Springboard Data Science Career Track 2020

Capstone Project 1: Project Proposal

Heart Disease Prediction

Mukkul Kurwa

April 2020

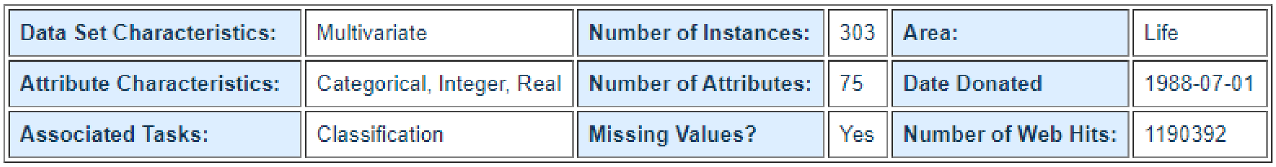
The project's **domain background** — the field of research where the project is derived is Healthcare. This project will focus on predicting heart disease using neural networks. Based on attributes such as blood pressure, cholesterol levels, heart rate, and other characteristic attributes, patients will be classified according to varying degrees of coronary artery disease. This project will utilize a dataset of 303 patients and distributed by the UCI Machine Learning Repository. I will be using some common Python libraries, such as pandas, NumPy, and Matplotlib. Furthermore, for the machine learning side of this project, we will be using sk-learn and Keras.

A problem statement — The task is to predict heart disease using neural networks based on attributes. This is a classification problem where the model takes attributes and detects whether the patient has heart disease or not.

The **datasets and inputs** — The dataset is available through the University of California, Irvine Machine learning repository. Here is the URL:
http:////archive.ics.uci.edu/ml/datasets/Heart+Disease

This dataset contains patient data concerning heart disease diagnosis that was collected at several locations around the world. There are 76 attributes, including age, sex, resting blood pressure, cholestoral levels, echocardiogram data, exercise habits, and many others. To data, all published studies using this data focus on a subset of 14 attributes - so we will do the same. More specifically, I will use the data collected at the Cleveland Clinic Foundation.

Abstract: 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach



Data Set Characteristics:	Multivariate	Number of Instances:	303	Area:	Life
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	75	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	1190392

A **solution statement** — I'm going to use Keras to build and train our network. This model will be relatively simple and will only use dense (also known as fully connected) layers. This is the most common neural network layer. The network will have one hidden layer, use an Adam optimizer, and a categorical cross entropy loss. I won't worry about optimizing parameters such as learning rate, number of neurons in each layer, or activation functions in this project.

A **benchmark model** — I’ll use the default vanilla model as the benchmark. Hyper parameter tuning my final model will result in significant improvements over this benchmark.

A set of **evaluation metrics** — Once the model is trained, I need to test its performance on the testing dataset. The model has never seen this information before; as a result, the testing dataset allows me to determine whether or not the model will be able to generalize to information that wasn't used during its training phase. I will use some of the metrics provided by Scikit-learn for this purpose such as classification reports and accuracy score.

An outline of the project design —The project design includes the following phases:

- **Data Preprocessing:** This dataset is going to require multiple preprocessing steps. First, I have columns in our DataFrame (attributes) that I don't want to use when training our neural network. I will drop these columns first. Secondly, I’ll remove the missing data which is indicated by a “?” and then drop the rows with NaN values from DataFrame. During the pre-processing, I will also split the dataset into X and Y datasets, where X has all of the attributes I want to use for prediction and Y has the class labels.

- **Data Splitting:** Split the data into a training set and validation set with an 80-20 split.

- **Model Training and evaluation:** I will start with the simple model architecture first before training and evaluation. Then iterate this process trying different architectures and hyper-parameters to reach an accuracy I’m happy with.

Source:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Relevant Papers:

1. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.
[Web Link]
2. David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."
[Web Link]
3. Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61.
[Web Link]