

Centre for Development of Advanced Computing, Mumbai

**REPORT ON**

**HEART ATTACK PREDICTION**

PG-DBDA SEP 2022

Submitted by:

**Project Team 17**

Chandrashekhar Bhavsar Shivangi Singh

Digvijay Pawar Rohan Nipurte

Sumeet Yadav

Mr. Abhishek Raghorte

**Project Guide**

**1. INTRODUCTION**

“CDC (Centre for disease control and prevention) is the nation’s leading science-based, data-driven, service organization that protects the public’s health. For more than 70 years. CDC is one of the major operating components of the Department of Health and Human Services.”

CDC works 24/7 to protect America from health, safety and security threats, both foreign and in the U.S. Whether diseases start at home or abroad, are chronic or acute, curable or preventable, human error or deliberate attack, CDC fights disease and supports communities and citizens to do the same.

Heart disease is the leading cause of death for people. Heart dataset collected by CDC to help people stay healthy. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient. There are several factors responsible for heart disease, these factors are need to be analysed for heart attack prediction.

**1.1 Dataset**:

The Behavioural Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviours, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

By collecting behavioural health risk data at the state and local level, BRFSS has become a powerful tool for targeting and building health promotion activities. As a result, BRFSS users have increasingly demanded more data and asked for more questions on the survey. Currently, there is a wide sponsorship of the BRFSS survey, including most divisions in the CDC National Centre for Chronic Disease Prevention and Health Promotion; other CDC centre’s; and federal agencies, such as the Health Resources and Services Administration, Administration on Aging, Department of Veterans Affairs, and Substance Abuse and Mental Health Services Administration.

This Dataset surveyed by CDC contains several parameters on which people have given their responses.

Dataset consists of 401958 records and 279 features, which is surveyed data of 2020 and 2021.

**Heart.csv**

Contains 401958 records and 279 columns, out of which following fields are selected based on Meta Data.

* SEXVAR : Sex of Respondent
* GENHLTH : Would you say that in general your health is:
* EXERANY2 : During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?
* SLEPTIM1 : On average, how many hours of sleep do you get in a 24-hour period?
* CVDINFR4 : (Ever told) you had a heart attack, also called a myocardial infarction?
* CVDSTRK3 : (Ever told) (you had) a stroke.
* ASTHMA3 : (Ever told) (you had) asthma?
* CHCSCNCR : (Ever told) (you had) skin cancer?
* CHCOCNCR : (Ever told) (you had) any other types of cancer?
* CHCCOPD2 : (Ever told) (you had) chronic obstructive pulmonary disease, C.O.P.D., emphysema or chronic bronchitis?
* ADDEPEV3 : (Ever told) (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)?
* CHCKDNY2 : Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
* MARITAL : Are you: (marital status)
* DIFFWALK : Do you have serious difficulty walking or climbing stairs?
* \_PHYS14D : 3 level not good physical health status: 0 days, 1-13 days, 14-30 days
* \_MENT14D : 3 level not good mental health status: 0 days, 1-13 days, 14-30 days
* \_MICHD : Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
* \_AGE\_G : Six-level imputed age category
* \_BMI5CAT : Four-categories of Body Mass Index (BMI)
* \_SMOKER3 : Four-level smoker status: Everyday smoker, Someday smoker, Former smoker, Non-smoker
* \_IMPRACE : Imputed race/ethnicity value

# **2. PROBLEM STATEMENT**

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 21 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

* 1. **Objective:**

Our aim in this project is to build a Machine Learning model which predicts coronary heart disease chances. With this model, people can take needed precautions and regularly visit doctor.

The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. In this paper different machine learning algorithms are applied to compare the results and analysis of CDC Dataset. Wherein we are predicting chances of heart attack on the basis of surveyed data.

Performing exploratory data analysis of CDC surveyed dataset.

Identifying the factors that impact coronary heart disease.

Developing machine learning models to predict Chances of coronary heart disease.

1. **FLOW OF EXECUTION**

**Data Preparation**

**Data Pre-processing**

**Feature Selection**

**Model Building**

**Data Analysis**

**Prediction**

**Data Visualization**

**3.1 Project Architecture:**

Data Storage

**Local Disk**

**Dataset**

**CDC**

Data Source

**Data Cleaning & Pre-processing using Pandas**

**EDA & Feature Selection**

Data Analysis

**Spark SQL**

Data Visualization

**Tableau**

**ML Model**

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Naïve Bayes
5. XGB Classifier

**Web UI**

**Flask / Render**

# **DATA PREPARATION**

**4.1 Missing Values**:

There are 229 columns that contain missing values. Since most missing values exist because there was no information available at a specific time, fields containing missing values are left as ‘NA’.

**4.2 Don’t Know and Refused Values:**

There are various columns that contain don’t know and refused response from the people. Replacing those values with mode of that columns.

* 1. **Feature Selection:**

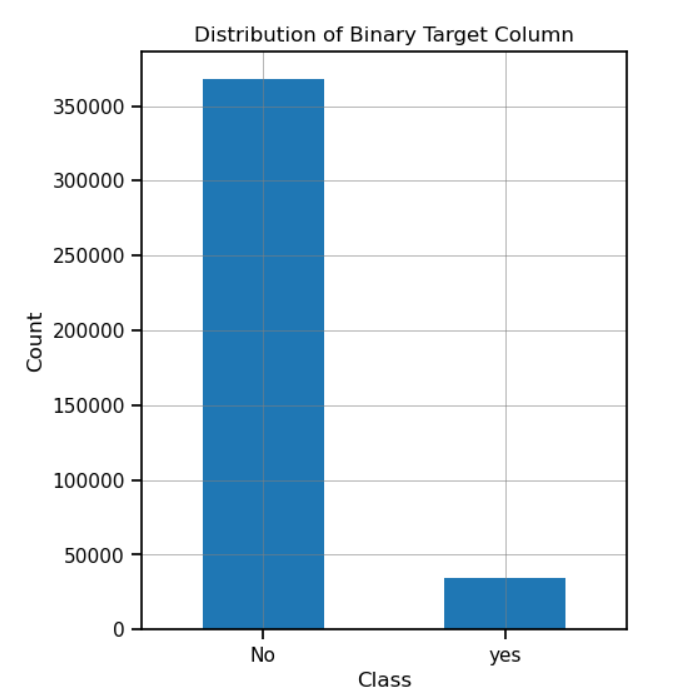
Based on Meta Data, there are 21 columns that are relevant to our model building and prediction. Out of these columns 20 columns are categorical data and 1 column is continuous data.

**4.4 Binary Target Columns Analysis**:

There is imbalance distribution of data in target Binary column. This bias in the training dataset can influence many Machine learning algorithms, leading some to ignore the minority class entirely. This is a problem as it is typically the minority class on which predictions are most important.

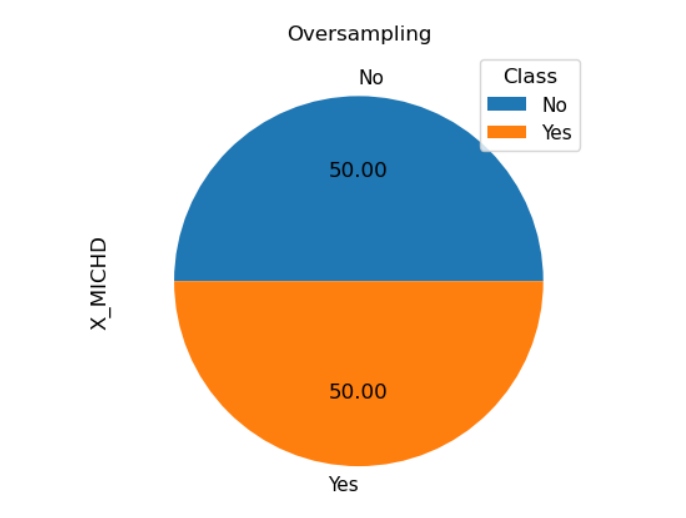
One approach to addressing the problem of class imbalance is to randomly resample the training dataset.

Resampling involves creating a new transformed version of the training dataset.



We have imbalance target data, for which we used Randomoversampling, to balance the data.

Random Oversampling involves duplicating instances of the minority class to balance the dataset. This technique can be used when the number of instances in the minority class is not sufficient to represent the underlying distribution accurately. It is done by randomly selecting instances from the minority class and adding copies of them to the dataset until the desired balance is achieved.



**4.5 Chi Square Test:**

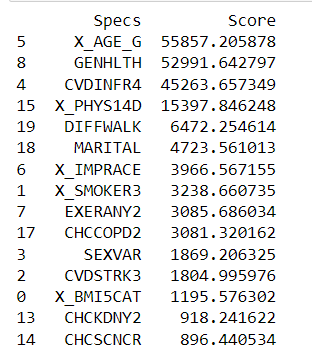
For our Algorithm we require some features, which is best for prediction, so we selected some of the columns on the basis of chi square test.

Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

Below are the scores of features:



Sort listed top 15 features for our algorithm, these features are:



**4.6 Data Splitting into Training, Testing:**

The main difference between training data and testing data is that training data is the subset of original data that is used to train the machine learning model, whereas testing data is used to check the accuracy of the model. The training dataset is generally larger in size compared to the testing dataset.

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model. A better option is to split our data into two parts: first one for training our machine learning model, and second one for testing our model. Split the dataset into two pieces: a training set and a testing set. Train the model on the training set. Test the model on the testing set, and evaluate how well our model did.

Here we use train test split function to split or data set we have split 20% data as test data and remaining 80% as training data.

**Advantages of train/test split:**

* Model can be trained and tested on different data than the one used for training.
* Response values are known for the test dataset; hence predictions can be evaluated.
* Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

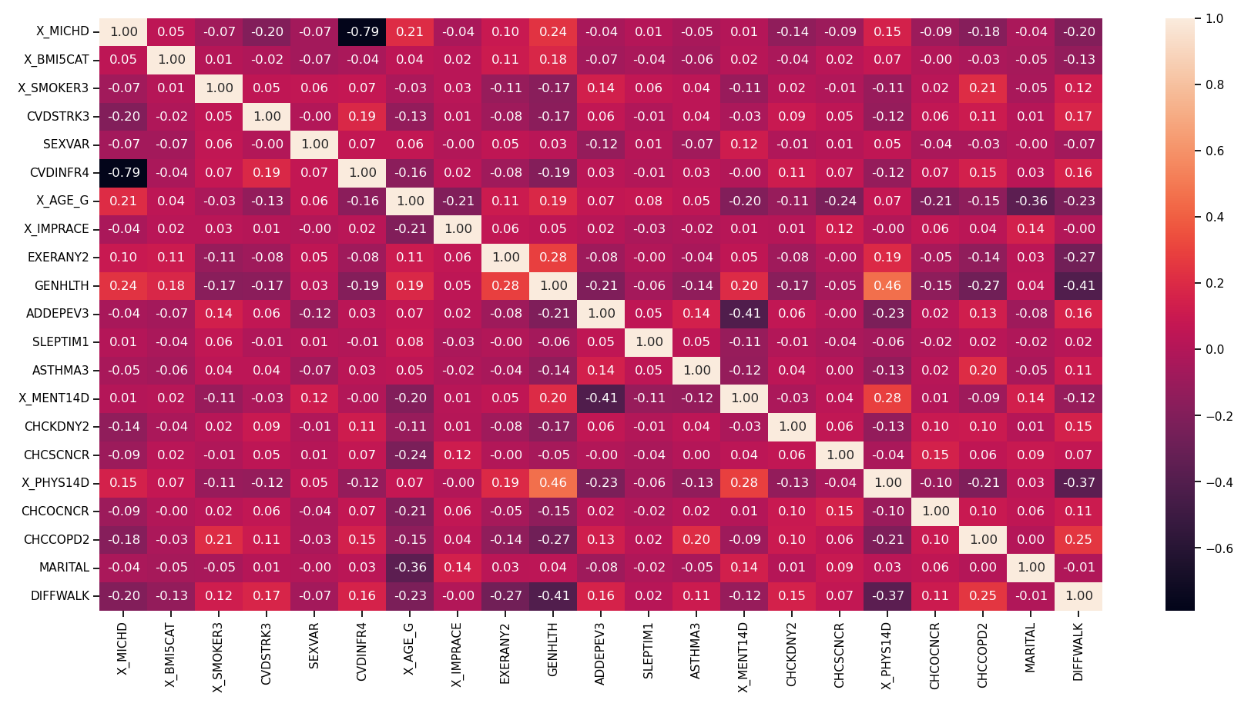
# **EXPLORATORY DATA ANALYSIS**

* 1. **Multivariate Data Analysis:**

A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using coloured cells to represent data from usually a monochromatic scale. The values of the first dimension appear as the rows of the table while of the second dimension as a column. The colour of the cell is proportional to the number of measurements that match the dimensional value. This makes correlation heatmaps ideal for data analysis since it makes patterns easily readable and highlights the differences and variation in the same data. A correlation heatmap, like a regular heatmap, is assisted by a colorbar making data easily readable and comprehensible.

The heatmap bellow shows that there is a strong correlation between the following variables:

* Coronary Heart Disease and CVDINFR4 (myocardial infarction)



**5.2 Conclusion On EDA:**

CDC Dataset consists of various features, based on Meta data we have drawn a co-relation matrix of features in that we found that CVDINFR4 (attribute consists “Yes” and “No” for myocardial infarction) is highly related to coronary heart disease. Based on the chi square test we have Analysed the important features, which are highly dependent on heart disease prediction.

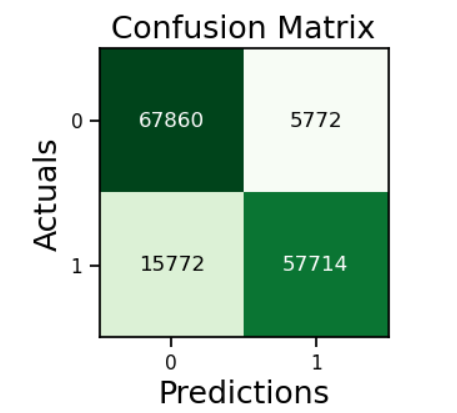
We have done the Analysis on Binary target Column, where we found the Imbalance dataset.

1. **MODEL TESTING**

**6.1 Linear Regression**:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

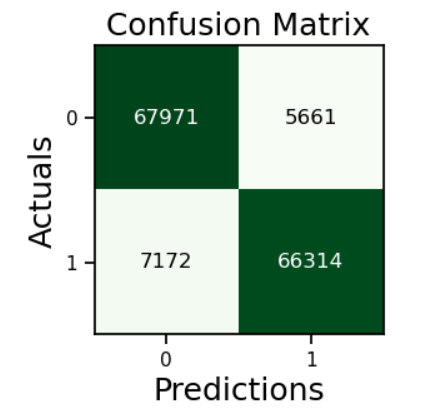
Confusion matrix Plot for Linear Regression



In this linear regression model, we got an accuracy is equal to 85.35%.

**6.2 Decision Tree:**

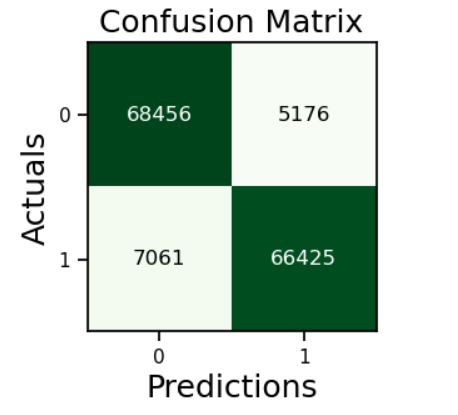
Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



In this model we got an accuracy is equal to 97.27%.

**6.3 Random Forest:**

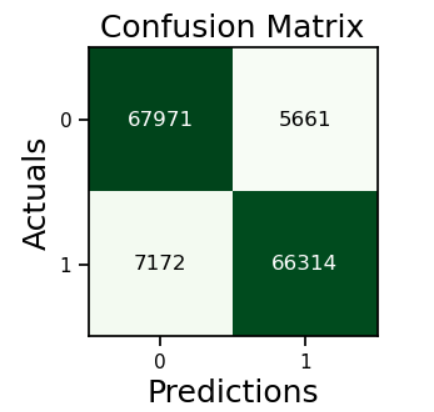
Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



In this model we got an accuracy is equal to 91.68%.

**6.4 Naïve Bayes:**

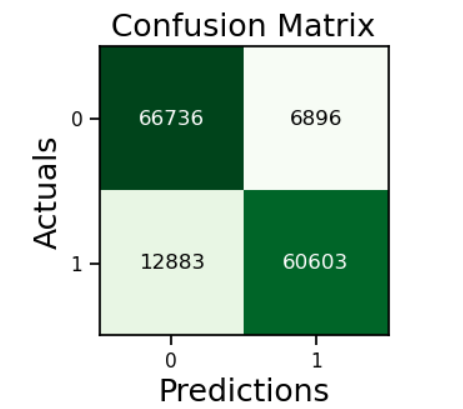
Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.



In this model we got an accuracy is equal to 91.27%

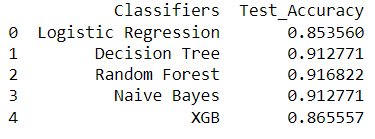
**6.5 XGB Classifier:**

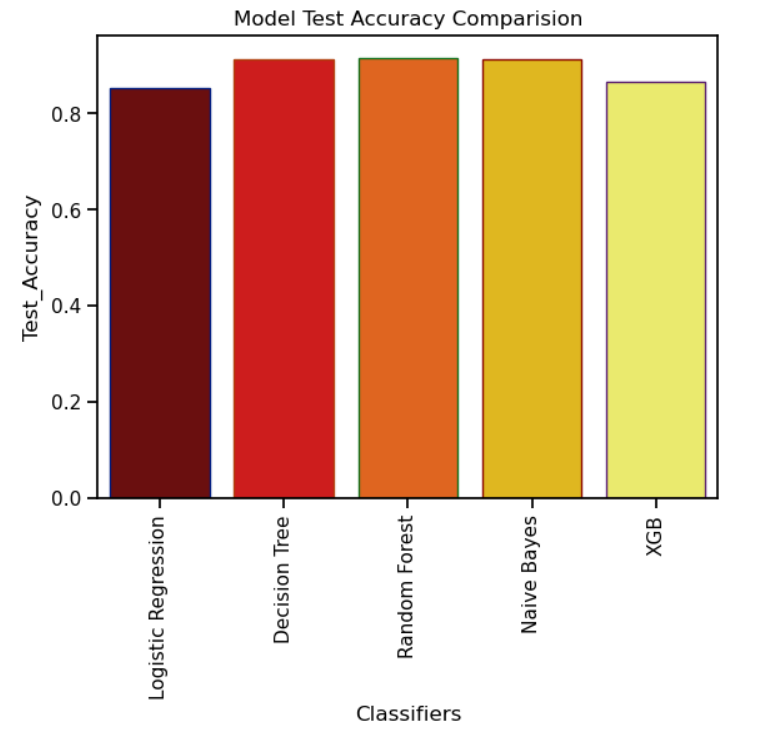
XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.



In this model we got an accuracy is equal to 86.5%

# **COMPARING MODELS**





From above figures we can see that random forest regression model has the highest accuracy

Which is 91.68 %

So, we will use random forest regression model for Heart Attack prediction of this datasets other models also perform well as their accuracies are:

Linear regression = 85.35 %

Decision Tree = 91.27 %

Naïve Bayes = 91.27%

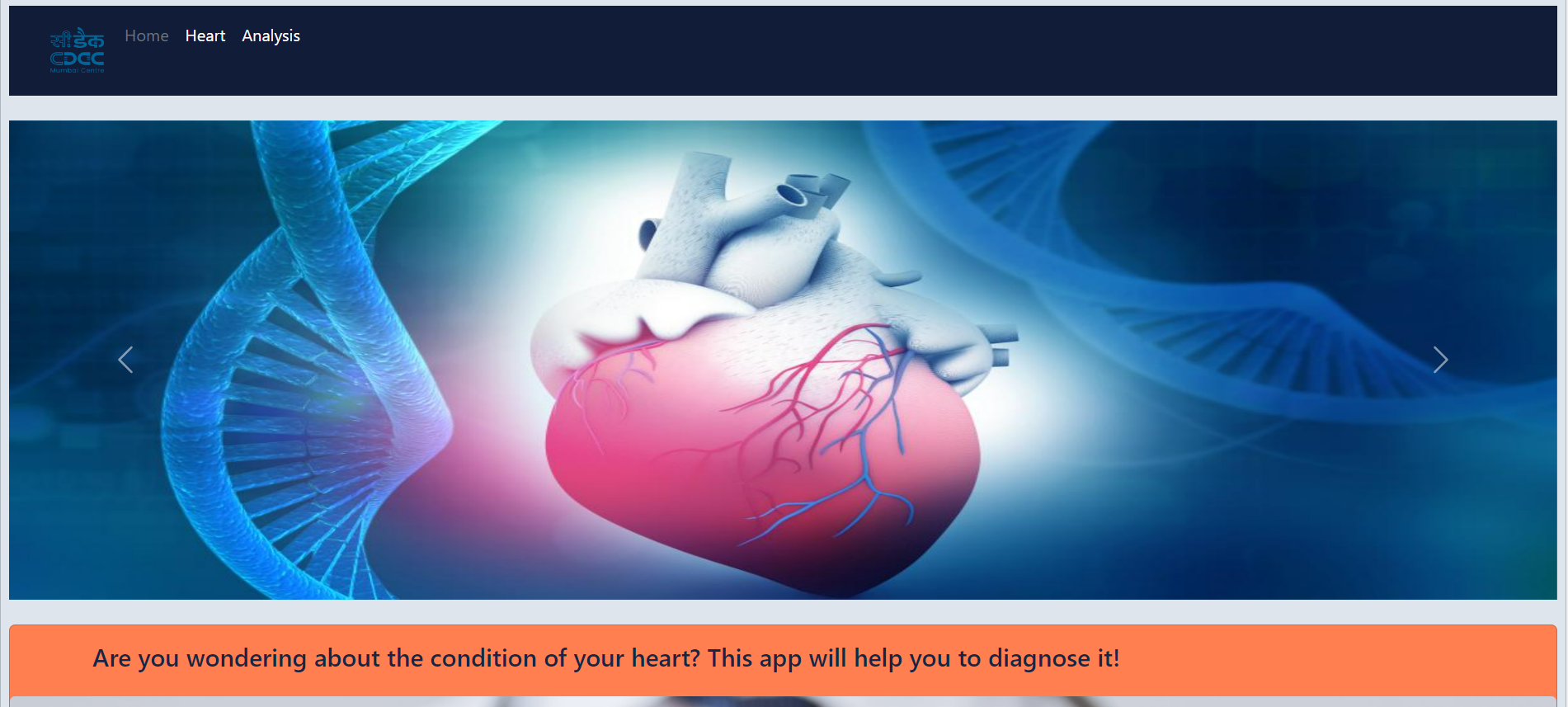
XGB regression = 86.55 %

1. **APP CREATION (FLASK)**

Flask is a lightweight web framework for Python, designed to make it easy to build web applications quickly and with minimum code. It provides a simple and flexible way to create web applications, with a powerful set of tools that can be used to build applications of varying complexity.

Flask provides a built-in development server and debugger, as well as support for unit testing and integration testing. It also supports third-party extensions, making it easy to add functionality to an application, such as support for databases, authentication, and caching.

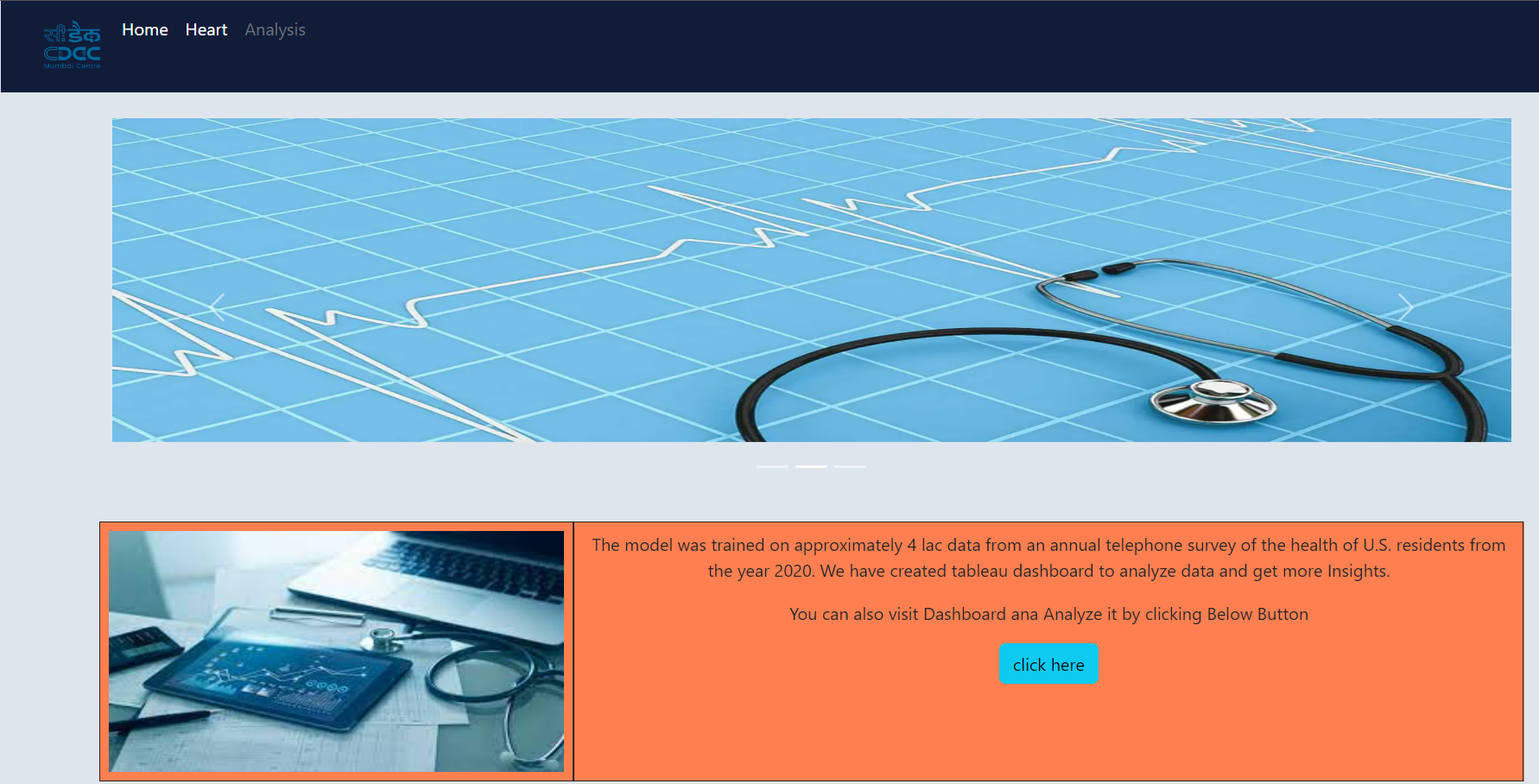
**USER INTERFACE OF WEB APP:**





RESULT:

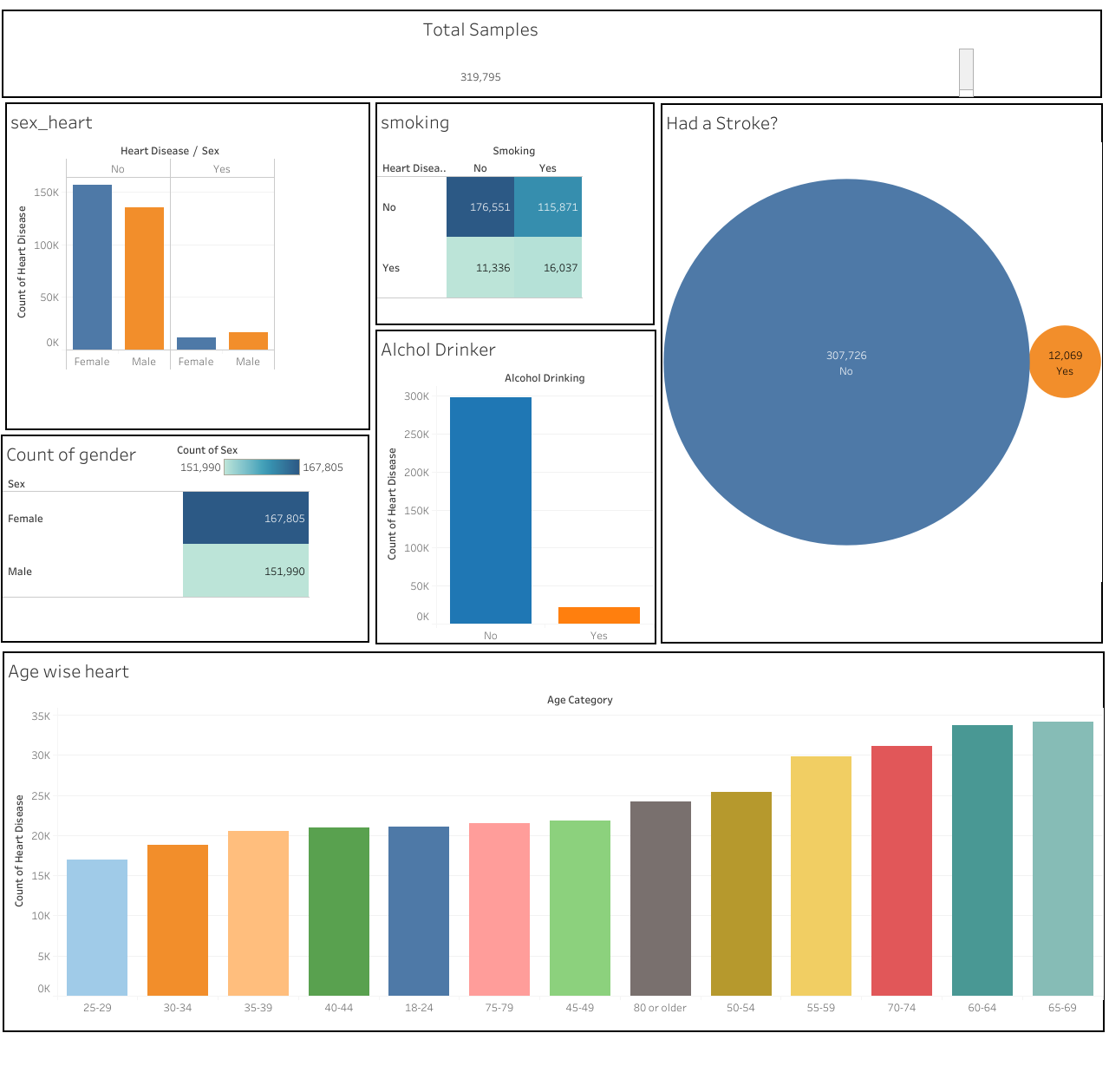




**ANALYSIS RESULT:**

**Data Visualization on Tableau-**

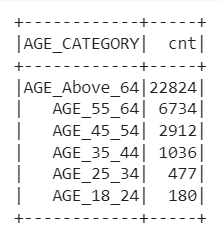
* Gender Visualization on heart Disease
* Smoker Visualization on heart Disease
* Stroke Visualization on heart Disease
* Age Visualization on heart Disease



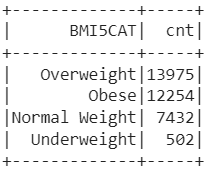
1. **DATA ANALYSIS (SPARK)**

**9.1 Univariate Analysis:**

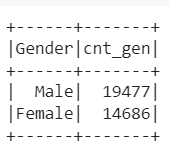
**Analysis on Age**



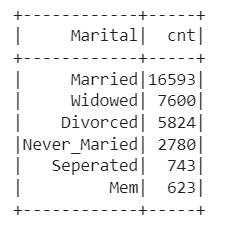
**Analysis on Body Mass Index**



**Analysis on Gender**



**Analysis on Marital**



# **FUTURE IMPROVEMENTS**

* We can make this cloud-based system.
* We can use real time sensors.
* We can use spark for model building as well as for analysis, which will increase the efficiency of our system.
* We can make the system in mobile units.
* More detailed feature engineering and feature selection will be done.
* More data can be used for better prediction.

# **CONCLUSION**

In conclusion, we find that our regression equation is quite accurate (91.68% accuracy) in predicting the heart disease with Random Forest Regressor. People can use it to Predict their Heart Health Status. People need to focus on their Physical Health, General Health, Smoking Habits, Exercises, BMI, Kidney Disease.

1. **BIBLIOGRAPHY**

Dataset link:

<https://www.cdc.gov/>

Models:

1. Linear Regression:

https://scikit-learn.org/stable/modules/linear\_model.html#ordinary-leastsquares

1. Random Forest Regression:

https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomizedtrees

1. Decision Tree:

[https://scikit-learn.org/stable/modules/tree.html#](https://scikit-learn.org/stable/modules/tree.html)

1. Naïve Bayes:

[https://scikit-learn.org/stable/modules/naive\_bayes.html#](https://scikit-learn.org/stable/modules/naive_bayes.html)

5. XGB:

https://scikit- learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html