**HEART ATTACK RISK PREDICTION**

**SDLC (Software Development Life Cycle)**

The Software Development Life Cycle is a systematic process for building software that ensures the quality and correctness of the software built. SDLC process aims to produce high quality software which meets customer expectations. The software development should be completed within the pre-defined time frame and cost.

**SDLC Phases**

The entire SDLC process is divided into the following stages

● Phase 1: Requirement collection and analysis

● Phase 2: Feasibility study

● Phase 3: Design

● Phase 4: Coding

● Phase 5: Testing

● Phase 6: Installation/Deployment

● Phase 7: Maintenance

**Platform Knowledge**

**What is data science?**

Data science is the study of data to extract meaningful insights for business.It is a multi disciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

**Importance of data science**

Solves real problem using data.Modern organizations have lots of data.Online systems and payment portals capture more data in the fields of e-commerce, medicine, finance, and every other aspect of human life. We can process the information and predict the results

1. **Problem statement**

Implementing ML algorithms to find the best model which accurately predicts the risk of getting heart attack in patients.

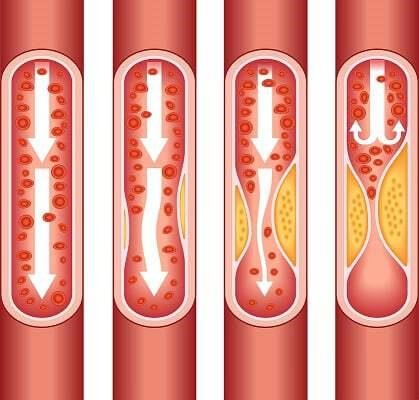
1. **Market need assessment**

A hospital(client) wants us to predict the probability of heart attack happening to their patients. This will help them to take proactive health measures such as promoting new health related schemes. The existing software used by the hospital is time consuming and is unable to adapt to new patterns that cause heart attack.Our ML model will help in making accurate predictions.This will help the hospital to save more patients.It will improve their infrastructure and facilities.

1. **Target specification**

Predicting and diagnosing heart attack is the biggest challenge in the medical industry and it is based on factors like physical examination, symptoms, and signs of the patient. Factors which influence heart attack are cholesterol levels of the body, smoking habits, obesity, family history of diseases, blood pressure and working environment. Machine learning algorithms play a vital and accurate role in predicting heart attack. With the use of this project, we can use classification like ML algorithms to predict the risk of a heart attack.

Heart attack is perceived as the deadliest disease in the human life across the world. This type of disease, the heart is not capable of pushing the required quantity of blood to the remaining organs of the human body to accomplish the regular functionalities. Heart diseases are concertedly contributed by hypertension, diabetes, overweight and unhealthy lifestyle.



As plague builds up in the arteries of a patient,inside of the arteries begin to narrow,which blocks the flow of blood.Plagues can also break.It causes blood clot on the plaque which blocks flow of blood.

1. **External Search(References/links)**

In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10-fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximum accuracy of 91.2 percent.

Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K Neighbors Classifier: 87%, Support Vector Classifier: 83%, Decision Tree Classifier:

79% and Random Forest Classifier: 84%.

1. **Bench marking alternate products**

The existing software used by the hospital is time consuming and is unable to adapt to new patterns that cause heart attack.Our ML model will help in making accurate predictions.This will help the hospital to improve their infrastructure and facilities.The client hospital can use our model in their database to analyse their patients.Based on the results,the hospital can provide special care and facilities to patients who are on the verge of getting an heart attack.Medical help like medicines and exercises will be prescribed to such patients,ultimately saving their life.

1. **Applicable constraints**

Lack of qualified staff: Finding qualified medical officers with in-depth understanding of heart attack risk management can be difficult.Medical institutions and hospitals must provide trainings to their staff so that they are up to date in the medical field.Regular checkups for risk prone patients is needed.It is difficult to execute as it depends on multiple factors like availability of premium patients,infrastructure and bill.

**DATA DESCRIPTION AND PREPROCESSING:**

**Data Dictionary:**

* + **Patient\_ID:** Unique ID of different patients.
  + **Gender:** Gender of the patient.
  + **Age:** Age of the patient.
  + **HyperTension:** A person has history of Hypertension or not •

**Heart\_Disease:** A person has history of heart disease or not.

* + **Is\_Married:** Whether the person is married or not.
  + **Employment\_Type:** Determines whether the patient is a working professional in a Private/Govt sectors, never worked or children.
  + **Residential\_type:** Specifies whether the patient is from Urban/Rural areas.
  + **Glucose\_Levels:** Average glucose levels of a patient.
  + **BMI\_Values:** Considering height and weight of a patient.
  + **Smoking\_Habits:** Classifies whether the patient is a regular smoker, past smoker or never smoked.
  + **Heart\_Attack:** Chances of getting heart attack (Dependent Variable)

**DATA PREPARATION:**

* + Null values in the column are replaced by median value and stored.
  + Null values in ‘Smoking\_Habits’ are filled with ‘never smoked’. Null value imputation done for both the independent variables.

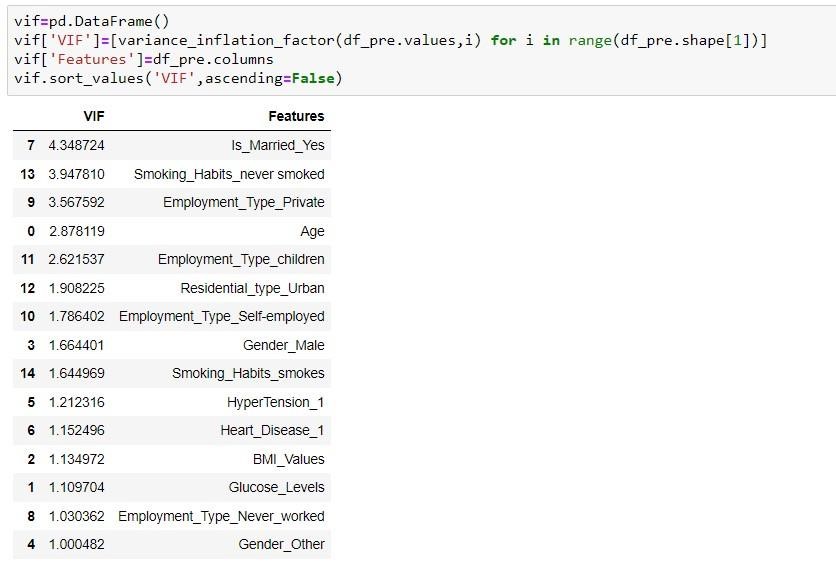
Outlier treatment:

* + We plot boxplot on log transformed values of BMI\_Values. Outlier values are also displayed.
  + Eliminating outliers based on IQR technique. When we plot boxplot, the outliers are neglected.
  + We plot distplot on outlier treated BMI\_Values and we get a near normal distribution.
  + Skewness of Glucose\_Levels is very high. We plot boxplot on the column. We infer that this column has many outliers too.
  + Glucose\_Levels is highly positively skewed. By using skew function.

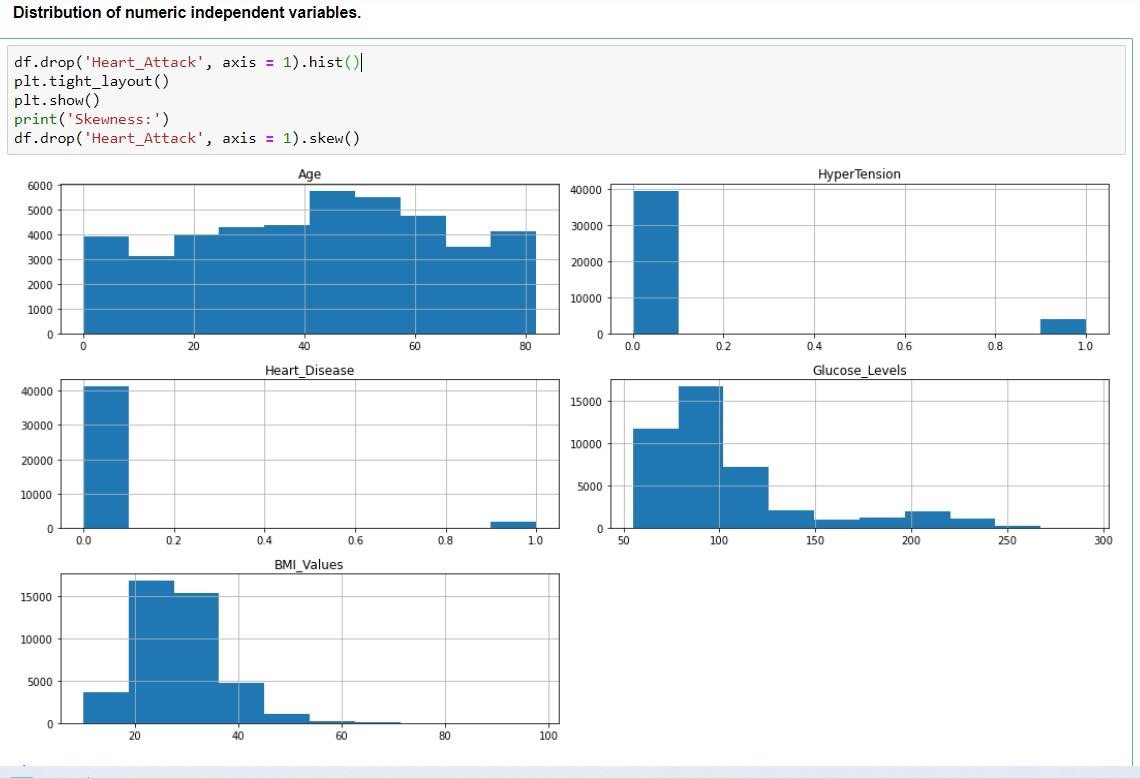
**EXPLORATORY DATA ANALYSIS & BUSINESS INSIGHTS:**

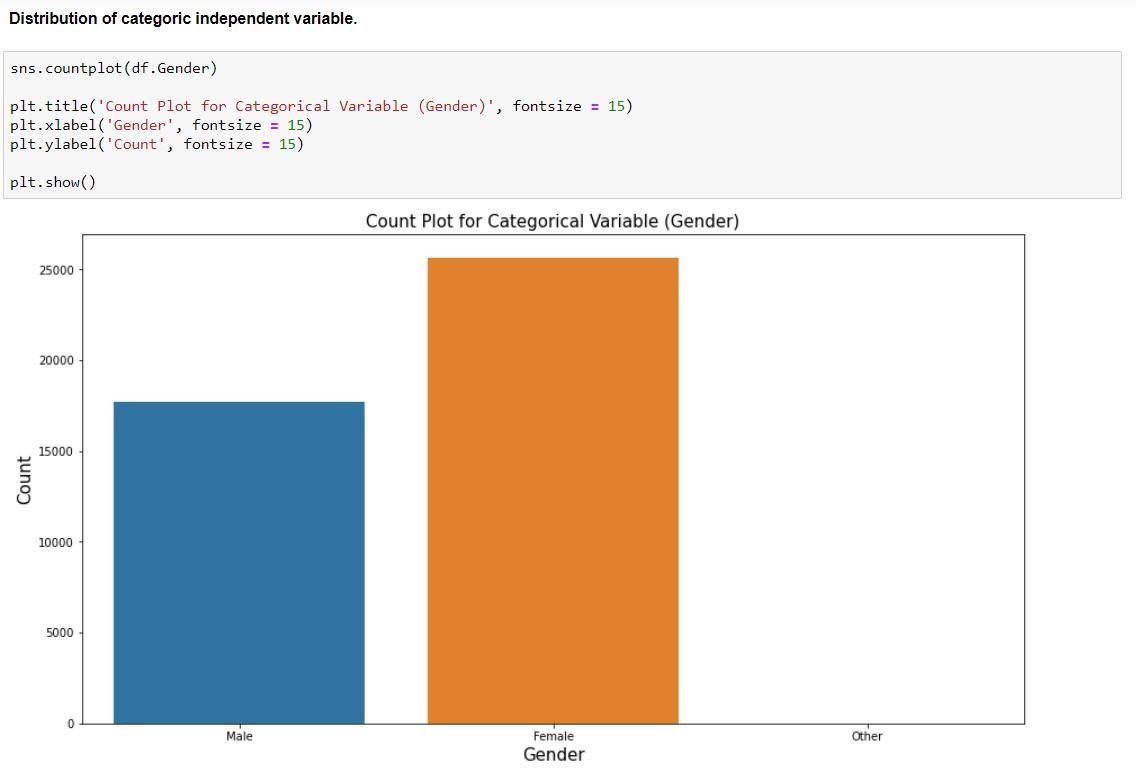
Multi-collinearity:

As VIF values are less than 5, we can conclude that there is no multi collinearity amongst independent variables.

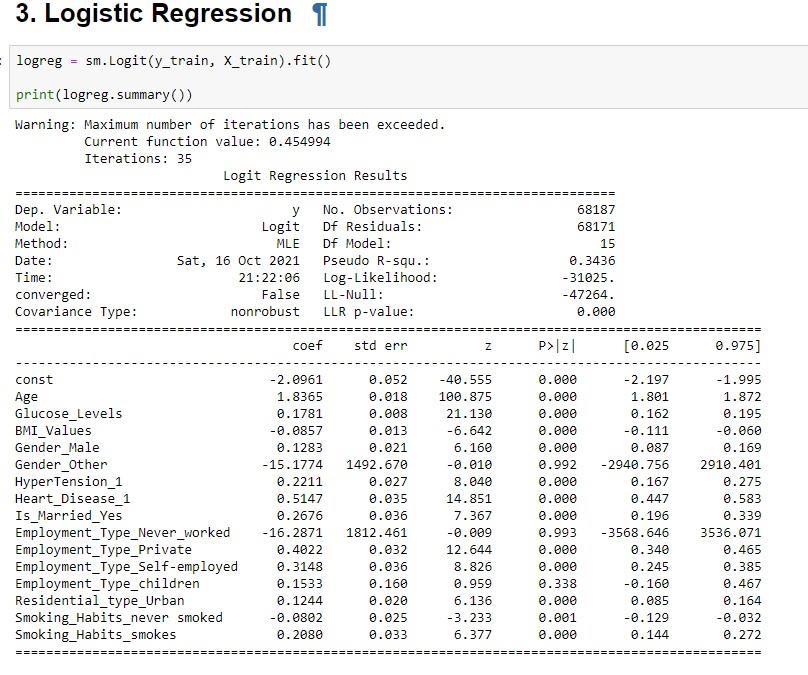


**Distribution of Variables:**



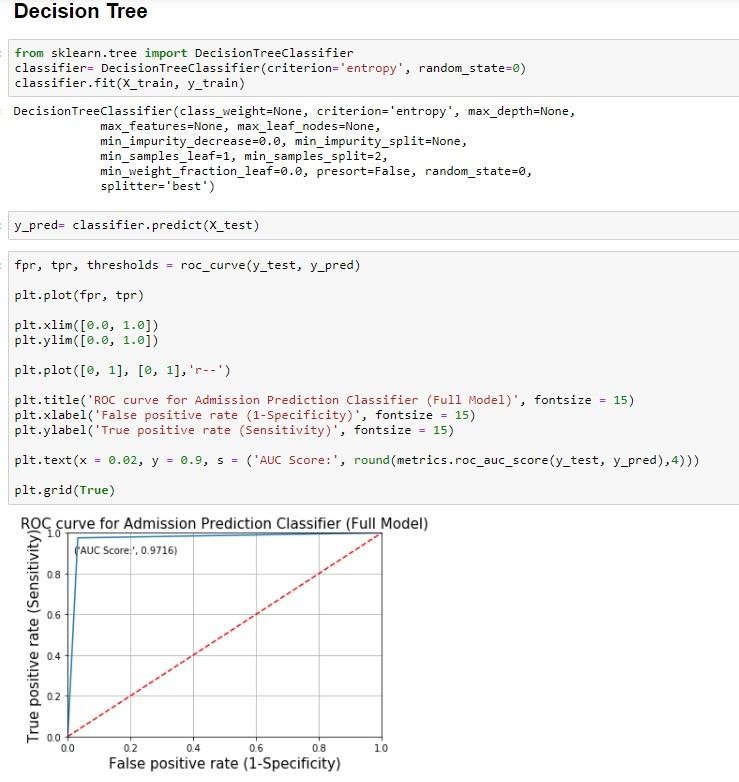


**BASIC MODEL:**



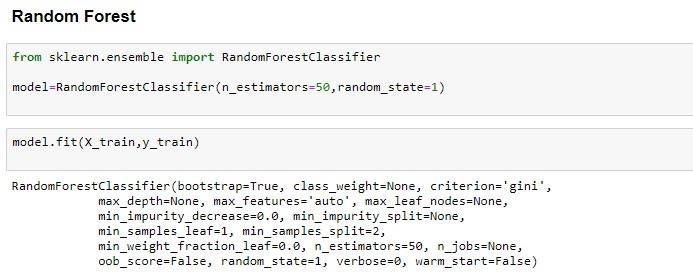


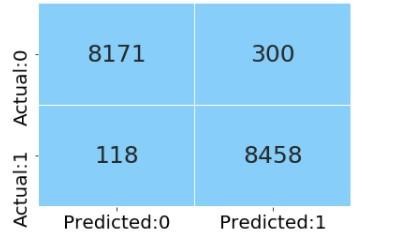
**Accuracy score for Logistic Regression method is 0.8578.**





**Accuracy score for Decision Tree method is 0.9716.**





**Accuracy score for Random Forest method is 0.9754.**

**FEATURE ENGINEERING & FEATURE EXTRACTION:**

**Transformation:**

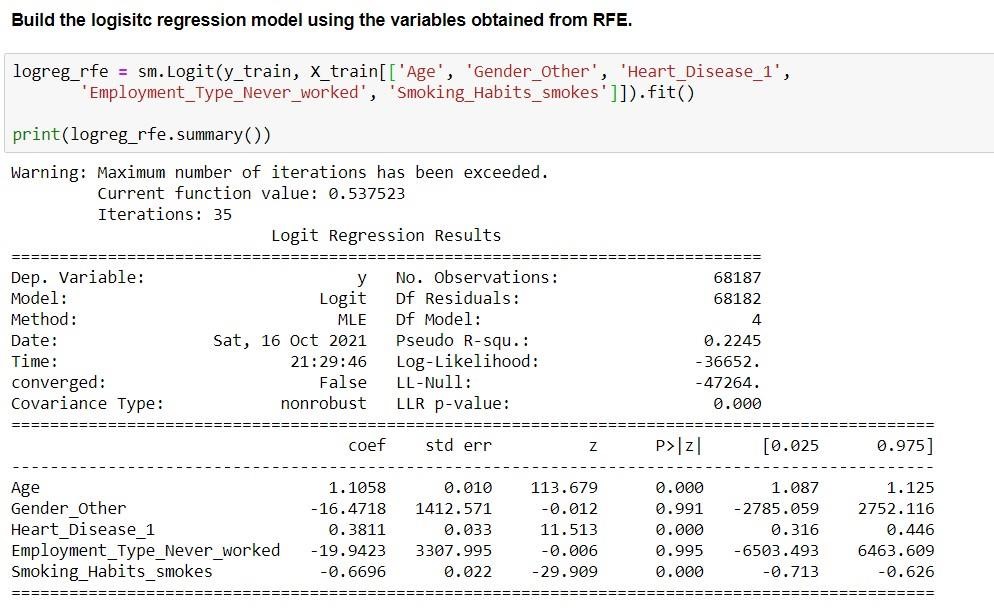
* All the categorical variables are stored in a new dataframe named df1\_cat.
* Dummy variables are created for all categorical variables using get\_dummies function from pandas and all encoded columns are stored in df1\_cat. .(one hot encoding).

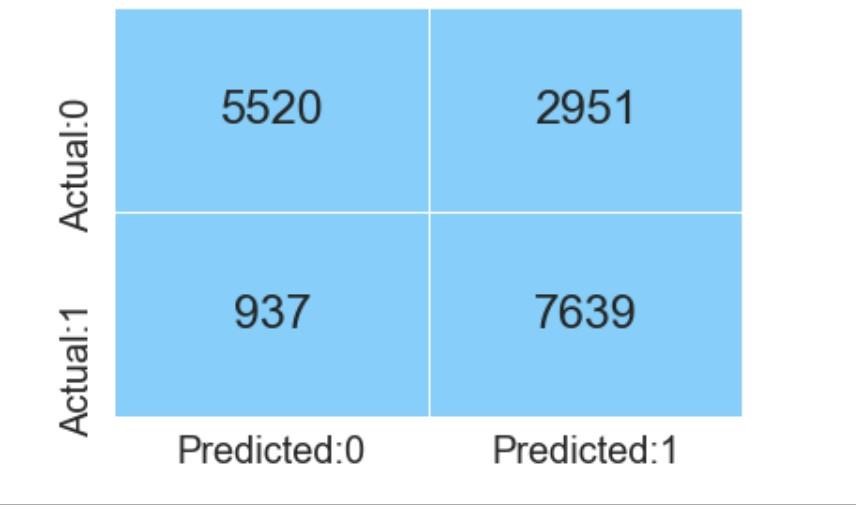
**Scaling the data:**

* Performing standardization on numerical dataframe num\_df using standard scaler function from sklearn.preprocessing module.
* We are concatenating the scaled numerical variables and encoded categorical variables and storing it in df\_pre.df\_pre is the preprocessed dataframe ready for modelling.
* Concatenating preprocessed df\_new with target variable to form a dataframe df\_corr.We are plotting heatmap on df\_corr to find correlation amongst two independent variables.

**Recursive Feature Elimination (RFE):**

* In the linear Classification module, we learn about various techniques for selecting the significant features in the dataset. In this example, let us consider the RFE method for feature selection.





**Accuracy score for RFE method is 0.8578.**

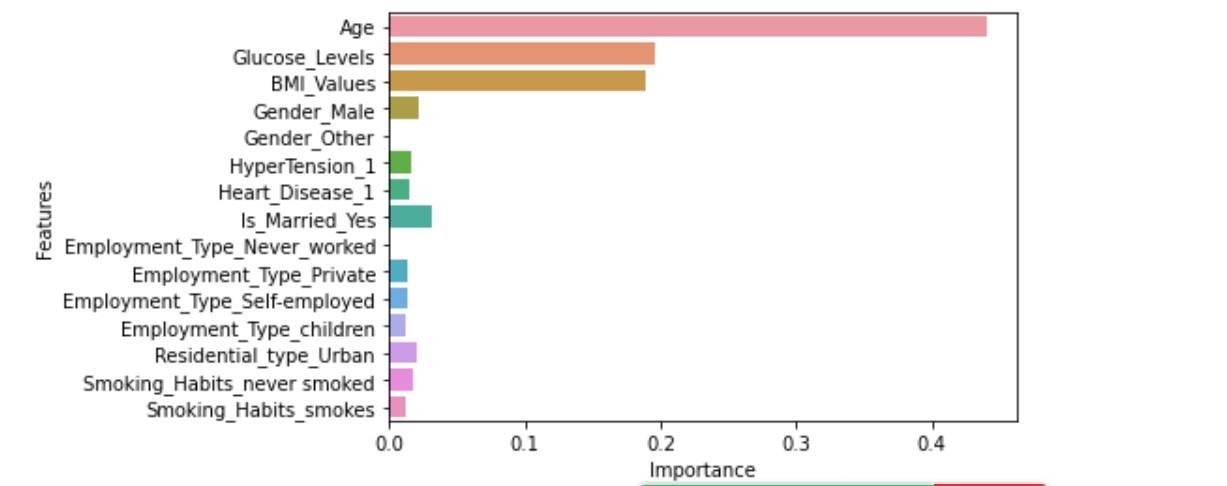
**HYPER PARAMETERS TUNING:**

We have found out the best parameters using Grid Search CV.

The parameters are as follows:

'criterion': 'gini', 'max\_depth': 10, 'max\_features': 'sqrt', 'max\_leaf\_nodes': 9, 'min\_samples\_leaf': 1,

'minsamples\_split': 2, 'n\_estimators': 30



Age is the most important feature in prediction of Heart attack of a patient.

Feature Engineering:

Transformation: All the categorical variables are stored in a new dataframe. One hot encoding:Dummy variables are created for all categorical variables using get\_dummies function from pandas and all encoded columns are created. Taking the best estimators, training algorithm is performed on Test data.

The min-max scaling method is used for numerical features.

Min-max scaling is similar to z-score normalization in that it will replace every value in a column with a new value using a formula. In this case, that formula is:

*m = (x -xmin) / (xmax -xmin)*

Where:

• *m* is our new value

• *x* is the original cell value

• *xmin* is the minimum value of the column

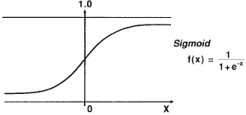
• *xmax* is the maximum value of the column

Using this formula, we will see that the values of each column will now be between zero and one.

**Machine Learning Algorithms:**

**1)Logistic Regression:**

It is one of the most simplest and popular machine learning model.



This model takes the input values as x and gives the output values as f(x) i.e. 0 or 1. If I need to built a machine learning model then each data point of independent variable will be x (i.e. sum of x1 \* w1 + x2 \* w2 . . . .so on ) and this will give a value that is between 0 to 1. If I consider that 0.50 as deciding value or threshold. Then any result above 0.5 would be taken as 1 and below that as 0.

**2.Decision Tree**

The first step is to split the labelled data into train and test data.The tree begins with the root node which consists of the entire training data.The best attribute is found using the Attribute selection measure.Entropy: (chances of being incorrect).Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. When the data is highly impure, or highly pure, the entropy is 0. Entropy can be calculated as:

Entropy(s)= -P(Yes) \* log2 P(Yes) - P(No) \* log2 P(No)

Entropy(S) = 1 when P(Yes) = P(No) = 0.5.

Entropy(S) = 0 when P(Yes) = 1 or 0

Information Gain:

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.It calculates how much information a feature provides us with a class.According to the value of information gain, we split the node and build the decision tree.A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain = Entropy(S)- [(Weighted Avg) \*Entropy(each feature)]

Select the feature with maximum information gain and divide the dataset. Recursively make new decision trees using the subsets of the dataset created in above step.Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

**3.Random Forest**

A random forest algorithm is a supervised learning algorithm that can be used for both classification and regression tasks. It is an ensemble learning method, which means that it combines the predictions of multiple individual models to produce a more accurate prediction.

Random forest algorithms work by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

Random forest algorithms are more robust to overfitting than individual decision trees because they average the predictions of multiple trees. This is because each tree in the forest is trained on a different bootstrap sample of the training data and a different random subset of features.

**Hardware Requirements:**

• CPU: A modern multicore processor, such as an Intel Core i5 or i7, is recommended for training and deploying machine learning models.

• RAM: A minimum of 8GB of RAM is recommended, although more may be required for larger datasets.

• GPU: A powerful graphics processing unit (GPU) can significantly speed up training and inference of deep learning models. Nvidia GPUs are the most commonly used for machine learning tasks.

• Storage: A large amount of storage is required to store and process large datasets. Solid-state drives (SSDs) are recommended for fast access to data.

**Software Requirements:**

• Operating System: Most machine learning frameworks are compatible with popular operating systems such as Windows, macOS, and Linux.

• Python: Python is the most commonly used programming language for machine learning, and most machine learning frameworks are built using Python. • Machine Learning Frameworks: There are many machine learning frameworks available, such as TensorFlow, PyTorch, Keras, and scikit-learn. These frameworks provide tools and APIs for building, training, and deploying machine learning models.

• Development Tools: IDEs such as PyCharm, Jupyter Notebook, and Visual Studio Code are commonly used for machine learning development.

• Libraries: Libraries such as NumPy, Pandas, and Matplotlib are commonly used for data manipulation, analysis, and visualization.

In summary, to get started with your project, you will need a computer with a powerful CPU and GPU, a minimum of 8GB of RAM, and large storage capacity. You will also need to install Python and various machine learning libraries and frameworks.

**COMPARISON AND SELECTION OF MODEL:**

Based on our comparison done on all the Classification related algorithms, we found out that Random Forest model gives us the best accuracy score of

0.9754 and has the least False Negative values. Random Forest has the highest accuracy in predicting the correct classes.

**Future work of the Project**

In the future, this research will be expanded by identifying and integrating new features of heart disease. It also intends to employ other classification methods, such as deep learning to optimize the prediction. The goal is to study and merge more datasets in order to create a more relevant dataset that encompasses a broad range of population types. The feature selection can be used to generate more relevant features and effective results for the prediction of heart disease.

**7. Conclusion:**

We proposed three methods in which comparative analysis was done and promising results were achieved. The conclusion which we found is that

Random Forest machine learning algorithm performed better in this analysis.

**DESCRIPTION OF CRITERION:**

The methods which are used for comparison are Confusion Matrix, Precision, Specificity, Sensitivity, and F1 score. For some features which were in the dataset, Random Forest and decision tree classifier algorithms performed better in the ML approach when data preprocessing is applied.

The dataset size can be increased and then Machine learning with various other optimizations can be used and more promising results can be achieved.

Machine learning and various other optimization techniques can also be used so that the evaluation results can again be increased. More different ways of normalizing the data can be used and the results can be compared. And more ways could be found where we could integrate heart-disease-trained ML models with certain multimedia for the ease of patients.