**By: Amisha and Ankita Guha**

**Project Proposal**

**Predicting Heart Disease with Machine Learning**

**Background**

In the United States, Heart disease is the leading cause of death for both men and women. More than half of the deaths due to heart disease in 2015 were in men according to Centers for Disease Control and Prevention (CDC).

About 610,000 people die of heart disease in the United States every year–that’s 1 in every 4 deaths. Coronary heart disease (CHD) is the most common type of heart disease, killing over 370,000 people annually. Every year about 735,000 Americans have a heart attack. Of these, 525,000 are a first heart attack and 210,000 happen in people who have already had a heart attack. Heart disease is the leading cause of death for people of most ethnicities in the United States, including African Americans, Hispanics, and whites. For American Indians or Alaska Natives and Asians or Pacific Islanders, heart disease is second only to cancer. Heart disease costs the United States about $200 billion each year. This total includes the cost of healthcare services, medications, and lost productivity.

**Problem Statement**

Preventing heart disease is important as Heart disease is the number one cause of death worldwide. Good data-driven systems for predicting heart disease can improve the entire research and prevention process, making sure that more people can live healthy lives. To learn how to prevent heart disease we must first learn to reliably detect it. Our dataset is from a study of heart disease that has been open to the public for many years. The study collects various measurements on patient health and cardiovascular statistics, and of course makes patient identities anonymous. Data is provided courtesy of the Cleveland Heart Disease Database via the UCI Machine Learning repository. Thus, developing a Machine Learning Predictive Model that could enhance the predictive power of not only historical patient health data but also with present and future patient health data, with less bias and variance in the model, is the need of the day.

**Objective**

Our goal is to predict the binary class heart\_disease\_present, which represents whether or not a patient has heart disease:

* 0 represents no heart disease present
* 1 represents heart disease present, along with the probabilities of the presence of the heart disease in the patient.

Identify statistically significant features contributing to the presence or to the absence of heart diseases in patient.

**Data source Description**

We are using the dataset from the LIVE competition, hosted by [driven data](https://www.drivendata.org/competitions/).

**There are 14 columns in the dataset, where the patient\_id column is a unique and random identifier. The remaining 13 features are described in the section below.**

* **slope\_of\_peak\_exercise\_st\_segment (type: int): the slope of the peak exercise** [**ST segment**](https://en.wikipedia.org/wiki/ST_segment)**, an electrocardiography read out indicating quality of blood flow to the heart**
* **thal (type: categorical): results of** [**thallium stress test**](https://www.ucsfbenioffchildrens.org/tests/007201.html) **measuring blood flow to the heart, with possible values normal, fixed\_defect, reversible\_defect**
* **resting\_blood\_pressure (type: int): resting blood pressure**
* **chest\_pain\_type (type: int): chest pain type (4 values)**
* **num\_major\_vessels (type: int): number of major vessels (0-3) colored by fluoroscopy**
* **fasting\_blood\_sugar\_gt\_120\_mg\_per\_dl (type: binary): fasting blood sugar > 120 mg/dl**
* **resting\_ekg\_results (type: int): resting electrocardiographic results (values 0,1,2)**
* **serum\_cholesterol\_mg\_per\_dl (type: int): serum cholestoral in mg/dl**
* **oldpeak\_eq\_st\_depression (type: float): oldpeak =** [**ST depression**](https://en.wikipedia.org/wiki/ST_depression) **induced by exercise relative to rest, a measure of abnormality in electrocardiograms**
* **sex (type: binary): 0: female, 1: male**
* **age (type: int): age in years**
* **max\_heart\_rate\_achieved (type: int): maximum heart rate achieved (beats per minute)**
* **exercise\_induced\_angina (type: binary): exercise-induced chest pain (0: False, 1: True)**

**Technical Summary**

We are planning to deploy our Prediction Model, using Python, and would be mostly using the following    libraries:

* NumPy, SciPy, Pandas for Data Wrangling and Exploratory Analysis
* Matplotlib, Seaborn for Data Visualization
* SciKit Learn, Keras, Tensorflow for building Predictive Models with Machine Learning

We would also like to use DL4J, to see the performance of our prediction model in terms of model’s performance in evaluation.

Our main goal is to reduce the loss in our Prediction Model by reducing the Stochastic Gradient in the deep neural network in such a way, so that the actual predicted values are maximized and the incorrect predicted values in our Confusion Matrix are minimized. We might also do a Feature Selection based on our Model outcome, since we shall be dealing with 13 attributes.

**Anticipated Challenges**

Since the dataset is small it may result in overfitting the model and we might face a lot of challenges in terms of tuning the hyperparameters and reducing the Stochastic Gradient.

Outliers might skew the data and ignoring them might either cause in Overfitting or Underfitting of the Model parameters.