew towards older age groups.

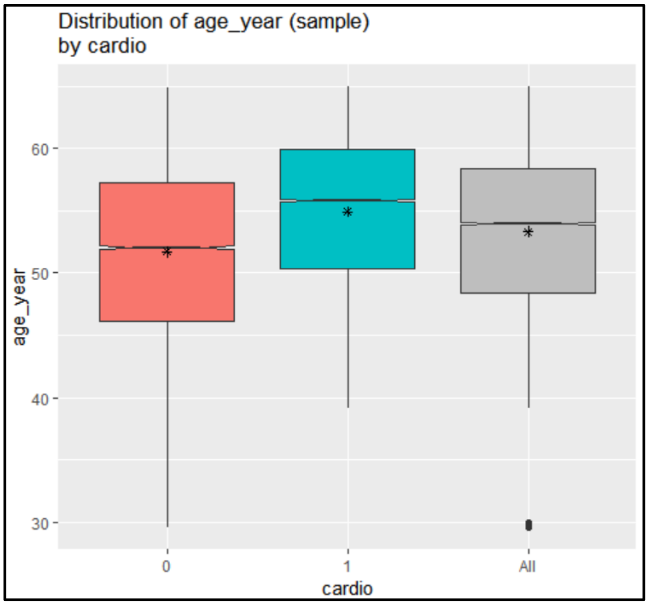


*Figure 13: Boxplot of age_year and cardio*

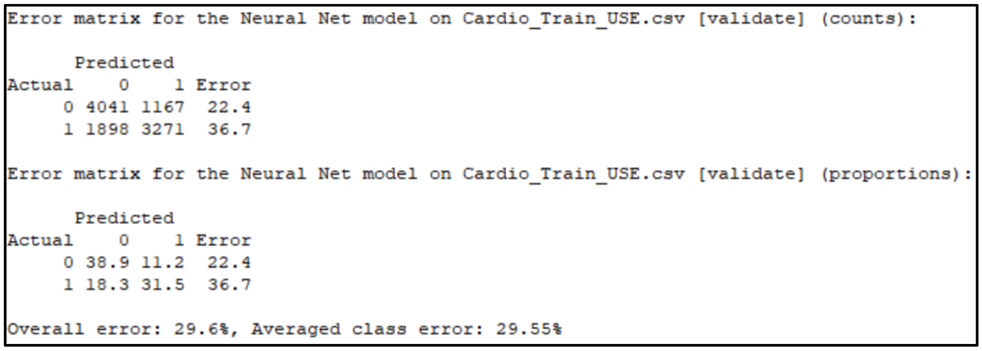
# 4. Modeling

As previously mentioned, a 70/15/15 partition was applied to the dataset. This has been a standard partition in our experience, and we saw no reason to use an alternative partition for this data.

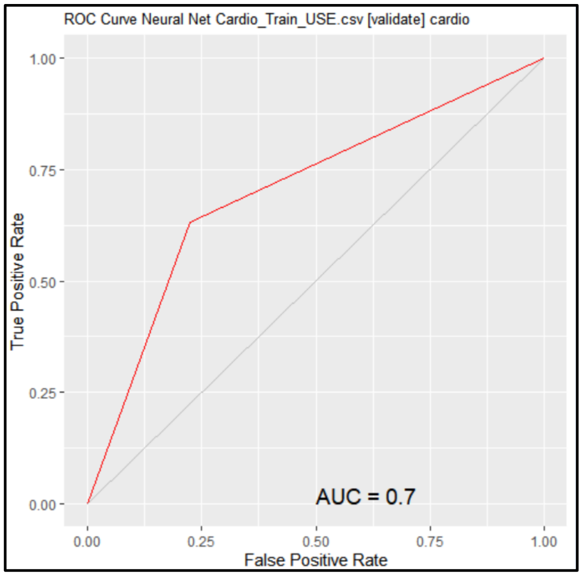
Most ANN models can be completed using only one hidden layer, but we were open to experimenting with using anywhere from 0 to 2 hidden layers. We determined in advance that the final model would use parameters that produced the best results in the validation phase.

## 4.1 ANN 1 & 2

ANN #1: Completed with 0 layers using Age in years, and ANN #2: Completed with 1 layer using Age in years, and ignoring Age in bins. The results for both were exactly the same [Fig. 14 & 15].



```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 4041 1167  22.4
     1 1898 3271  36.7

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 38.9 11.2  22.4
     1 18.3 31.5  36.7

Overall error: 29.6%, Averaged class error: 29.55%
```



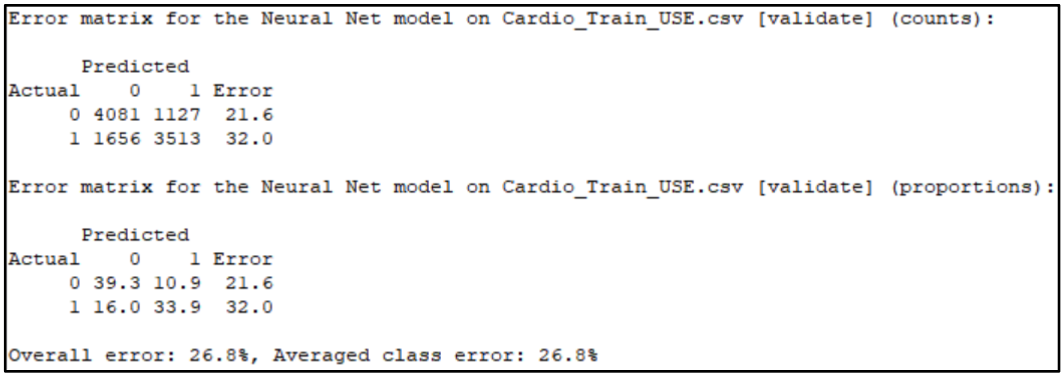
*Figures 14 & 15: Error matrix and ROC curve for ANN 1 & 2*

Initially, we assumed that Age being represented in years would produce very undesirable results since it would input dozens of different values for the ANN to analyze (the variable would be too complex). Surprisingly, this was not the case, since the final model was able to successfully predict the presence of cardiovascular disease 70% of the time. This is insufficient accuracy for real use in a medical setting, but it can still be utilized with a respectable confidence in results and shows the potential of the model.

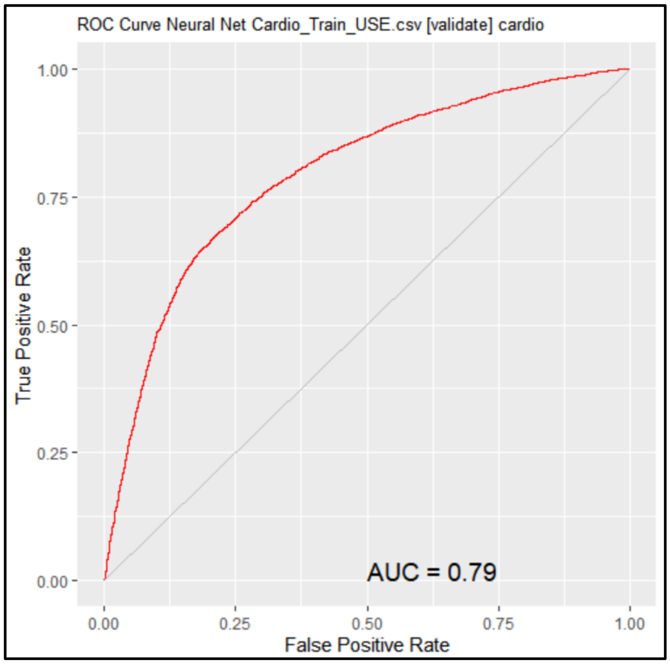
The next step was to create a model using Age as bins to evaluate whether it would improve the model accuracy.

## 4.2 ANN 3

ANN #3: Completed with 0 layers using Age in bins [Fig. 16 & 17].



```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 4081 1127  21.6
     1 1656 3513  32.0

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 39.3 10.9  21.6
     1 16.0 33.9  32.0

Overall error: 26.8%, Averaged class error: 26.8%
```

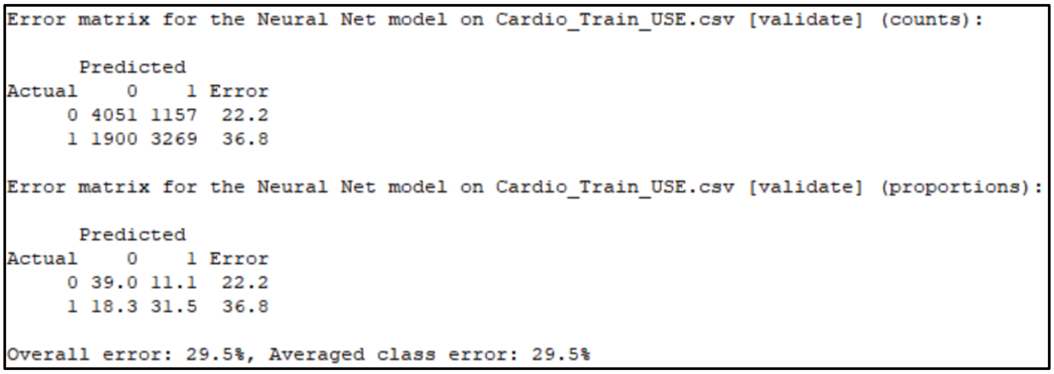


*Figures 16 & 17: Error matrix and ROC curve for ANN 3*

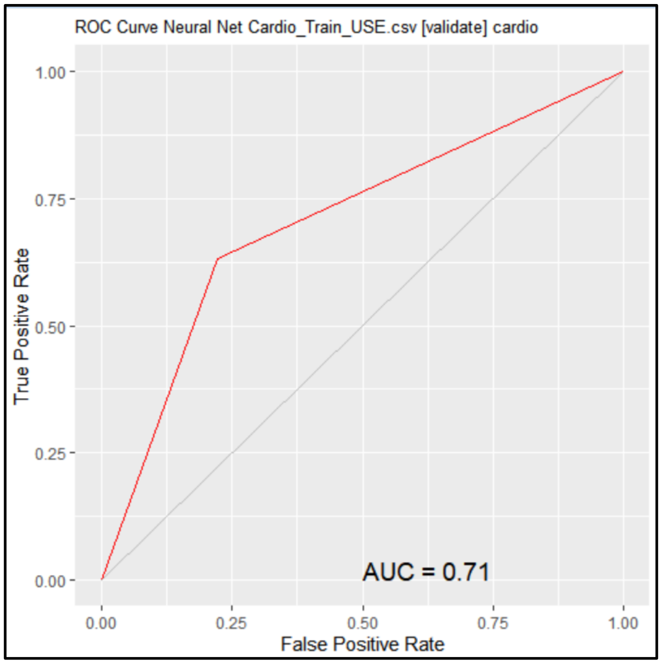
This was clearly an improvement over the zero-hidden-layer model, as the AUC increased from 0.7 to 0.79. Interestingly, the curve also became significantly smoother, as opposed to the sharp point present on the previous AUC. This means that there is a much less distinct tradeoff point between specificity and accuracy.

### 4.3 ANN 4

ANN #4: Completed with 1 layer using Age in bins [Fig. 18 & 19].



```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 4051 1157  22.2
     1 1900 3269  36.8

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 39.0 11.1  22.2
     1 18.3 31.5  36.8

Overall error: 29.5%, Averaged class error: 29.5%
```



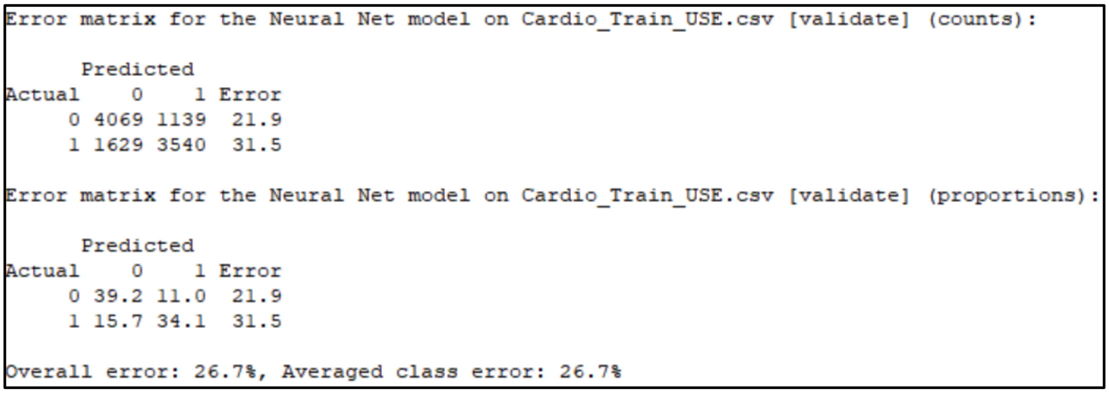
*Figures 18 & 19: Error matrix and ROC curve for ANN 4*

Clearly, adding a hidden layer when using the Age in bins variable took our results backwards in terms of desirability–the AUC dropped from 0.79 to 0.71, indicating a less reliable performance from this model than the one with no hidden layers.

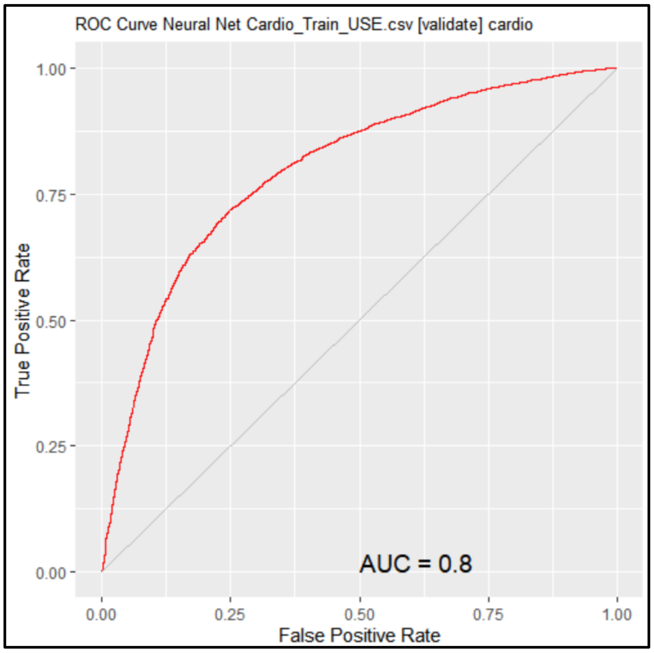
At this point, ANN #3 remained the best option; the next step was to evaluate the merit of including both the Age in years and Age in bins variables in the model.

### 4.4 ANN 5

ANN #5: Completed with 0 layers using both Age in bins and Age in years [Fig. 20 & 21].



```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 4069 1139  21.9
     1 1629 3540  31.5

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 39.2 11.0  21.9
     1 15.7 34.1  31.5

Overall error: 26.7%, Averaged class error: 26.7%
```

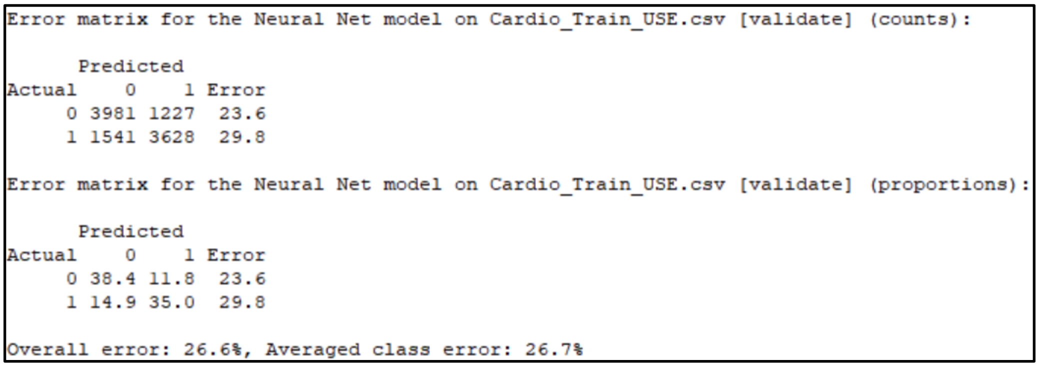


*Figures 20 & 21: Error matrix and ROC curve for ANN 5*

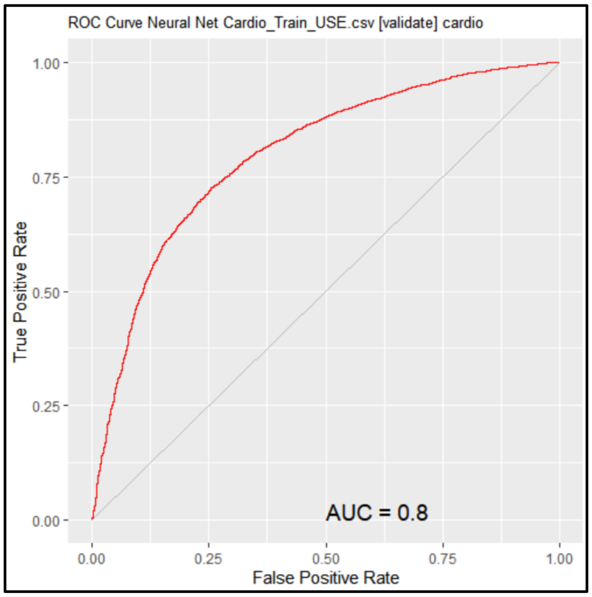
Interestingly, using both Age variables in years and bins resulted in a model that is better than the previous ANNs. This is likely because having both variables included allows the model to make better sense of the age information–the general categories of the bins variable help to balance out the specificity of the Age in years variable. Notably, the AUC curve returned to a smooth curve.

### 4.5 ANN 6

ANN #6: Completed with 1 layer using both Age in bins and Age in years [Fig. 22 & 23].



```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 3981 1227  23.6
     1 1541 3628  29.8

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 38.4 11.8  23.6
     1 14.9 35.0  29.8

Overall error: 26.6%, Averaged class error: 26.7%
```

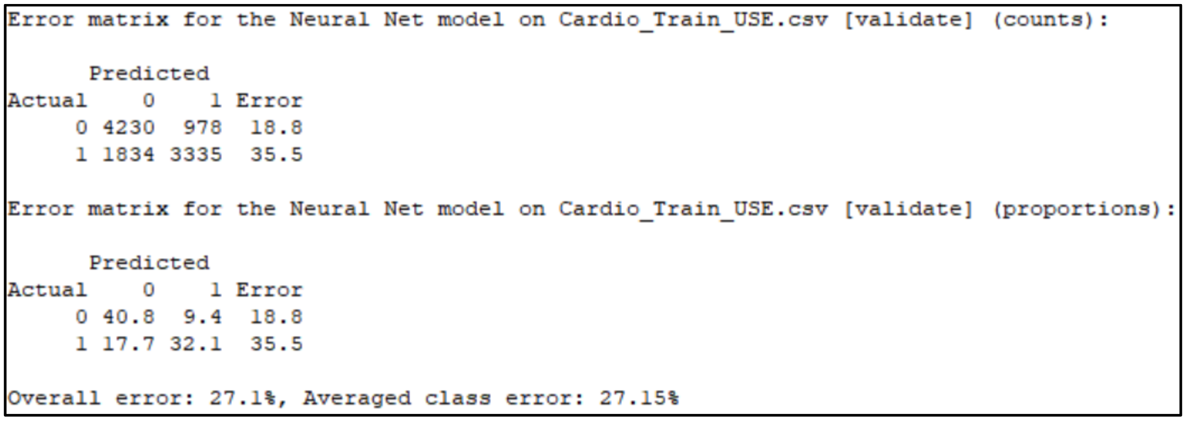


*Figures 22 & 23: Error matrix and ROC curve for ANN 5*

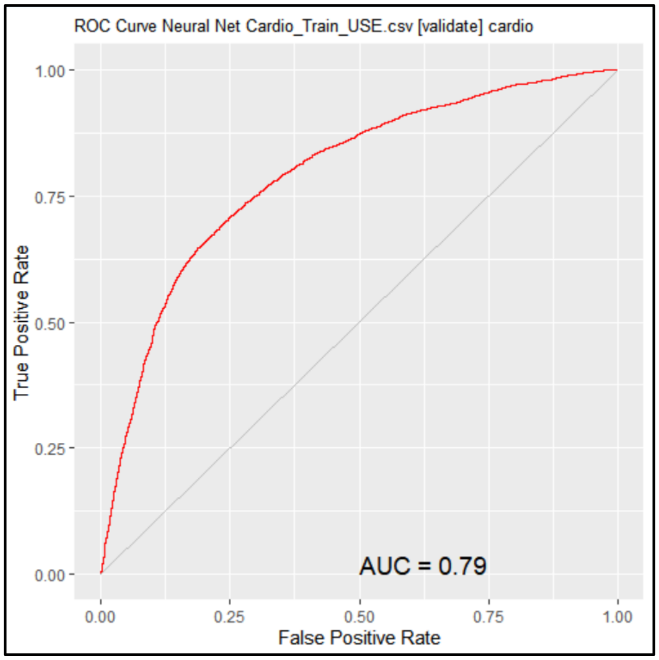
While very similar, the AUC for this final model was ever so slightly higher than that of the previous model (this model's AUC was 0.7991, vs. a value of 0.7964 for ANN #5). This meant that this model was our preferred model with which to proceed.

### 4.6 Testing with new Input Modifications

From this point, all further adjustments involved changing the data inputs to the model. We decided to evaluate whether utilizing weights would improve the model by emphasizing certain more significant variables. We began exploring potential effects by changing the "smoke" binary variable from an input to a weight.

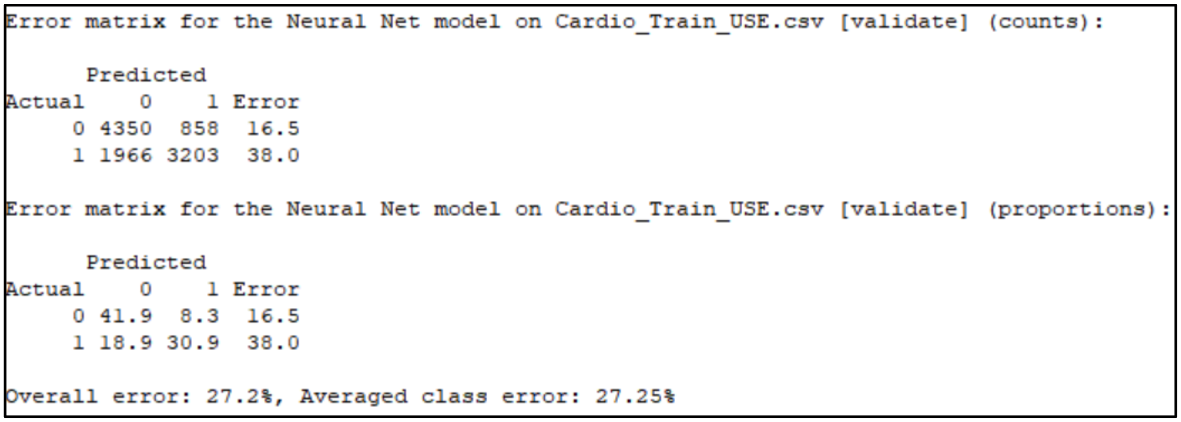


```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 4230  978  18.8
     1 1834 3335  35.5

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 40.8  9.4  18.8
     1 17.7 32.1  35.5

Overall error: 27.1%, Averaged class error: 27.15%
```

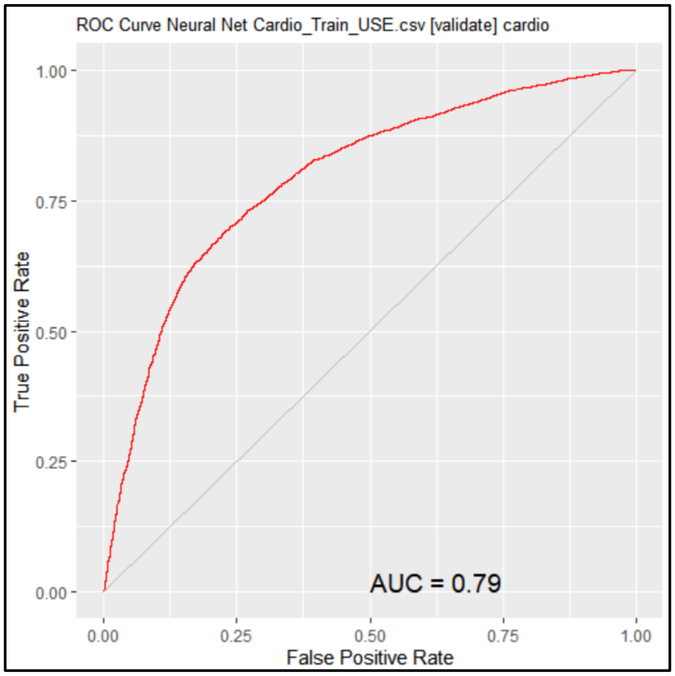


*Figures 24 & 25: Error matrix and ROC curve for ANN with smoke as weight*

No improvement in the model was observed. Next, the Alcohol binary variable was made as a weight for the model [Fig. 26 & 27].

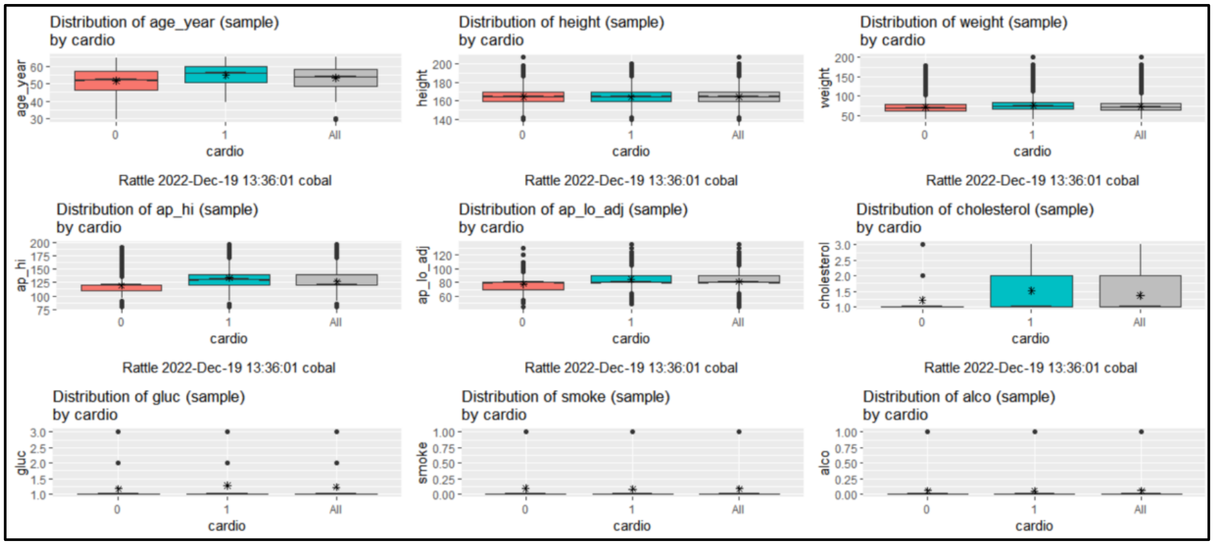


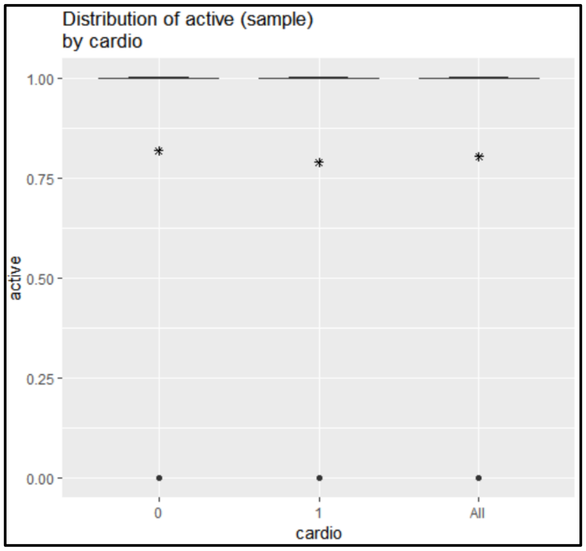
```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 4350  858  16.5
     1 1966 3203  38.0

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 41.9  8.3  16.5
     1 18.9 30.9  38.0

Overall error: 27.2%, Averaged class error: 27.25%
```



*Figures 26 & 27: Error matrix and ROC curve for ANN with alcohol as weight*

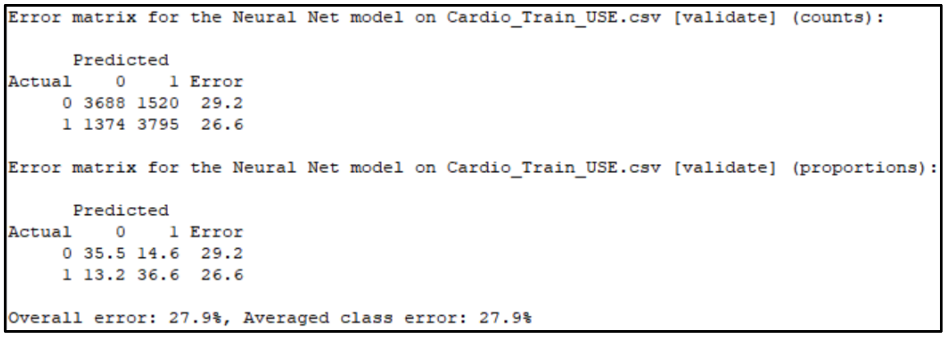
Combining these two weights through a weight calculator formula yielded no better results, and was still slightly worse than the best ANN model. We realized this likely stemmed from a lack of understanding on our part regarding the implementation of weights; we further recognized that this initial experimentation was rather baseless, as we did not have distinct reasoning for selecting the variables we chose to use as weights. We decided to analyze the data further and continue experimenting to see if our results would improve with an intentional selection of variables.





*Figures 28 & 29: Boxplots of all variables and cardio*

From Figures 28 and 29, it is clear that glucose, smoke, and alcohol are all concentrated at the lowest levels, whether or not the observation is one with cardiovascular disease; the active variable is concentrated at the highest level for both conditions. Height and weight are also mostly similar. However, cholesterol does show a significant difference between the conditions, which led us to believe that it may act as a suitable weight; it is a variable with integer values between one and three, which from our understanding would provide a clear basis for the ANN to use as a weight. Thus, an ANN was tested with cholesterol as a weight [Fig. 30].



```
Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (counts):

      Predicted
Actual    0    1 Error
     0 3688 1520  29.2
     1 1374 3795  26.6

Error matrix for the Neural Net model on Cardio_Train_USE.csv [validate] (proportions):

      Predicted
Actual    0    1 Error
     0 35.5 14.6  29.2
     1 13.2 36.6  26.6

Overall error: 27.9%, Averaged class error: 27.9%
```

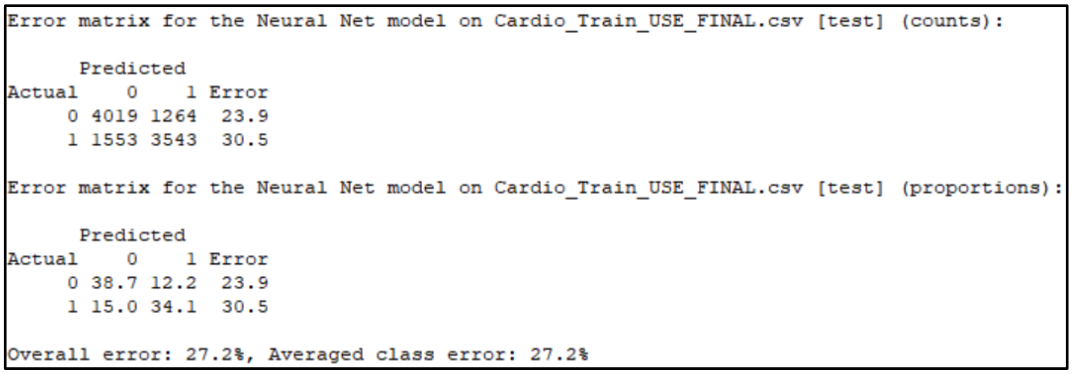
*Figures 30: Error matrix for ANN with cholesterol as weight*

Several iterations, with various combinations of the above-mentioned variables ignored and included, were tested; as can be seen in Figure 30, the best result (ignoring alcohol and glucose, including blood pressures, smoking, and activity, cholesterol as weight) still yielded worse results than our previous model.

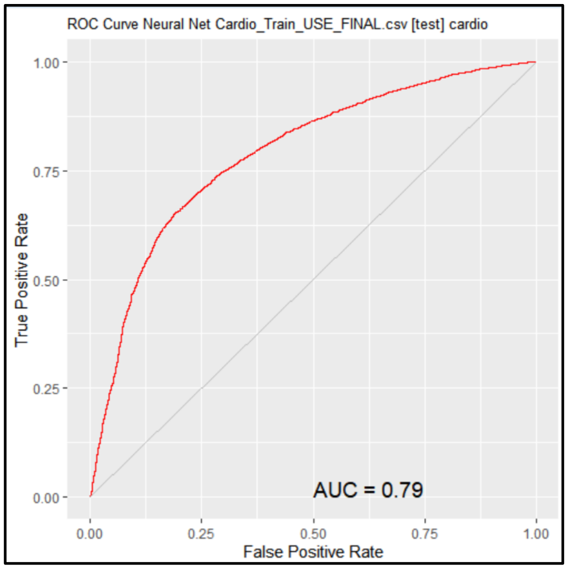
Finally, we attempted to create a weight variable, in the dataset itself, that would give more weight to those variables with high age, blood pressures, and cholesterol, as these were the variables we determined to be most significant based on their boxplots. This variable was created by summing age and blood pressures, multiplying this sum by cholesterol, dividing this by a large number to get the values into the single digits, subtracting the mean of the dataset from each result, and then taking the absolute value. Despite this effort, the results were still not improved over the initial model's error of 26.6%, yielding a minimum overall error of 27.2%.

### 4.7 Final Model Testing

Since the best model produced was from ANN 6, it was decided to be the final model for this project. To ensure the model's accuracy, it was tested [Fig. 31 & 32]



```
Error matrix for the Neural Net model on Cardio_Train_USE_FINAL.csv [test] (counts):

      Predicted
Actual    0    1 Error
     0 4019 1264  23.9
     1 1553 3543  30.5

Error matrix for the Neural Net model on Cardio_Train_USE_FINAL.csv [test] (proportions):

      Predicted
Actual    0    1 Error
     0 38.7 12.2  23.9
     1 15.0 34.1  30.5

Overall error: 27.2%, Averaged class error: 27.2%
```



*Figures 31 & 32: Error matrix and ROC curve for Final Model*

These results are quite close to our validation results, meaning we did not overfit our model to the data. This is what we expected, because we did not remove any variables from the model in its final configuration.

## 5. Conclusion

Initially, we had thought that the model would be more accurate than it currently is (about 80% accurate) given our experience with examples from the EM-489 curriculum. However, given that there were many mistakes in the original dataset, alongside natural irregularities that are a result of differences in humans, it is understandable for our final model to not be perfectly accurate. Even though the final model has a fairly high accuracy, it is not suitable for important decision-making in the real world (e.g. giving someone medication for heart disease based on model results). Especially for medical use, an acceptable model would likely have no less than 95% accuracy. However, the model can still potentially be used for low-risk settings to help doctors to preliminarily evaluate whether a patient is likely to have cardiovascular disease or not during a normal physical examination. To build a better model in the future, additional data should be included, and perhaps an alternative ANN model produced. Additionally, the Type I error should be prioritized for reduction; currently, the Type II error (false positive) is lower than Type I (mistaken clearance). In reality, it is more preferable to have a false positive, which would likely then be checked by a subsequent test to verify its veracity (and would then be determined to be in error), rather than to be unaware of an existing health condition (as once cleared, an additional test is unlikely to be pursued).

Even though the Baked Cauliflowers have little-to-no experience in the medical field, we were still able to produce a decently accurate model for predicting cardiovascular disease. This is a true testament to how powerful data science and analysis can be, as it opens the door to many industries and opportunities through a single multidisciplinary skillset.