Machine learning model predicting Cardiovascular disease

Mapalo Kasapo

July 31, 2023

1 Abstract

This project sets out to build a machine learning model that can predict whether or not someone has cardiovascular disease based on their lifestyle choices. To achieve this, we used different machine-learning techniques such as logistic regression and neural network models. This dataset was taken from Kaggle [4]. It has 70,000 participants, whose lifestyles are assessed to see their probability of contracting cardiovascular diseases. Our highest performing model was the neural network with a final accuracy of 73.89 percent.

2 Background

According to the World Health Organization, Cardiovascular disease is estimated to take the lives 17.9 million people globally each year; it is the leading cause of death in the world, but most cardiovascular diseases can be prevented by changing behavioral risk factors, like smoking, physical inactivity, and poor diet [3]. All these are included as features in this dataset, which makes it even more possible to predict if someone has cardiovascular disease based on their health and behavioral attributes.

3 Materials, Methods

3.1 Dataset Description and Data Cleaning

This dataset was taken from Kaggle[4]. It has 70000 participants. The dataset contains numerical features like physical activity, whether or not they smoke, and alcohol intake, and categorical features like height glucose levels, and weight, just to name a few. It is split into two groups, one for males and the other for females. This dataset was chosen because it contains information about people's lifestyles, and so it is highly effective for this project. It did not contain any missing values, so no data cleaning was necessary.

3.2 Exploratory Dataset Analysis

The following are the main findings from this dataset

- Fifty percent of the respondents in the dataset have cardiovascular disease.
- Only 35 percent of the respondents are male while 65 percent are female.
- Fifty percent of the women in the dataset have cardiovascular disease while 51 percent of the men have cardiovascular diseases.
- One important finding is that 61 percent of respondents who were active did not have cardiovascular diseases, showing that physical inactivity is positively correlated with a higher risk of contracting cardiovascular disease.

Figure 1 shows the distribution of weight among people in this dataset. Since more men have cardiovascular diseases compared to women, the overlaid histogram below illustrates that an increase in weight may be related to more cases of cardiovascular disease. To further explore this relationship, we plotted a graph for weight against cardiovascular diseases. Figure 2 shows two

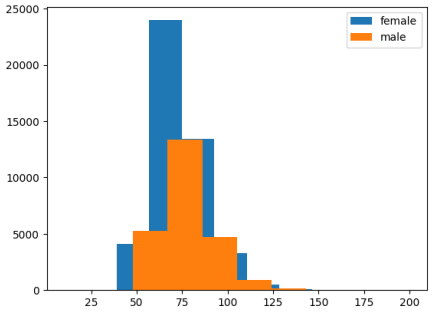

Figure 1: Weight of men and women in Kilograms

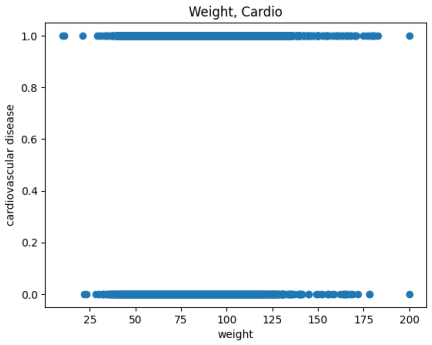

Figure 2: Weight against cardiovascular disease

possibilities of one and zero for either having or not having cardiovascular disease. This graph *mildly* demonstrates that the number of people without cardiovascular diseases who weigh above 130 KG reduces. This means that there is a mild, positive correlation between weight gain and cardiovascular disease.

3.3 Modeling

To model the relationship between behavioral change and cardiovascular disease, we used logistic regression and neural network modeling. These were used because we can model a binary output - making it easier to distinguish between those that have and those that do not have cardiovascular disease.

3.3.1 Logistic Regression

Logistic regression estimates the probability of an event happening based on a given dataset of independent variables [2]. In this project, logistic regression was used to determine the probability of an individual having or not having cardiovascular diseases based on their behavioral patterns. We first defined the features in the dataset using X and y: X being all the factors that were being assessed and y the target, cardiovascular disease. All the models had a train/test split of 70/30, respectively. Then, we experimented with different combinations of features in X to see which ones provide us with the highest accuracy. Through this, we found that the best model used the following features: *Cholesterol*, *Glucose*, *Smoking*, *Aphi (Systolic Blood Pressure)*, and *Alcohol Consumption*. Our final test accuracy was 72.15 percent.

Figure 3 is a confusion matrix. It shows how the logistic regression model performed by comparing its predictions with the actual state of an individual either having or not having cardiovascular diseases. According to the figure, the model correctly predicted whether or not someone had cardiovascular disease 72.15 percent of the time. 11.76 percent of the predictions are false positives while 16.09 percent are false negatives.

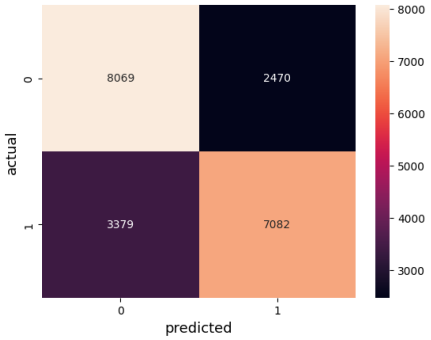

Figure 3: Confusion matrix showing the distribution of predictions of the logistic regression model and the actual value of cardiovascular disease.

3.3.2 Neural Network

Neural networks operate in a similar way as the brain as they signal information from a node layer, which contains an input layer, through hidden layers, to an output layer. They rely on training data to learn and improve their accuracy over time [1]. In this model, we used all available features for X, except the *Cardiovascular Disease* column which was the y. We had a train-test split of 70/30. The selected model has three layers. In the first layer there are 250 nodes, 110 in the second one, and 35 in the third. The output layer has 2 nodes, one for each possible class. It was trained over 15 epochs.

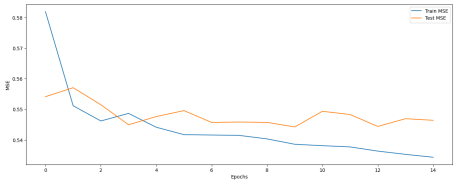

Figure 4: Accuracy over epochs

Figure 4 shows accuracy over time. The training accuracy keeps reducing over the epochs while it sees little decline for the test. This model has a final test accuracy of 73.89 percent. The confusion matrix in Figure 5 shows that 73.17 percent of the predictions were accurate, 14.15 percent are false negatives, and 12.68 percent are false positives.

4 Discussion

False negatives mean that the machine learning system predicts that a person does not have cardiovascular disease when they actually do. False positives, on the other hand, occur when the model predicts that someone has cardiovascular disease when they do not. False positives can

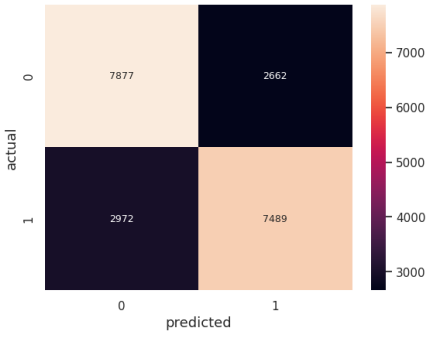

Figure 5: Confusion matrix for neural network

make a patient start taking medication for a disease that they do not have. This can have some side effects but they can not be compared to worsening cases caused by lack of timely treatment. Therefore, it is more dangerous to have a machine learning model with more false negatives than false positives.

4.1 Findings

Since the neural network model had a smaller percentage of false negatives than the logistic regression model, the neural network performed better. This reduces the possibility of individuals missing out on treatment due to a lack of awareness of the disease. Furthermore, the neural network had a higher percent of accurate predictions compared to the logistic regression.

Model	Test Acc.
Logistic Regression	72.15%
Neural Network	73.89%

Table 1: Summary of model performance

4.2 Project Limitations and Future Steps

The limitations of the project are as follows:

- The time to do the project was limited due to the fact that we would only meet twice a week.
- The dataset we used had no proper citation showing the source of the information and where the sample was from, so it is hard to figure out how reliable it is.
- Some of the variables in the Kaggle dataset, such as outliers in height and weight were unrealistic, so the quality of the dataset is not assured.

Possible furture steps include the following:

- Using this model in a real life situation, where real cardiovascular disease patients and other individuals are sampled to test the accuracy of the model.
- Furthermore, this model can be used to help predict peoples chances of contracting cardiovascular disease by using it in hospitals or in as an application on mobile phones.

4.3 Human Context and Ethics

If this model were put into medical practice, we should first think about the source of the information: which part of the world the dataset was from, and how responses were collected from the participants. This is important to know because it can give more clarity on what the specifics of the respondents are: if the sampled respondents were selected at random and not from similar groups of people if the countries that were sampled are high, middle, or low-income countries, and if non-communicable diseases like cardiovascular disease are prevalent in those countries. It would not be a good idea to use this model without finding these things out first because the data could be giving biased results due to a limited population and lack of diversity in sampling.

References

[1] International Business Machines. What are neural networks?

[2] International Business Machines. What is logistic regression?

[3] World Heath Organisation. Cardiovascular diseases. 2021.

[4] Svetlana Ulianova. Cardiovascular disease dataset. 2019.