**"Heart Attack Prediction Proposal"**

**Introduction:**

Cardiovascular disease or heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease. From WHO statistics, every year, 17.9 million die from heart disease. Medical studies suggest that human lifestyles are the main reason behind this heart problem. Apart from this, there are many key factors that contribute to an individual increasing their chances of heart disease, such as obesity, diabetes, and high blood pressure.

The Cleveland dataset was defined, cleaned, and analyzed to answer pertinent questions that are vital to human life as regards the possibility of a heart attack. This dataset stemmed from a survey donated on 7/1/1988. However, this could serve as a template for learning why individuals could have a heart attack. We plan on using the heart attack prediction dataset to create a machine learning model to accurately predict heart attacks and analyze the contributing factors.

**Statement of Problem**

This dataset gives information concerning heart disease diagnosis. The dataset contains 14 columns. The target is the class variable which is affected by the other 13 columns. Here the aim is to classify the target variable to (disease\non disease) using different machine learning algorithms and build a model which demonstrates meaningful predictive power.

**Dataset Source**

https://www.kaggle.com/datasets/nareshbhat/health-care-data-set-on-heart-attack-possibility?select=heart.csv

**Plan of Action**

* Clean the data
* Normalize and standardize the data
* Store the cleaned data in SQL or Spark
* Build various models using various ML algorithms and evaluate the most applicable model
* Optimize the model using an iterative process to achieve meaningful predictive power

**Visualizations**

* ROC- curve
  + In order to visualize the true positives created by the model
* Heat map
  + We will use a heat map to visualize the correlation between variables in the dataset
* Decision tree
  + A decision tree can be used to gain insights into the decisions the model is making as it lays out all possibles outcomes

**Model Implementations**

* K-Nearest-Neighbor
* Decision Tree
* Random Forest Classifier
* Logistic Regression