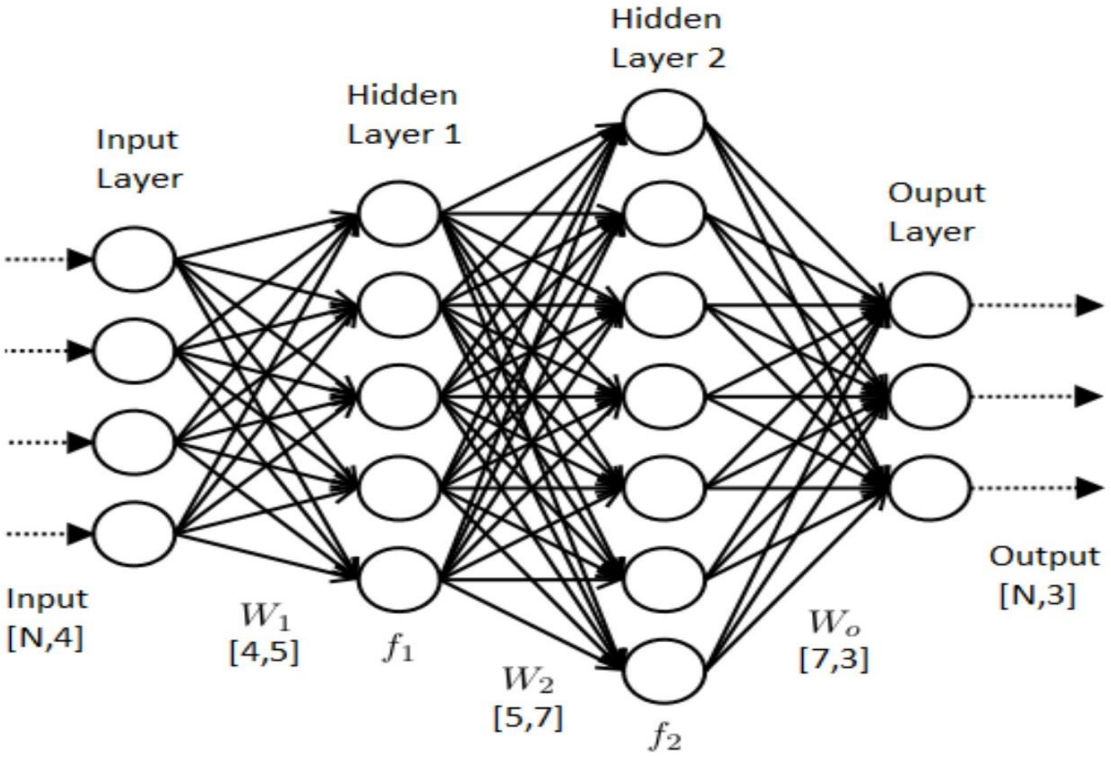
2021

Heart Disease Prediction

PROJECT 2021

WANDEY B. ADEYEYE

ANN | 21



CARDIOVASCULAR DISEASE CLASSIFICATION

BY:

WANDE B. ADEYEYE

E-MAIL:

ADEYEYEWANDE@GMAIL.COM

Project Overview

Aim of the project is to detect the presence or absence of cardiovascular disease in person based on the given features.

The dataset consists of 70 000 records of patients' data, 11 features + target. Features available are:

Age (Objective Feature)	Age of individual patients int (days)
Height (Objective Feature)	Height of individual patients int (cm)
Weight (Objective Feature)	Weight of individual patients in float (kg)
Gender (Objective Feature)	Gender of the patient (categorical code)
Systolic blood pressure (Examination Feature)	ap_hi (int)
Diastolic blood pressure (Examination Feature)	ap_lo (int)
Cholesterol (Examination Feature)	Cholesterol level ranging from 1 to 3 (1: normal, 2: above normal, 3: well above normal)
Glucose (Examination Feature)	Glucose level ranging from 1 to 3 (1: normal, 2: above normal, 3: well above normal)
Smoking (Subjective Feature)	smoke (binary)
Alcohol intake (Subjective Feature)	alco (binary)
Physical activity (Subjective Feature)	active (binary)
Presence or absence of cardiovascular disease (Target Variable)	cardio (binary)

NOTES ON BLOOD PRESSURE

- **Blood Pressure notes:**

    i. Blood pressure is represented by 2 numbers systolic and diastolic (ideally 120/80 mm Hg).

    ii. These two number are critical in assessing the heart health.

    iii. The top number represents **systolic** and the bottom number representing the **diastolic**.

    iv. Systolic pressure indicates the blood pressure in the arteries when the blood is pumped out of the heart.

    v. The diastolic pressure indicates the blood pressure between beats (at rest, filling up and ready to pump again).

    vi. If these numbers are high, that means that the heart is exerting more effort to pump blood in the arteries to the body.

NOTES ON CHOLESTEROL

- **Cholesterol notes:**

    i. Cholesterol is a waxy material found in human's blood.

    ii. Normal level of cholesterol is necessary to ensure healthy body cells but as these levels increase, heart disease risk is elevated.

    iii. This waxy material can block the arteries and could result in strokes and heart attacks.

    iv. Healthy lifestyle and regular exercises can reduce the risk of having high cholesterol levels.

    v. More information here

NOTES ON GLUCOSE

- **Glucose notes:**

    i. Glucose represents the sugar that the human body receive when they consume food.

    ii. Glucose means "sweet" in Greek.

iii. Insulin hormone plays a key role in moving glucose from the blood to the body cells for energy.

iv. Diabetic patients have high glucose in their blood stream which could be due to two reasons:

   i. They don't have enough insulin

   ii. Body cells do not react to insulin the proper way

v. Read more here

With that been said, we will build our artificial neural network using TensorFlow, and Keras as API. However, first we need to perform some exploratory data analysis to understand our dataset properly and to put our dataset in good shape to get the best accuracy possible with our model

- The dataset contains 11 features and a target
- The age column is in days but was converted to years
- The dataset contains no null values
- The dataset correlations are not strong (direction and strength) can affect the model accuracy

BUILD AND TRAIN AN ARTIFICIAL NEURAL NETWORK MODEL

- split the dataframe into target and features
- Feature Scaling to find the appropriate weights(w) for each feature.
- Split the data into test and train sets

After we've successfully implemented the above three procedures, it's time to build our artificial neural network model.

We start by adding hidden layer to our network and specify how many neural par layer using 'tf.keras.models.Sequential()' as classifier then add a dense layer and specify the number of neural, activation function 'relu' or Rectified Linear Unit, and input shape which is equals to '11' in this case after we add a dropout layer to try to ensure the network generalization capabilities. The Dropout layer randomly sets input units to '0' with a frequency of rate at each step during training time, which helps prevent overfitting.

After that has be done, we will add two more dense layer that contains '400' neural each and in the output, we will use one unit of output, because our output will either be '0' or '1', that's why we used sigmoid activation function instead of 'relu' because we want the output to be saturated

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_4 (Dense)              (None, 400)               4800
_________________________________________________________________
dropout_1 (Dropout)          (None, 400)               0
_________________________________________________________________
dense_5 (Dense)              (None, 400)               160400
_________________________________________________________________
dense_6 (Dense)              (None, 400)               160400
_________________________________________________________________
dense_7 (Dense)              (None, 1)                 401
=================================================================
Total params: 326,001
Trainable params: 326,001
Non-trainable params: 0
_________________________________________________________________
```

From the above output, here is the summary of our model:

- It consists of 326,001 parameters
- Dense layer and parameters
- Dropout layer and parameters
- Output layer and parameters

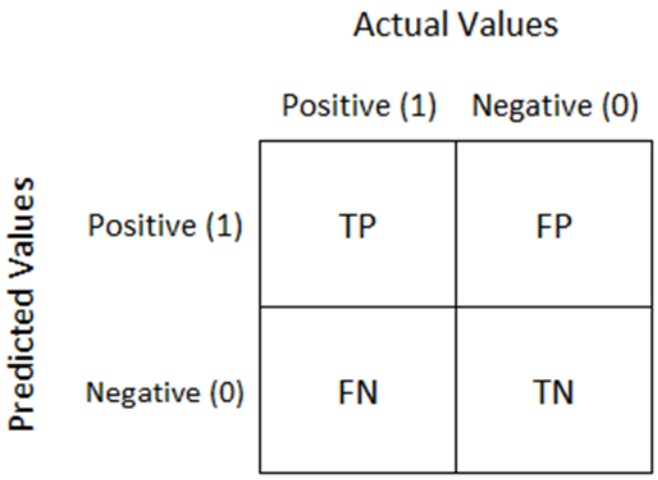
If we sum up the total number of the parameters it will add up to the number of parameters.

Next, we will compile our model and specifier optimizer as ‘Adam’, loss as 'binary_crossentropy', and metrics = ['accuracy'] after that we can fit our classifier with train dataset and specifier the number of epochs which is a hyperparameter that defines the number of times that the learning algorithm will work through the entire training dataset. ANN learn through experience, learn over time similar to humans, so every time we take the data and feed it to the network then we update the weights, we go back and repeat again and we keep repeating over and over again until the number of epochs, it will take couple of minutes depending on the number of epochs.

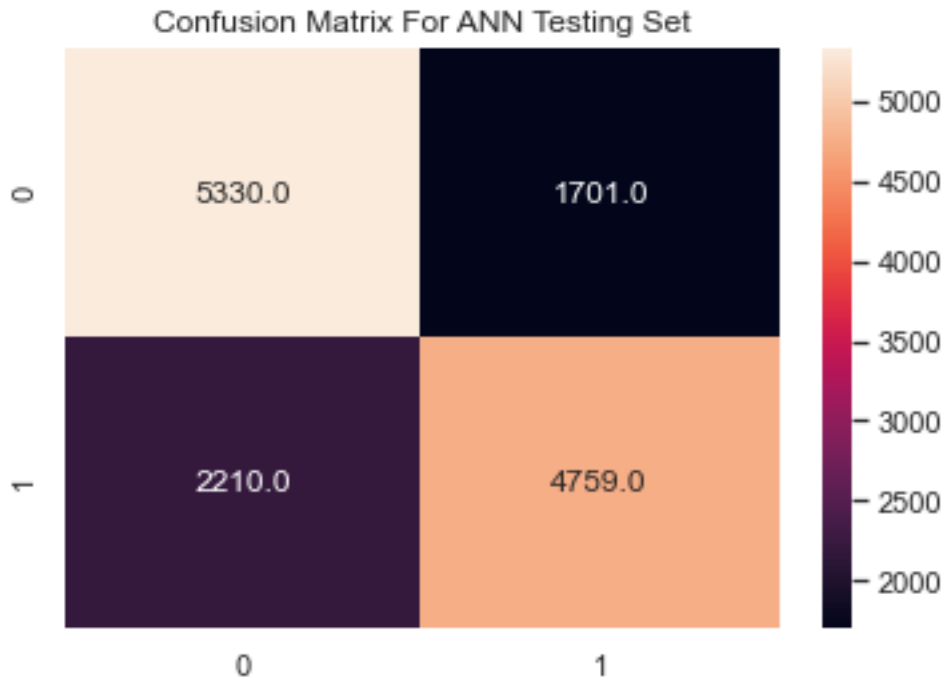
After our network is trained, we will take our classifier model apply predict method and pass the testing data which the model hasn’t seen yet, so then the model will generate ‘y_pred’ and then set a **threshold** of **0.5** which means any value that’s above **0.5** belongs to class ‘1’ and any value less belongs to class ‘0’

EVALUATE THE ARTIFICIAL NEURAL NETWORK MODEL

The evaluation metric used is the confusion matrix. The confusion matrix displays the correctly predicted as well as incorrectly predicted values by a classifier. The sum of TP and TN, from the confusion matrix, is the number of correctly classified entries by the classifier. As shown in the figure below:



Similarly let us look at the confusion matrices for ANN below:



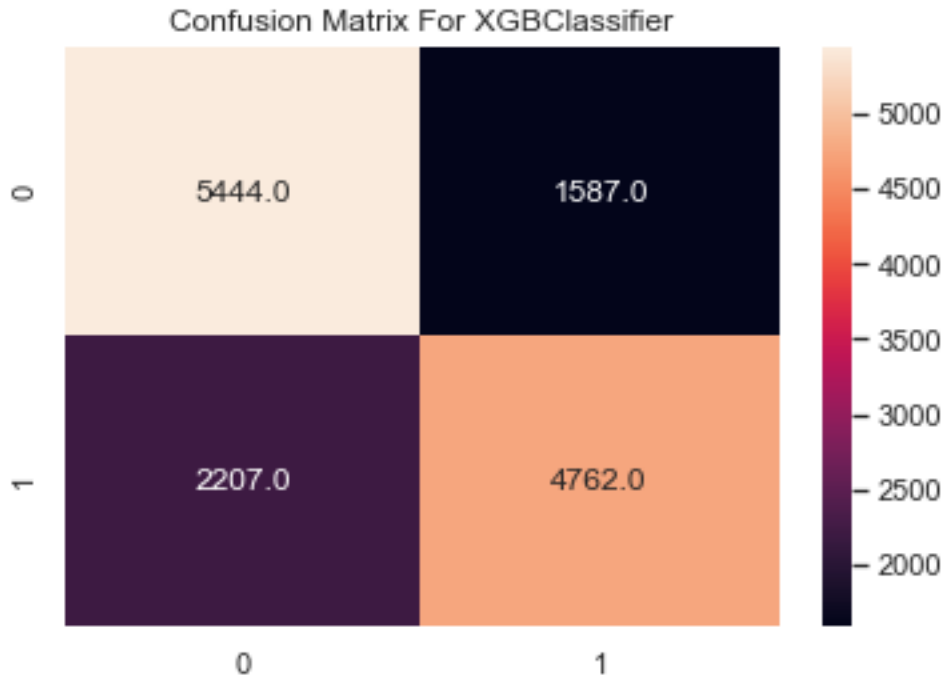
Accuracy for ANN testing set = ((5330+4759)/(2210+1701+5330+4759))*100 = 72.06%

Accuracy for ANN for test set = 72%

72% accuracy is quite strong base on our data, the data overlap on each other and can be difficult to separate, when we look for correlation, we focus basically on two things, the strength and the direction of the correlation and both seems to be week on this dataset that's why we end up with 72% accuracy.

Next, we will try XGBoost classifier to see maybe we can get a higher accuracy.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data.



Accuracy for XGBoost training set = ((5444+4762)/(2207+1587+5444+4762))*100 = 73%

Accuracy for XGBoost for test set = 73%

This algorithm performs slightly better than the ANN model

We see that the XGBoost gives us a higher accuracy for the test set

Conclusion

Heart Disease is one of the major concerns for society today.

It is difficult to manually determine the odds of getting heart disease based on risk factors. However, we can harness the advanced in technology today such as machine learning techniques that are useful to predict the output from existing data. AI can help clinicians to make more accurate predictions for patients

Useful links:

Dataset link here

More info on cholesterol here

More info on glocuse here

My Github profile here