**Jonathan Pham**

**CS 4300 AI**

**Due: 05/02/2019**

**Heart Disease Predictor:**

**Project Report**

**Background**:

Heart disease is a general term for several different conditions and the cause of an individual’s case depends on the type they have. Unfortunately, heart disease is remains extremely prevalent in our society; being the leading cause of death for both men and women in the United States. It is estimated that 90% of cardiovascular disease (CVD) is preventable. There are many risk factors for heart diseases that we will examine in our dataset.

The objective of this project is to build a model that may predict heart disease prevalence. After loading our dataset, heart.csv, a little bit of sanity checking will be performed to ensure the proper data is being fed into our model. Next, heart disease occur will be predicted based on a combination of risk factors describing the disease. Different neural network techniques will be implemented, models/graphs (pie-chart, bar graph, pair plots) will be utilized for visualization, and data will be verified based on accuracy of the target, heart disease.

The reasoning for selecting this dataset is because I currently work in the healthcare field, specifically pharmacy, and a large majority of patients that I see suffer from heart disease. There are numerous medications on the market used to treat said disease including ACE-Inhibitors, Beta-Blockers, Calcium-Channel Blockers and diuretics. But if there is a way to build a model that may predict heart disease occurrence based on input factors that work with machine learning, the healthcare field may forever be changed for the better.

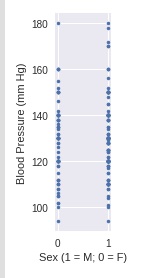
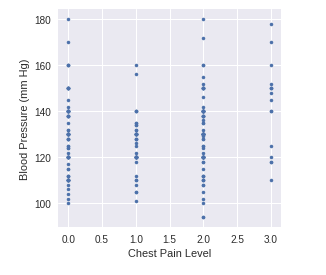
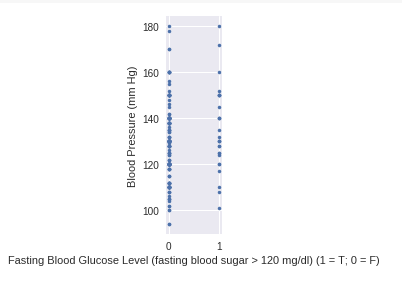
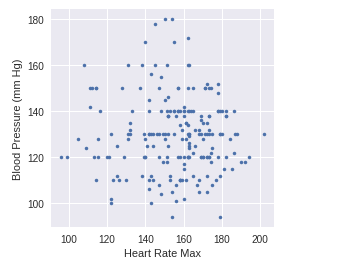
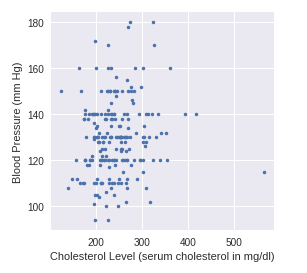
**Dataset:**  <https://www.kaggle.com/ronitf/heart-disease-uci>

The dataset that I used contains 14 different columns and 920 total individual data. They are outlined in the table below.

|  |  |
| --- | --- |
| **Dataset Columns** | |
| **Age** | age of the individual |
| **Sex** | 1 = male  0 = female |
| **Chest-pain level** | 1 = typical angina  2 = atypical angina  3 = non-anginal pain  4 = asymptotic |
| Resting Blood Pressure | value of an individual in mmHg (unit) |
| **Serum Cholesterol** | value in mg/dl (unit) |
| **Fasting Blood Sugar (FBS)** | If FBS > 120mg/dl:  1 = true  0 = false |
| Resting ECG | 0 = normal  1 = having ST-T wave abnormality  2 = left ventricular hypertrophy |
| **Max Heart Rate (MHR)** | value of max heart rate |
| Exercise induced angina | 1 = yes  0 = no |
| ST depression induced by exercise relative to rest | ST depression value |
| Peak exercise ST segment | 1 = upsloping  2 = flat  3 = down sloping |
| Number of major vessels (0-3) colored by fluoroscopy | Number of major blood vessels |
| Thal | 3 = normal  6 = fixed defect  7 = reversable defect |
| Target  (diagnosis of heart disease) | 0 = absent **< 50% diameter narrowing**  1 = present > **50% diameter narrowing**  The target variable is feature number 14, which is a narrowing in any major blood vessel due to cholesterol and plaque deposits, as detected through the use of an angiogram |

I decided to examine factors that I deemed would be thought-provoking to look at; sex, fasting blood glucose levels, chest pain level, max heart rate and cholesterol levels (in bold above and will be located on the x-axis). Let’s see how these factors affect heart disease by graphing them against the patients’ blood pressure (y-axis). Graphs of the five inputs versus blood pressure can be seen below:

Although blood pressure isn’t the target, it is a strong predictor of whether an individual will develop heart disease (Obara).

**Model Training & Splitting Data into Training/Test Sets:**

We can train our model by analyzing existing data because we already know the target; whether each patient has heart disease or not; supervised learning. The trained model will then be used to predict if users suffer from heart disease.

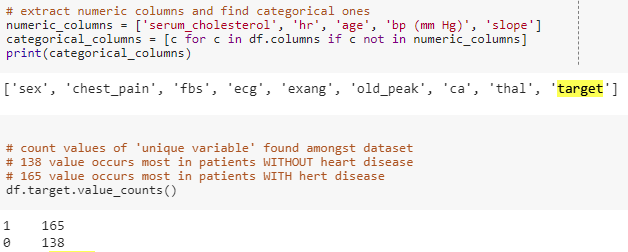
In order to get an accurate model with logistic regression, we must choose only the features that will make a real impact on the decision of the classifier. Moreover, this is the target, which is whether the person has heart disease or not. In this case I've chosen all the columns as input features. I split the dataset into training and test sets, based on a ratio of 75:25 of the total dataset size (76 training).

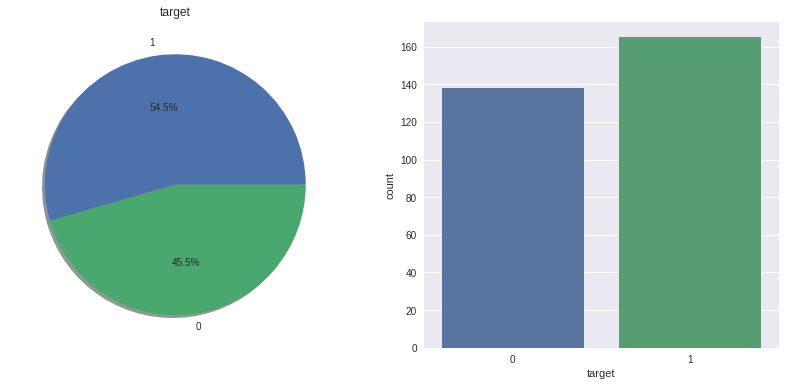
**Statistics:**

In addition, I analyzed the features, identified patterns and explored the data itself my using built-in Python functions that displays data in a table format and find the input columns with numeric columns (sex, chest pain, fasting blood glucose levels, EKG, exang, ST peak, ca, thal and target).

Target values were obtained and displayed in a pie-chart (percentage) and bar graph. All these useful features aid in visualization of the dataset because it is presented in a picture format as opposed to looking at raw numerical data values.

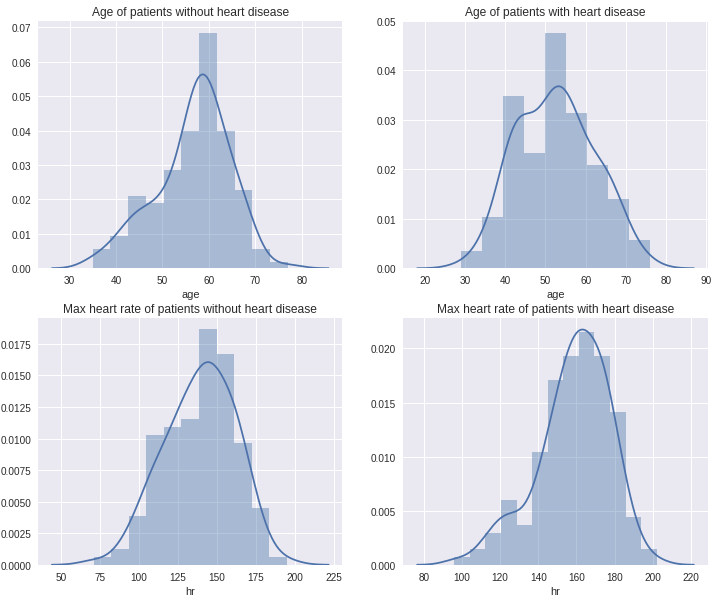
As seen below, the value\_countsfunction was utilized. It returns a series containing counts of unique values. In doing so, the blood pressure values 165 and 138 were most prevalent for the presence and absence of heart disease, respectfully.



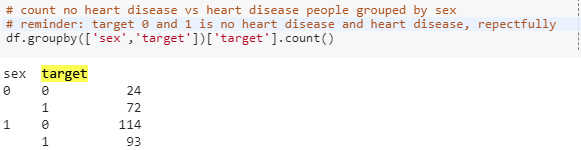


Graph max heart rate of patients without and with heart disease showed a fairly *normally skewed distribution*.

In addition, graphic age of patients with and without heart disease showed a fairly negative skewness (right-modal).

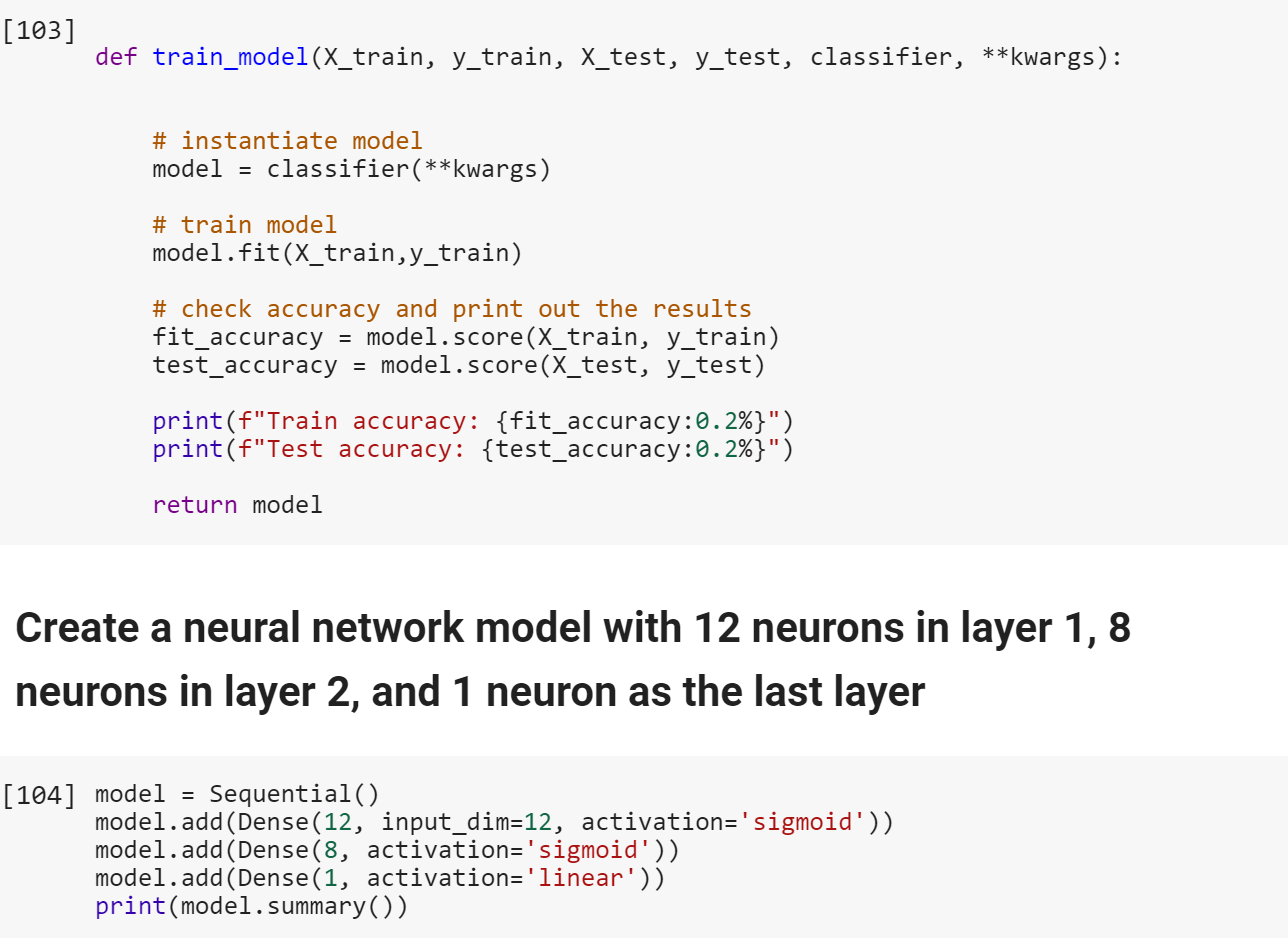


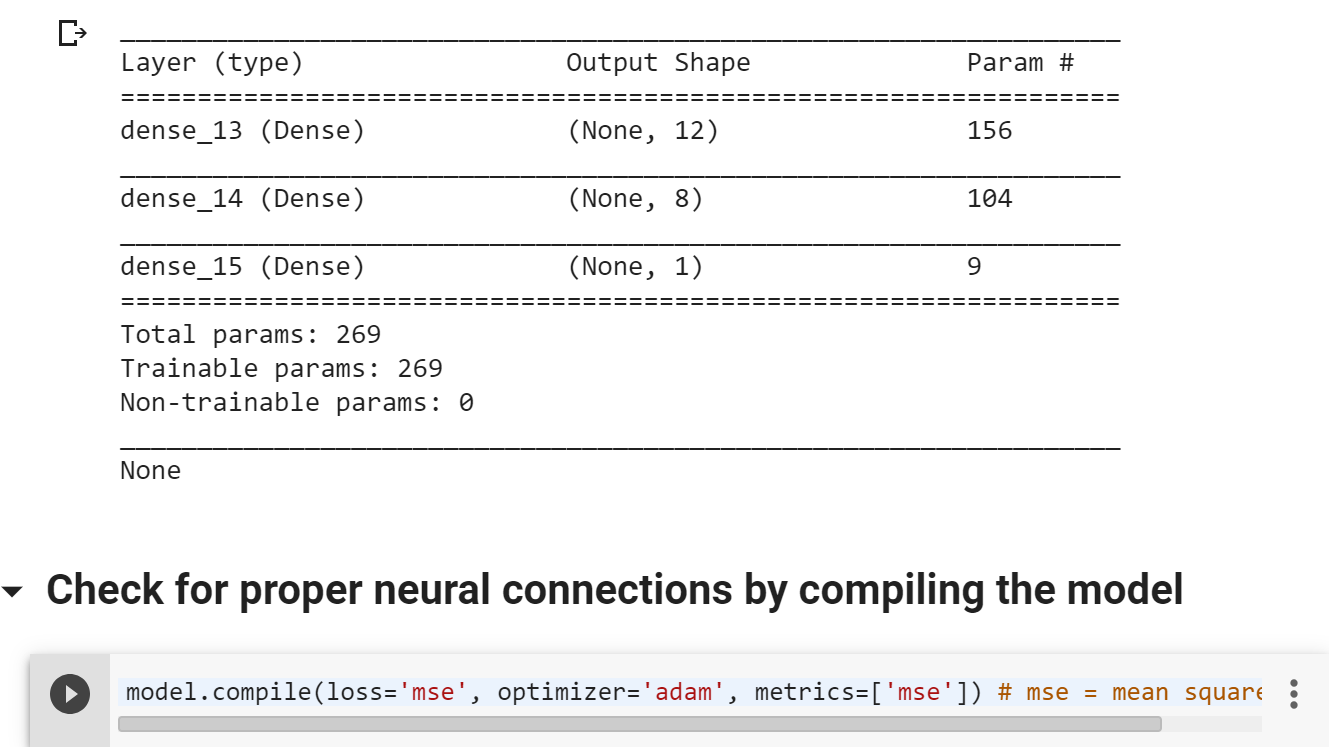
Grouping sex of individuals with and without heart disease yielded for following results:

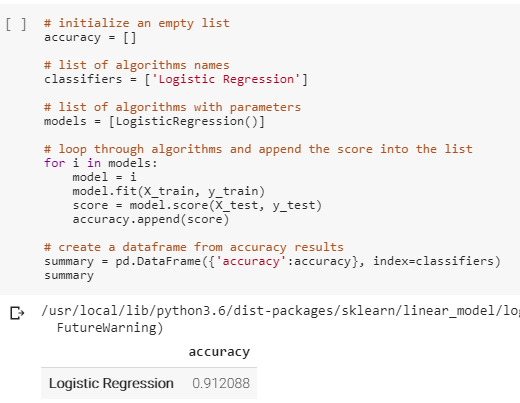


Males accounted for 55% and females 45% for the prevalence of heart disease.

**Modelling, predicting and analyzing with Neural Learning**







It is important to note that logistic regression is *not* a classifier but rather a probability estimator. It describes the data and to explain the relationship between one dependent binary variable (the target) and one or more nominal, ordinal, interval or ratio-level independent variables (input variables).

Overall, the logistic regression accuracy of 0.912088 deems our model fairly accurate. To increase the model’s performance, we can increase the number of epochs, incorporate additional layers into our neural network and increase or decrease rows in our training and validation set.

**Conclusion:**

All in all, neural learning has a bright future in the healthcare field. Good data driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that individuals can live healthy lives. Imagine living in a world where heart disease experts are not available. Conversely bright individuals with coding and neural networking experience will be able to predict whether a disease will occur or not accurately. This may be achieved with information about a patient's medical history.

**Citations**:

* Obara, Fumio, et al. “Influence of Hypertension on the Incidence of Cardiovascular Disease in Two Rural Communities in Japan: the Tanno-Sobetsu [Corrected] Study.” *Hypertension Research : Official Journal of the Japanese Society of Hypertension*, U.S. National Library of Medicine, Aug. 2007, [www.ncbi.nlm.nih.gov/pubmed/17917314](http://www.ncbi.nlm.nih.gov/pubmed/17917314).
* Adhikari, Badri. “Badriadhikari/2019-Spring-DL.” GitHub, 2 Apr. 2019, github.com/badriadhikari/2019-Spring-DL/blob/master/course\_content/module1\_intro2ML/06\_NNs\_for\_Regression.ipynb.