Udacity Machine Learning Nanodegree 2019

Capstone Proposal

Classifying Heart Disease Using Neural Networks

Manish Garg

August, 2019

[1 Domain Background 3](#_Toc15124336)

[2 Problem Statement 3](#_Toc15124337)

[3 Dataset and Inputs 3](#_Toc15124338)

[4 Solution Statement 5](#_Toc15124339)

[5 Benchmark Model 5](#_Toc15124340)

[6 Evaluation Metrics 6](#_Toc15124341)

[7 Project Design 6](#_Toc15124342)

[8 References 8](#_Toc15124343)

# Domain Background

Heart diseases are often underestimated, but, in reality, they are the leading cause of death

in the world. Among them, coronary artery disease (CAD) accounts for about a third of all

deaths worldwide in people over 35 years of age. CAD is the result of arteriosclerosis,

which consists in the narrowing of the blood vessels and the hardening of its walls. In some

cases, CAD can completely block the influx of oxygen-rich blood to the heart muscle,

causing a heart attack.

CAD is caused by an accumulation of waxy grease deposits on the inner walls of the

arteries. These deposits consist of cholesterol, calcium, and other substances that travel in

the blood; the product of their accumulation is called atherosclerotic plaque. This plaque

can clog the coronary arteries and make them rigid and irregular, causing the so-called

hardening of the arteries or atherosclerosis. These obstructions can be single or multiple

and present various levels of gravity and different locations. Gradually, the deposits restrict

the lumen of the coronary arteries, thus reducing the supply of blood and oxygen to the

heart muscle. This reduction in blood flow can cause chest pain (angina), difficulty in

breathing (dyspnoea), and other symptoms, while complete obstruction can induce a heart

attack.

Coronary angiography is used to diagnose CAD. Angiography is the diagnostic

representation of the blood or lymphatic vessels of the human body through a technique

that involves the infusion of a water-soluble contrast agent within the vessels and the

generation of medical images through various biomedical imaging techniques.

# Problem Statement

The main objective of this project will be to use Neural Networks techniques try to predict a condition of heart disease through a classification Algorithm.

Similar problem and solutions approach were cited in the *Journal of Intelligent Learning Systems and Applications*. Check the reference sections for more details.

# Dataset and Inputs

For this project we will use the Heart Disease Data Set, which is available in the UCI Machine Learning Repository.

*Source:*

[*https://archive.ics.uci.edu/ml/datasets/Heart+Disease*](https://archive.ics.uci.edu/ml/datasets/Heart+Disease)

*file: processed.cleveland.data*

More specifically we will use data that was made available by the CCF (edited by Robert Detrano, MD, PhD). This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The goal is to predict the presence of heart disease in the patient. The target is an integer value from 0 (no presence) to 4.

Experiments with the Cleveland database have concentrated on simply attempting to

distinguish presence (values 1, 2, 3, 4) from absence (value 0). We will further refine the target to have 0(no presence) and 1(presence).

* *Number of instances to be used: 302*
* *Number of feature attributes to be used : 14 continuous attributes (including the class attribute*

*HeartDisease):*

* *Target class distribution: Cleveland: Class 0: 164, Class 1: 139*

The target class in dataset seems to be balanced and there are no missing attribute values.

As the dataset is small and we have balanced target class while splitting the data into train and test we will use StratifiedShuffleSplit.

[StratifiedShuffleSplit](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html) : This module creates a training/testing set having equally balanced(stratified) classes.

Only 14 attributes used:   
1. #3 (age)   
2. #4 (sex)   
3. #9 (cp)   
4. #10 (trestbps)   
5. #12 (chol)   
6. #16 (fbs)   
7. #19 (restecg)   
8. #32 (thalach)   
9. #38 (exang)   
10. #40 (oldpeak)   
11. #41 (slope)   
12. #44 (ca)   
13. #51 (thal)   
14. #58 (num) (the predicted attribute)

Each of the attributes are detailed as follows:

* age: Age in years
* sex: Sex (1 = male; 0 = female)
* cp: Chest pain type (Value 1: typical angina; Value 2: atypical angina; Value 3: non-anginal pain, Value 4: asymptomatic)
* trestbps: Resting blood pressure (in mm Hg on admission to the hospital)
* chol: Serum cholesterol in mg/dl
* fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
* restecg: Resting electrocardiographic results (Value 0: normal; Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria)
* thalach: Maximum heart rate achieved
* exang: Exercise induced angina (1 = yes; 0 = no)
* oldpeak: ST depression induced by exercise relative to rest
* slope: The slope of the peak exercise ST segment (Value 1: upsloping; Value 2: flat; Value 3: down sloping)
* ca: Number of major vessels (0-3) colored by fluoroscopy
* thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
* Heart Disease: Diagnosis of heart disease – angiographic disease status (Value 0: < 50% diameter narrowing; Value 1: > 50% diameter narrowing)—in any major vessel: attributes 59 through 68 are vessels

# Solution Statement

The proposed solution to this problem is to apply Neural Network techniques that have proved to be of high accuracy in classification.

The neural network will be **deep learning model**. A neural network takes in inputs, which are then processed in hidden layers using weights that are adjusted during training. Then the model spits out a prediction. The weights are adjusted to find patterns in order to make better predictions. The user does not need to specify what patterns to look for — the neural network learns on its own.

First, we will extract the required data and do exploratory analysis to understand how the data is distributed and extract preliminary knowledge. Then we will scale the feature values under different features so that they fall under a common range.

Once the data is scaled we will do train test split where 30% of the data is divided up as test

data. The next step will be to train a Neural Network with these data sets and make predictions. For workflow steps check the Project Design section.

I believe that this will be very effective at finding patterns and predict a condition of heart disease.

We will use the evaluation metrics described in later sections to compare the performance of these solutions against the benchmark models in the next section.

In case if we do not get the desired results with deep learning model we will try to use the Ensemble Learning technique – XGBoost.

XGBoost (Extreme Gradient Boosting) is an advanced implementation of the Gradient Boosting. This algorithm has high predictive power and is ten times faster than any other gradient boosting techniques. Moreover, includes a variety of regularisation which reduces overfitting and improves overall performance. We will use - xgb.XGBClassifier()

# Benchmark Model

The given dataset is a supervised classification learning problem for which we will first try with simple K-nearest neighbors.

KNN is an algorithm that's used in pattern recognition for object classification based on the characteristics of the nearest objects. An object is classified according to the majority of the votes of its neighboring k cluster.

The next step will be to train a Neural Network with these data sets and make better predictions.

# Evaluation Metrics

The performance of each classification model is evaluated using three statistical measures; classification accuracy, sensitivity and specificity. As this is medical problem so high recall will be more valuable than precision.

We will also summaries the performance of a classification algorithm using confusion metrics as calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

|  |  |  |
| --- | --- | --- |
|  | **Predicted - Positive** | **Predicted - Negative** |
| **Actual - Positive** | True Positive | False Negative |
| **Actual - Negative** | False Positive | True Negative |

* “True Positive” for correctly predicted event values.
* “False Positive” for incorrectly predicted event values.
* “True Negative” for correctly predicted no-event values.
* “False Negative” for incorrectly predicted no-event values.

# Project Design

The workflow of solving this problem will be in the following order:

* Exploring the Data
* Loading Libraries and data
* Peek at the training data
* Dimensions of data
* Overview of responses and overall response rate
* Statistical summary
* Data preprocessing/cleaning
* Preprocess feature columns
* Identify Feature and Target columns
* Data cleaning
* Feature Scaling - Standardization/Normalizing data
* Training and Validation data split
* Evaluate Algorithms
* Build models
* Select best model
* Make predictions on the validation set
* Feature importance and feature selection
* Model Tuning to Improve Result
* Final conclusion

# References

1. [*https://archive.ics.uci.edu/ml/datasets/Heart+Disease*](https://archive.ics.uci.edu/ml/datasets/Heart+Disease)
2. [*https://keras.io/*](https://keras.io/)
3. [*https://en.wikipedia.org/wiki/K-nearest\_neighbors\_algorithm*](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm)
4. [*https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/*](https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/)
5. [*https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62*](https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62)
6. *S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, “Innovative artificial neural networks-based decision support system for heart diseases diagnosis,” Journal of Intelligent Learning Systems and Applications, vol. 5, no. 3, pp. 176–183, 2013.* [*http://www.scirp.org/journal/paperinformation.aspx?paperid=35396*](http://www.scirp.org/journal/paperinformation.aspx?paperid=35396)
7. [*https://www.ahajournals.org/doi/full/10.1161/CIR.0b013e31820a55f5*](https://www.ahajournals.org/doi/full/10.1161/CIR.0b013e31820a55f5)
8. *K. Vanisree and J. Singaraju, “Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks,” International Journal of Computer Applications, vol. 19, no. 6, pp. 6–12, 2011.*

[*https://scholar.google.com/scholar\_lookup?title=Decision+support+system+for+congenital+heart+disease+diagnosis+based+on+signs+and+symptoms+using+neural+networks&author=K.+Vanisree&author=J.+Singaraju&publication\_year=2011*](https://scholar.google.com/scholar_lookup?title=Decision+support+system+for+congenital+heart+disease+diagnosis+based+on+signs+and+symptoms+using+neural+networks&author=K.+Vanisree&author=J.+Singaraju&publication_year=2011)

1. *David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database.*
2. [*http://scikitlearn.org/stable/modules/generated/sklearn.model\_selection.StratifiedShuffleSplit.html*](http://scikitlearn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html)
3. [*https://medium.com/diogo-menezes-borges/boosting-with-adaboost-and-gradient-boosting-9cbab2a1af81*](https://medium.com/diogo-menezes-borges/boosting-with-adaboost-and-gradient-boosting-9cbab2a1af81)