HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Applying Machine Learning to Heart Disease Risk Analysis

Sanaa Bhorkar[1] and Guillermo Goldsztein[#]

[1]The Harker School, USA
[#]Advisor

ABSTRACT

Heart disease is known to be deadly; about 700,000 Americans die from it annually. It is essential for young adults to calculate their risk early on, as many factors could lead to a high probability of heart disease. Some factors, such as diet or weight are obvious, but several others, such as mental health, hours of sleep, and even race, can be crucial indicators. This is where machine learning can help, as AI models can look at a large number of factors together to calculate risk by drawing upon real-world data. This paper discusses the implementation of one such model. The model described below is trained on data from about 320,000 patients in the United States. The model uses a Sequential Neural Network algorithm and was created using TensorFlow. It analyzes each patient's data and calculates their risk for heart disease, performing reasonably well with a 77% accuracy. This model can replace a linear-regression-based model, such as the ASCVD Risk Estimator, as it accounts for more factors in its analysis. Further testing with new and diverse datasets and different algorithms, like a RandomForestClassifier, could potentially improve this model.

Introduction

Machine Learning has recently become one of the most widely used data analysis and forecasting methods. It has been used in several industries to make critical decisions and optimize processes such as stock trading and medical diagnosis. Machine learning uses algorithms to identify patterns in data and uses those patterns to create a model that makes predictions. With increased data and experience, the results of machine learning increase in accuracy, similar to humans improving with more practice. Machine learning is instrumental in scenarios where data or the nature of the task is constantly shifting or coding a solution would be impossible (Alpaydin, 2010; Guido, 2016; Mohri et al., 2012; Murphy, 2013).

Calculating Heart Disease Risk Using Machine Learning

The nature of machine learning makes it useful in several industries, but this paper focuses on the health industry, specifically in estimating heart disease risk factor. Unlike illnesses like pneumonia or injuries like a broken ankle, heart disease cannot be immediately diagnosed through a visual image; other data is required for diagnosis, ranging from blood tests to lifestyle characteristics. It is vital to keep track of these characteristics as the risk factor of this deadly disease can continue to increase over the years of a person's life. In the United States alone, about 700,000 people die of heart disease, almost one in every five deaths (New York Department of Health, 2022). These unfortunate numbers only increase the need for an application to help adults track their risk. This paper focuses on building such a model and the results that come with it.

Calculating heart disease risk is a complicated process; it cannot be done only with a simple weighted average or numerical data. This is where machine learning can intervene and optimize this process, as various factors such as mental health and race could affect a person's risk for heart disease. Organizations like the South

Asian Heart Center exist to help account for and combat these effects and help their patients reduce their risk. These organizations can also help with other factors, such as mental health and physical activity. Therefore, a simple linear regression/weighted average model cannot be used; a classification model would work as it can account for these several factors and accurately assess the risk.

Dataset Used in the Model

The first step in creating a machine learning model is finding the right dataset. The dataset used in this model is below, split into Tables 1a and 1b. This dataset comes from a general research study with the CDC (Center for Disease Control) in 2020 for about 320,000 adults (Centers for Disease Control and Prevention (CDC), 2020). While the study was not specifically related to heart disease, the specific data collected does impact heart disease diagnosis in adults.

Table 1a. Healthcare dataset obtained from the CDC with the general features

Patient	Heart Disease	Smoking	Alcohol Drinking	Mental Health	Sex	Age	Race	Sleep Time	Physical Health	BMI
1	No	Yes	No	30.0	F	55-59	White	5.0	3.0	16.6
2	No	No	No	0.0	F	80+	White	7.0	0.0	20.34
3	No	Yes	No	30.0	M	65-69	White	8.0	20.0	26.58
4	No	No	No	0.0	F	75-79	White	6.0	0.0	24.21
5	No	No	No	0.0	F	40-44	White	8.0	28.0	23.71

Table 1b. Healthcare dataset obtained from the CDC with the health features

Patient	Heart Disease	Kidney Disease	Skin Cancer	Diabetic	Asthma	GenHealth	Stroke	Difficulty Walking	Physical Activity
1	No	No	Yes	Yes	Yes	Very good	No	No	Yes
2	No	No	No	No	No	Very good	Yes	No	Yes
3	No	No	Yes	Yes	Yes	Fair	No	No	Yes
4	No	No	No	No	No	Good	No	No	No
5	No	No	No	No	No	Very good	No	Yes	Yes

Tables 1a and 1b show the dataset used in this ML model. The full dataset consists of information from about 320,000 patients. Information about only five patients is shown.

One category of machine learning problems is called supervised learning. The data in supervised learning problems consists of a collection of examples that is already labeled by a human. In this case, the examples

are the patients. The information about each example is split into features and labels. With this dataset, the label is whether the patient has heart disease, and the features are every other column, ultimately determining whether the patient has heart disease. This dataset has 17 features: gender, age category, race, number of days in a month with poor mental health, number of days in a month with poor physical health, BMI (body mass index), diagnosis of skin cancer, diagnosis of kidney disease, diagnosis of asthma, current diagnosis of diabetes, general health, hours of sleep per day, physical activity, difficulty walking, personal history of stroke, over 7 or 14 alcoholic drinks per week (depending on gender), and whether or not the patient smokes regularly.

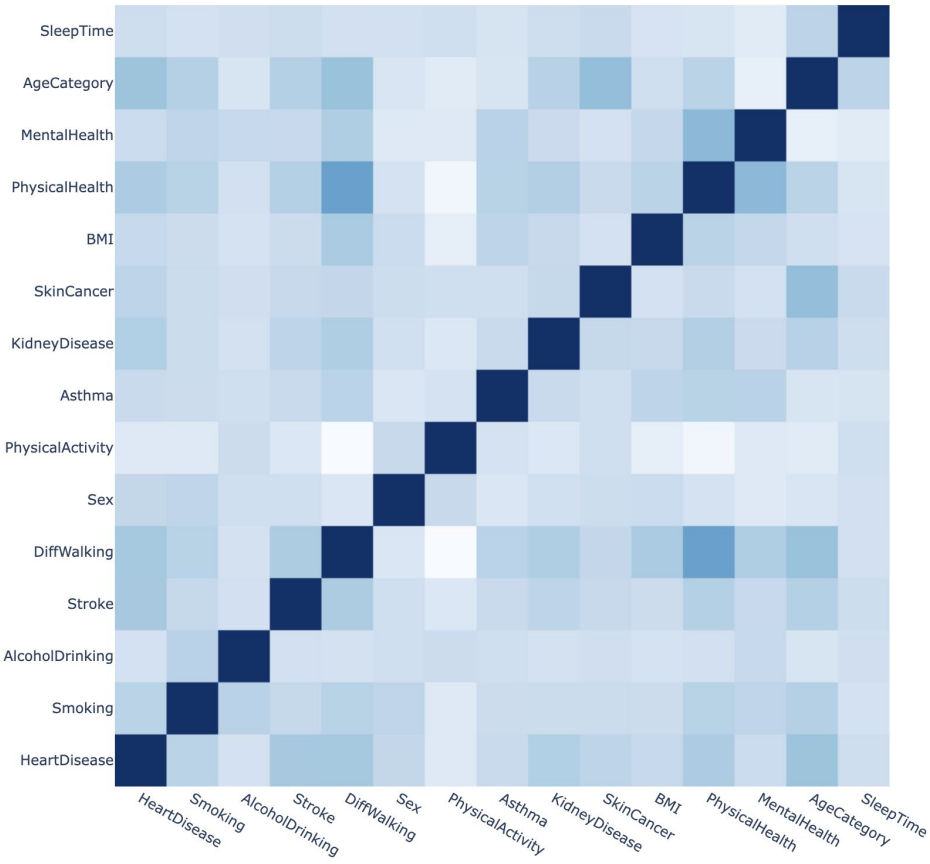

Figure 1. A correlation matrix that shows how different data set features relate to others

Figure 1 depicts a correlation matrix for the dataset. As the features become more correlated with one another, their associated square's color becomes a darker shade of blue. This shows which features are the most connected and have similar values. For example, 'DiffWalking' (difficulty walking) and 'PhysicalHealth' are the most correlated features. This means most patients with difficulty walking are in poor physical condition. The complexity of the correlation matrix shows why it is impractical to use linear regression for this case.

The dataset that is used to develop the model is called the training set, and the examples of this set are called training examples. Both the features and the labels are known for all the examples in the training set. The training set is used to "train" an ML model. The purpose of having this model is to use it later to predict the labels of new examples when only the features are known. Once the model is created, it will be used to predict the risk that a patient has heart disease or not.

Data Preprocessing

Data preprocessing is an important step in creating an ML model. The data must be as straightforward as possible to ensure the model understands the data it is receiving. Different datasets require different steps, but the specific one used in this model required replacing strings with numbers and hot encodings.

In the column "gender," "male" by 0 and "female" are replaced by 1. The same applies to eight other columns: heart disease, smoking, alcohol drinking, asthma, skin cancer, kidney disease, diabetes, and physical activity. In the "age" column, there are several categories, such as 40-44 or 65–69. These categories are replaced with the average age in each set; in our example, 42 and 67, respectively.

The next step is to apply hot encodings to each feature that requires it. For example, for each patient, the "race" feature has one of six possible values: "White," "Asian," "Hispanic," "American Indian/Alaskan Native," "Black," and "Other." One-hot-encoding is applied to this feature. More precisely, the feature "race" is replaced by six new features: "White," "Asian," "Hispanic," "American Indian/Alaskan Native," "Black," and "Other." If the "race" of a patient was "White," then the value of "White" for this patient is 1 and 0 for the other five races.

The model uses these features and returns a numerical prediction using a method known as binary classification. In binary classification, the labels of the examples take either the value 1 or the value 0. In our dataset, 1 corresponds to the patient having heart disease, and 0 is the opposite. Once the model has analyzed a patient's data, it returns a number between 0 and 1.

The data set has been split to train the model, with 80% of patients acting as training data and the remaining 20% as testing data. Different computational models will be developed using only the training set. The validation set will be used only to evaluate the performance of the models.

It is important to note that this specific dataset is not balanced between patients with heart disease and patients without it. Only 27,400 patients (9% of the data set) have heart disease, while the remaining 292,000 do not. Unfortunately, this can cause issues in the model, as it can become biased or skewed. The best and most ethical way to fix this issue is to obtain more data to increase the ratio from only 9% to at least 35%.

Creating the Model

The binary classification model is developed using a machine-learning technique called logistic regression. Logistic regression develops a probability of heart disease for each patient, which can then be rounded to 0 or 1. Part of the code for this application is below, from which the exact model, functions, and parameters used in this analysis are seen.

```
model = Sequential()
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy')
model.fit(X_train_scaled,y_train,epochs=20,verbose=0)
J_list = model.history.history['loss']
plt.plot(J_list)
```

The so-called "Accuracy" measures the performance of the model. For binary classification, the accuracy is calculated by looking at the ratio of total number of correct predictions, whether positive or negative, over the total number of predictions. The accuracy varies from 0 to 1. The accuracy with a value close to 1 indicates that the model works well. In the opposite case, where accuracy close to 0 means the model has a poor performance.

This model's training dataset has been trained at different epochs, or complete passes, through the data set. With only five epochs, the loss, or errors of the model, is 0.53009456, which is about 0.47 accuracy. At 20 epochs, the accuracy increases significantly to 0.77. The precision and recall are also slightly higher than the model that ran for 20 epochs. However, the number of epochs cannot be exorbitant, as there is a possibility of overfitting the training set.

Overfitting is a term that refers to a modeling error that occurs when a function corresponds too closely to a particular set of data. As a result, overfitting may fail to fit additional data, affecting the accuracy of predicting future observations. Overfitting can be identified by checking validation metrics such as accuracy and loss. If the model performs about the same on the validation dataset as on the training dataset, it is unlikely to have an overfitting issue in the model. In this investigation, the trained model performs about the same on the two datasets, the training dataset and the validation dataset. Therefore, there should be no serious overfitting problem in the model trained in this data analysis to identify potential patients with heart disease.

Results

After running this model for 20 epochs, TensorFlow returned with an accuracy of about 77%. The testing data given to the model was taken from the original data set, a study done by the CDC. The testing data, about 20% of the complete data set, yielded 63,959 patients. The model's accuracy means it could accurately classify about 48,600 patients using physical characteristics and other essential metrics mentioned earlier.
Below is a confusion matrix to visualize the data:

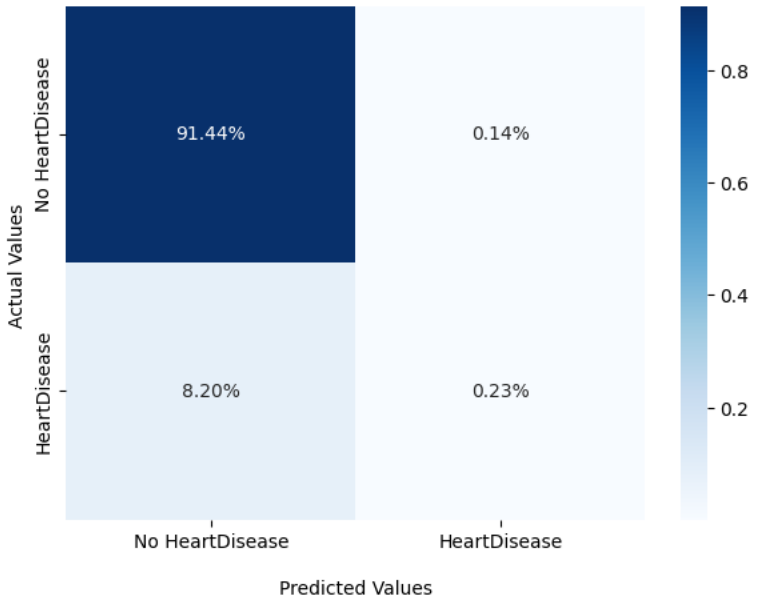

Figure 2. A confusion matrix that shows the results of the model

There are two separate accuracies to consider within this confusion matrix: accurately predicting that the patient has heart disease and accurately predicting does not have heart disease. One can find these accuracies by dividing the true positive value, matching 'Actual Values' and 'Predicted Values,' by the sum of the true positive and false negative values. In the case of predicting heart disease, the accuracy is 62% (0.23 divided by the sum of 0.14 and 0.23), while the accuracy of predicting no heart disease is about 92% (91.44 divided by the

sum of 8.20 and 91.44). The true accuracy is found by averaging these two values, leading us to a true accuracy of 77%, validating the error given by the TensorFlow.

Discussion

As the 'Results' section identifies, the data imbalance affects the true accuracy. This data imbalance resulted in a skew, as the model is clearly biased towards 'No Heart Disease' and, therefore, could inaccurately analyze a patient as one without heart disease. This skew can be minimized in two ways: using a more balanced dataset and/or trying different algorithms like a RandomForestClassifier.

This model can be useful in real-world applications. An example of such application is integrating it with an AI chatbot, implemented using GPT, to solicit user input about the features required by the model and output their risk of heart disease.

Overall, this model has a reasonably high accuracy and does a decent job of analyzing patient data and accurately predicting their risk.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

Alpaydin, E. (2010). *Introduction to Machine Learning*. MIT Press.

Centers for Disease Control and Prevention (CDC). (2020). Behavioral Risk Factor Surveillance System Survey Data.

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Guido, S. (2016). *Introduction to Machine Learning with Python*. O'Reilly Media.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

Murphy, K. P. (2013). *Machine Learning: A Probabilistic Perspective*. MIT Press.

New York Department of Health. (2022, August). Heart Disease in the United States. Heart Disease and Stroke Prevention. https://www.health.ny.gov/diseases/cardiovascular/heart_disease/