**Project Report**

**Introduction**

Heart Disease is a major problem which we are facing in our day to day life. The heart disease rate is increasing day by day. If we see the record in 1980 the death rate was 1.8 million and it has increased to 2.8 million till 2008. Heart Disease is caused due to various factors such as Lifestyle of a person, Health, BP, and stress is one of the major factor for this. So based on these factors we want to find whether the person will have heart disease in the upcoming 10 yrs. This is an interesting project where we will be predicting heart disease will be caused in future or not. So if we know that are going to have heart disease in future, we can take precautions and medical treatment to cure it so that the death rate can be decreased. Here we are given the dataset of Framingham that contains detailed info about heart disease. Dataset Contains :- 4240 records and 15 features By using ML Algorithm we will be doing prediction for heart disease in upcoming 10 yrs.

**Understand and Define the Problem**

**Problem Statement**:-

Heart Disease is a major problem world-wide and lot of times people do not get the risks of it on time.

**Objectives of the project**:-

To determine possible chronic heart disease in coming next ten years for a person.

**Data Sources:-**

In this project we have used the Framingham Dataset of Ten Year Chronic heart disease study.

**Tools and Techniques used:-**

We have used the Random Forest Classification model for training the model, also as the data is imbalanced we have done Oversampling of the model.

For checking the accuracy of the model we have used the Accuracy\_score function and F1\_score function.

**Limitations:-**

The imbalanced data causes a bit of fluctuation in training of model and affects a bit of accuracy too, thus oversampling is required.

**Dataset Preparation and Preprocessing**

**Data Collection**

The data was obtained from internet which is freely available provided by Scientist Framingham, this data was taken by various physical methods eg: blood pressure, heart rate, current smoker and diabetes.

**Data Visualization**

The amount of data used in ML projects is large in size. When the large data is plotted i.e. visualized, it makes it easy to understand and analyze.

Data visualization was done using matplotlib and seaborn, we plotted Barplot and Histogram. Histogram to compare all the columns with the target variable.

**Labeling**

Regression and Classification type of prediction is done using Supervised machine learning technique. In this technique, the data points are labeled i.e. target values of the data points are known. If the data is not labeled, it needs to be done which takes lot of efforts and time. But in our case the data was already labeled and there was no need of Label Encoding.

Also the whole data was numerical in type.

**Data Selection**

In our case we had to delete the unwanted data i.e Education, current smoker column and glucose column. It was done because some of the columns were unwanted and other columns information was already present in other columns like current smoker was represented by cigsPerDay and glucose was represented by diabetes.

**Data Preprocessing**

The purpose of preprocessing is to convert raw data into a form that is useful in training and testing the ML model. The structured and clean data produces more precise results. In short, good quality data when fed to the ML model, it produces better results.

The Preprocessing technique includes data formatting, cleaning, and sampling techniques.

**Data Formatting :** Usually all the categorical values are changed to numerical values in this stage but our dataset already had all the numerical values so there was no need of data formatting.

**Data cleaning:** As the data was imbalanced we had to do oversampling of data. Also all the null values were deleted and others replaced by median of other corresponding data.

**Data Scaling**

In this stage, the data is transformed into the form which is appropriate for machine learning. The scaling and normalization is usually used to transform the data.

**Scaling:** The different attributes in the dataset may have different ranges i.e. data values may vary over different values. Scaling is used to correct this problem.

We used StandardScaler technique to normalize our data and reduce the difference in the overall min and max values of columns.

**Dataset Splitting**

The given dataset is split into three parts: training, testing, and validation sets. The ration of training and testing sets is typically 80 to 20 percent. The 20 percent of the training set is further split as a validation set.

**Model Training**

In this stage, the training data is fed to the ML algorithm to build and train a model. The purpose of training is to develop a model.

We used the Random Forest Classification algorithm to train our model because this model had greater accuracy of 83% as compared to other models.

**Model Testing and Evaluation**

The goal of this step is to develop the simplest, reliable and efficient model. This requires model tuning. Depending on your project, you may use a number of algorithms to test and ultimately select the best model.

We fed the testing data to the testing variable i.e X\_test and then we calculated the accuracy and confusion matrix on “target\_test” data.

**Model Deployment**

When the reliable model is selected and validated, the model is put into production. Model

Deployment means putting the model in use (production).

We did the following:

**Batch based Deployment:** In this type, the prediction was done in batch of observations rather than on continuous basis. We provided values by ourselves and not in any continuous manner nor in a dynamic way using IoT devices.

**Conclusion and Further Development**

In conclusion it was observed that the available Framingham dataset is quite unstable and imbalanced. The dataset had values which was biased to majority of result that is No Chronic heart disease and minority of data was Chronic Heart Disease, so the model was biased only towards majority of value so it was giving wrong results initially.

Thus we used Oversampling of data using SMOTE and Standardized it according to the expected result.

Now the model predicts accurately if the person will have a Chronic Heart Disease in coming 10 years.