**Group No: 11**

|  |  |  |
| --- | --- | --- |
| **Roll Number** | **Contribution in the project** | **Name** |
| CB.SC.P2CSE23011 |  | Sirajuddin Khan |
| CB.SC.P2CSE23014 |  | Jose Ignacio Gil |
| CB.SC.P2CSE23020 |  | Patil Sanket Shashikant |

**Title:** Heart Attack Prediction and Analysis

**Problem Statement:**

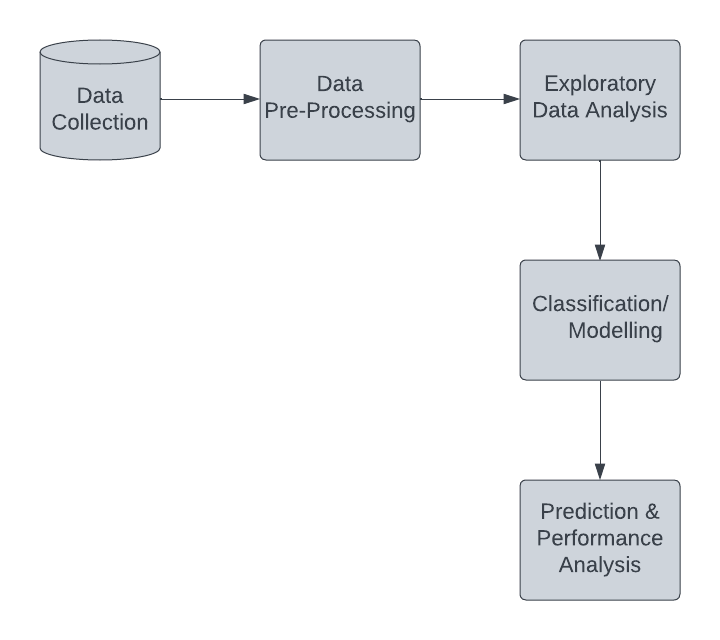
The objective of this project is to develop a machine learning model that can predict the risk of a heart attack for individuals based on their health attributes and medical history. The dataset contains a collection of features such as age, gender, cholesterol levels, blood pressure, smoking habits, and other relevant health indicators. The model should be able to classify individuals into two groups: those at high risk of a heart attack and those at low risk. This predictive model will assist healthcare professionals in identifying individuals who may require preventative measures or early intervention to reduce the risk of a heart attack. The project aims to achieve high accuracy and reliability in heart attack prediction while also providing insights into the factors that contribute to heart disease.

The dataset for this project would include both input features (independent variables) and the target variable (whether or not an individual has experienced a heart attack). Some common features in such datasets might include age, sex, blood pressure, cholesterol levels, chest pain type, electrocardiographic results, exercise-induced angina, and more. The target variable would be binary, where '1' could represent individuals who have experienced a heart attack, and '0' could represent those who have not.

The success of the project would be determined by the accuracy and performance of the machine learning model in making accurate predictions, as well as the insights gained from the analysis of the dataset regarding the risk factors associated with heart attacks.

**Technology Details**: Category -1 (Web Application)

**Methodology**:



1. **Data Collection**:

The dataset employed is taken from the Kaggle Website. It contains 14 attributes, including the predictor attribute. The predictor feature, usually referred to as the ‘target’ refers to the cardiovascular condition of every patient. If the target variable is assigned 0 in the dataset, it means that the patient is normal and if it is assigned 1, it indicates a heart condition.

1. **Data Pre-Processing**:

Various pre-processing steps have been carried out on the dataset to remove null and garbage values, as they aren’t required in the classification and do not add any importance significantly. After pre-processing, the clean data is obtained and is carried forward for the subsequent processes.

1. **Exploratory Data Analysis**:

It is just a way of analysing any given database visually to obtain a clearer understanding. Here, preliminary analyses of the data have been conducted to look for trends, identify any problems, probabilistic reasoning, and validate presumptions. It makes it simple to comprehend the structure of a dataset, making data modelling easier. The main objective is to clean the data, implying that it should be devoid of redundancies.

This kind of analyses assists in identifying the incorrect data points, so that it may be easily changed to make the data manageable for further processing. It also helps in grasping the relationship between the variables, providing us with a wider view of the data utilized in the process, and making allowances to dilate it. Having a data analysis also aids in the evaluation of the dataset’s statistical measurements. Outliers or abnormal occurrences can impact the accuracy of machine learning models. Thus, exploratory data analysis is used to eliminate or resolve all of the dataset’s undesirable qualities.

1. **Classification / Modelling**:

The classification has been conducted using the following four machine learning algorithms.

* **Logistic Regression**: A computerized classification scheme called logistic regression is employed to forecast the probability of a feature of interest. The target variables in binary logistic regression should always be binary, and the successful result must be represented by 1 or 0
* **Decision Tree**: As a supervised learning technique, this predictor can be applied both in classification and regression issues. It is a tree-structured classifier, with nodes in the network standing in for the parameters, branches for the rules of classification, and leaf nodes for the results.
* **Random Forest**: The Random Forest is a collective learning method accustomed to solve regression and classification problems. It consists of a variety of decision trees on numerous subsets and takes the common to boost the predictive database performance. It anticipates the result using the recommendations out of each tree and thus the majority votes of predictions. Higher accuracy and overfitting result from the forest’s larger number of prominent trees.
* **Support Vector Machine**: Additionally, classification and regression issues are addressed using the SVM. Finding a hyperplane that creates a border between the overall description is the major goal of this classifier. This hyperplane in a two-dimensional space is only a line. In an N-dimensional space, where N is the total number of features in the dataset, we plot each data point.

1. **Performance Analysis**:

Performance metrics like accuracy, sensitivity, specificity, precision, and F1-Score are attained with the help of confusion metrics.