**CIS 4930**

**Predictive Modeling: Heart Disease Analysis Written Report**

Desiree Bethelmie, Emory Ellis, Navjit Singh, Seyoung Koo

A person holding his chest with a drawing of a heart

Description automatically generated

**Table of Contents**

Project Description..........................................................................................................................2

Data Exploration and Processing …...............................................................................................3

Models

Logistics Regression...........................................................................................................4

K-Nearest Neighbors..........................................................................................................5

Neural Network..................................................................................................................6

Results and Discussion...................................................................................................................7

Summary.........................................................................................................................................8

**Project Description & Business Problem:**

The Beaumont health clinic is focused on proactive healthcare strategies and aims to assess the risk factors associated with heart disease among its patient population. Understanding these risk factors is crucial for early intervention and personalized healthcare management, aligning with the clinic's commitment to preventive care. The primary goal is to leverage the extensive dataset encompassing health indicators, lifestyle factors, and medical history to develop a predictive model for identifying individuals at higher risk of developing heart disease. The clinic seeks to deploy this model to proactively identify and intervene with patients susceptible to heart disease, potentially reducing adverse health outcomes and healthcare costs.

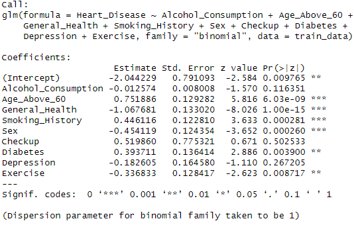
The project aims to construct a robust predictive model leveraging comprehensive patient data to effectively identify individuals at an elevated risk of developing heart disease within the Beaumont health clinic's patient population. Its primary objectives encompass the development of a sophisticated predictive framework that accurately discerns risk factors associated with heart disease. By categorizing patients into risk groups, the model seeks to enable proactive healthcare management strategies tailored to varying risk levels. Through this approach, the project endeavors to facilitate early intervention, precise risk stratification, and personalized healthcare initiatives. Ultimately, the overarching goal is to improve patient outcomes by initiating timely interventions, thereby emphasizing preventive care and enabling the clinic to deliver targeted and proactive healthcare services, enhancing the overall quality of patient care provided.

**Data Exploration and Processing**

In the context of heart disease risk assessment within the Beaumont health clinic, we began our Data Cleaning steps by Analyzing the data for missing values or Nas, which the dataset did not contain any. We transformed the Age, Diabetes, and General Health variables by setting thresholds, condensing the information from various intervals into broader, more easily interpretable categories, and aiding in data analysis and interpretation. For instance, we decided to set the threshold at the age “60” mark for the Age variable, which was originally in multiple intervals containing 6 age groups (18-24,24-30...etc). Identifying age group below 40 to 60 to No, if below age of 60 and yes otherwise. Then, finally renaming the variable to Age Above 60 or over". The Heart Disease variable serves as the presumed target, while predictors such as Exercise, Diabetes, Sex, Checkup, Depression, Smoking History, and Age Above 60 were identified as predictors for our Analysis. The correlation coefficients demonstrate relationships between Heart Disease and other dataset variables. Stronger correlations, such as General Health, hint at better-reported health correlating with lower heart disease likelihood, while diabetes shows a weak positive association. Moderate correlations, including Arthritis and Smoking History, suggest somewhat stronger links to heart disease. Weaker correlations in variables like Sex and several others imply fainter linear relationships with Heart Disease, warranting further investigation for clearer insights or causal inferences. We decided to remove other variables such as Exercise, Other cancers, Weight, Height, Alcohol Consumption, Vegetable Consumption that show weaker correlations (close to 0), indicating a very weak linear relationship with Heart Disease. We chose to partition the data into a 60:40 ratio, reducing the sample size from 300,000 to 187,000 observations. Following this, we created 5 dummy variables to conduct our analysis. This sequential approach was adopted to streamline our modeling process and derive comprehensive insights from the dataset.

**Logistic Regression**

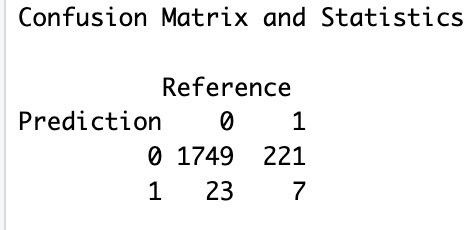
Logistic regression provides a good overview of what to expect in terms of what variables create the most impact on predicting who had a heart attack and who did not. By analyzing each coefficient, we can determine if it had an impact based on if the coefficient’s value was closer to negative 1 or positive 1. If a coefficient did not have much impact, it would be closer to null value. This is because in the equation we developed the variables are correlated based on their value in the coefficient and each are determinants of the heart disease predictor. It is also notable to mention that all variables are predicted equally without any extra weight given. By analyzing which values furthest from the null, we discovered that Age Above 60, General Health, Smoking History, Sex, Diabetes, and Exercise contributed most to heart disease. On the other hand, “alcohol consumption, and “depression” did not significantly contribute to the prediction of the presence of heart disease.

With further analysis into the summary of the model we also discovered that the logistic model provided an accuracy of 88.8%, as well as a false positive rate of 12.84%, we also had a limited range of false negatives and true negatives which meant most predicted values were indicated to be correct or falsely correct. This means that our model should predict individuals who are at higher risk as opposed to not predicting them at all putting them at higher risk of heart disease. Listed below are the specific values of each coefficient and their impact.

**K-Nearest Neighbors**

We decided to utilize the K-Nearest Neighbors model because while it’s one of the more basic models, it can also prove to be one of the best-performing models for certain datasets. To begin, we transformed the data to allow the values of each attribute to be more easily read by the KNN model. We transformed some of the variables into binary values, as well as took variables that had their values in groups (i.e. the Age Category variable) and transformed them from an unstructured format to a structured and numerical value. After transforming the data, we then ran the KNN model with k=5, to show how new data points would be compared to their five closest neighbors with respect to the variables. After transforming the variables, we then partitioned our dataset into training and validation sets at a 60/40 ratio.

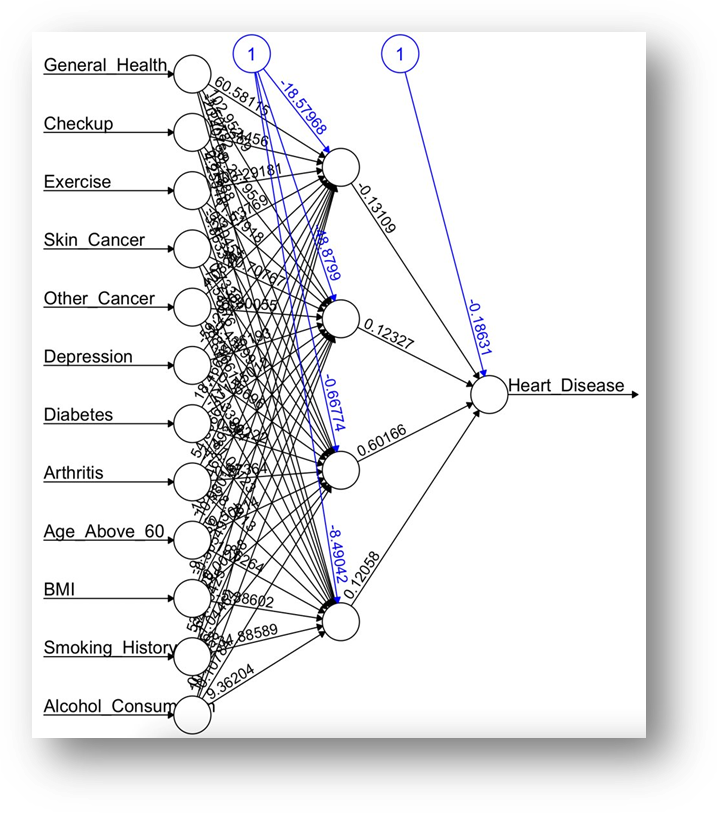
Our KNN model provided us with an impressive 87.8% accuracy rate along with an optimistic 1.3% false positive rate. For our project, we wanted to focus on the false positive rate because we felt it was of greater importance to limit the number of future patients who would be incorrectly predicted to not have heart failure when, in fact, they would. It’s of note that our original dataset contained an overwhelming rate of positive values for the heart disease variable, so we believe that this led our model to predict this to be positive for a large percentage of our KNN model.



\*Confusion Matrix For Our KNN Model

**Neural Network**

Neural networks excel in heart disease prediction, leveraging their adaptability to discern intricate patterns in complex medical datasets. Their ability to capture non-linear relationships and automatically extract relevant features makes them well-suited for the nuanced nature of heart disease analysis. The model's continuous refinement through iterative learning enhances accuracy, crucial for early intervention. In terms of formulation, a neural network consists of layers with nodes and weighted connections. The input layer receives features, processed through hidden layers via weighted sums and activation functions, culminating in the output layer's prediction. The model refines predictions by adjusting weights during training, blending linear algebra with nonlinear transformations for robust pattern recognition.

Our model incorporates twelve input nodes, four hidden nodes, and one output node, a design aimed at mitigating overfitting and complexity. Throughout the learning process, weights and biases are updated, enabling the network to adapt by fine-tuning these parameters. Positive weights denote causal connections between neurons, while negative weights signify inhibitory connections. Notably, our neural network model identified checkup, other cancers, smoking history, and alcohol consumption as prominent variables influencing the outcome.

**Results**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Regression** | **K-Nearest Neighbors** | **Neural Network** |
| **Accuracy** | **88.8%** | **87.8%** | **80.8%** |
| **Sensitivity** | **99.8%** | **98.7%** | **89.3%** |
| **Specificity** | **0.45%** | **3.1%** | **16.9%** |
| **False Positive Rate** | **12.84%** | **1.3%** | **11.1%** |

**Logistic Regression**: Results continue to vary along a 1-3% margin of error and especially considering different sample sizes as our original dataset was over 100,000 records. It has been discovered that logistic regression is most likely the best model used for accurate prediction of the presence of heart disease.

**K-Nearest Neighbors**: While this model has demonstrated comparable accuracy to our logistic regression model, it is important to note that our initial data set had a substantial number of positive values. This abundance may have influenced the model to lean towards more positive predictions. Despite logistic regression boasting the highest accuracy, this model exhibits a notably low false positive rate, suggesting potential for improvement.

**Neural Network**: As this model is unsupervised and incorporates random initialization and seed, the results varied with each run. The model presented in this report reflects the outcome when the weights stabilized, and changes became minimal. While it demonstrated reasonable accuracy and sensitivity, it lacks insights into the relationship between predictors and outcomes, and it exhibits a high false positive rate.

**Summary**

Based on an overview of the results of our various models, we feel confident recommending the logistic regression model to Beaumont as the optimal model to predict future heart failure in new patients. We have based this decision on several factors including accuracy rates and false positive rates. As we stated earlier in this report, we felt it was necessary to put more weight on the false positive rate as we do not want to incorrectly predict that a person is not at risk of heart failure when they actually are. These consequences would be much more severe than the opposite, which would be telling someone they’re prone to heart failure when they aren’t. While we would recommend the logistic regression model at this time, as we move forward, we would continue to look into fine tuning how we collect our data as well as aim to improve our data integrity, at which time we may find that another model would be better suited for our data.