**CARDIOVASCULAR RISK PREDICTION**

It is Heart and blood vessel disease (also called heart disease) includes numerous problems, many of which are related to a process called atherosclerosis

Atherosclerosis is a condition that develops when a substance called plaque builds up in the walls of the arteries. This buildup narrows the arteries, making it harder for blood to flow through. If a blood clot forms, it can block the blood flow. This can cause a heart attack or stroke.

Cardiovascular disease is a significant and ever-growing problem in all over nations. It is of pressing interest as developing countries experience a change in lifestyle which introduces novel risk factors for cardiovascular disease. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk

We started our project by importing data by an almabetter platform. The dataset was from an ongoing cardiovascular study on residents of the town of Framingham,Massachusetts. After data importing we moved to missing values where we found the only glucose column had a considerable amount of missing values we replaced then with glucose median.As there were no duplicates we directly moved toward EDA. During EDA on all the features of our dataset.. We first analyzed our dependent variable, 'TenYearCHD'. Next we analyzed categorical variables and dropped the variables who were not correlated, we also analyzed numerical variables, found out the correlation, distribution and their relationship with the dependent variable. We also removed some numerical features and hot encoded the categorical variables.

During EDA we found that number of females showing cardiovascular risk was more than number of male, A people showing cardiovascular risk as well as prevalent stroke were very few, Age was the most important feature for our machine learning models

Next we implemented 3 machine learning algorithms K neighbor classifier, Random Forest and Logistic regression. We did hyperparameter tuning to improve our model performance. In general, all models have comparable accuracy. Where as Random Forests gave the accuracy of 91% on test dataset & KNeighbors classifier, Logistic Regression are gave us the accuracy 79%, 88% among other algorithms We could conclude that these two algorithms KNeighbors classifier, Random Forests are the best to predict whether the person showing cardiovascular risk or not from our analysis. We picked up the KNeighbors classifier to deploy model on heroku

Contributerts Roles:

1.Lalit Ahirrao:

1.Data Wrangling

2.Handeling missing values

3.EDA of TenYearCHD variable

4.HeatMap

5.scaling down variables

6.Model training

2.Aniket Gajmal:

1.Data Wrangling

2.Hadeling outliers

3.EDA of Age & diabetes variable

4.Model training

5.Building a Predictive System

6.web app development(|Flask)

3.Rushikesh Pawar:

1.Data Wrangling

3.EDA of Education and sex

4.Model training

5.Creating a Pickle File

6.web app development(Flask)

4.Prasad Ghegade:

1.Data Wrangling

2.Handling Duplicates

3.EDA of bp\_meds and prevalent stroke variable

4.Model training

5.web app deployment on Heroku

5.Samarth Gangurde:

1.Data Wrangling

2. EDA of heart rate variable

4.Creating dummy variables

5.Model training

6.Comparison of models performance

7.web app deployment on Heroku