**SUMMARY**

**CARDIOVASCULAR RISK PREDICTION**

* **INTRODUCTION**

Cardiovascular disease has been regarded as the most severe and lethal disease in humans. The increased rate of cardiovascular diseases with a high mortality rate is causing significant risk and burden to the healthcare systems worldwide. Cardiovascular diseases are more seen in men than in women particularly in middle or old age , although there are also children with similar health issues . According to data provided by the WHO, one-third of the deaths globally are caused by the heart disease. CVDs cause the death of approximately 17.9 million people every year worldwide and have a higher prevalence in Asia.

* **PROBLEM STATEMENT**

Predicting coronary heart disease in advance helps raise awareness for the disease. Preventive measurements like changing diet plans and exercise can slow down the progression of CHD. Early prediction can result in early diagnosis. So, we can treat the disease at an early stage and avoid more invasive treatment.

**The goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).**

* **TOOLS REQUIRED**

To do analysis on Cardiovascular data we used language python. In python we have various libraries which we must import like NumPy, Pandas, Matplotlib and Seaborn**,** Scikit-Learn.

* **ACTIVITY**
* First step involved is understanding the data and getting answers to some basic questions like; What is the data about? How many rows or observations are there in it? How many features are there in it? What are the data types? Are there any missing values? And anything that could be relevant and useful to our investigation.
* Handling missing values is an important skill in the data analysis process.
* Exploratory data analysis is a crucial part of data analysis. It involves exploring and analysing the dataset given to find out patterns, trends and conclusions to make better decisions related to the data, often using statistical graphics and other data visualization tools to summarize the results.
* Data manipulation involves manipulating and changing our dataset before feeding it to various regression machine learning models. This involves keeping important features, outlier treatment, feature scaling and creating dummy variables if necessary.
* In statistics, an outlier is a data point that differs significantly from other observations.
* Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is done to prevent biased nature of machine learning algorithms towards features with greater values and scale.
* **CONCLUSION**
* As age increases the risk of getting diagnosed with heart disease also increases.
* Cigarette consumption is also a major factor that causes CHDs.
* Patients having Diabetes and cholesterol problems show a higher risk of CHDs.
* Patients having high glucose levels are more prone to CHDs.
* Patients with a history of “strokes” have a higher chance of developing CHDs.
* Patients with high BMI(Body Mass Index) are at more risk of getting diagnosed with CHDs.
* Finally, we can say that, XGBoost Classifier has performed best among all the models with the accuracy of 76% & f1-score of 0.71. It is by far the second highest score we have achieved. So, It’s safe to say that XGBoost Classifier provides an optimal solution to our problem.