PREDICTIVE MODELING OF CORONARY HEART DISEASE : INSIGHTS FROM THE FRAMINGHAM HEART STUDY DATASET

## INTRODUCTION

Coronary heart disease (CHD) remains a top concern in the health sector due to its high fatality rate. This has created the need to have risk assessment strategies to ensure early detection and timely intervention and prevention.

In this study, the primary objective is to perfom a comprehensive analysis of the Framingham Heart Study Dataset and make use of machine learning algorithms to predict the ten year risk of coronary heart disease occuring in a person.

In particular, we will make use of Random Forests, a supervised machine learning algorithm, to learn both simple and complex patterns in our data, and as a result create a model that will be used to assess which individuals are at risk of getting diagnosed with CHD. Random Forests is preffered over other machine learning models beacause:

1. It is an ensemble of the decision tree algorithm hence it can learn complex relationships that individual trees would otherwise find it difficult to learn. This can improve the model’s accuracy.
2. Random Forests algorithm is robust in nature. Their nature allows them to ignore outliers and noise, making them to have more reliable results
3. It is able to identify relevant features by providing a feature importance score of each feature. This can help to understand which features are more important in determing the risk of one acquiring CHD.
4. Ease of use. The algorithm has