CHAPTER 1

INTRODUCTION

Abstract

According to the WHO, approximately 17.9 million deaths result from heart disease which counts for about 32% throughout the world in 2021. In non-infective disease, heart disease become a leading cause of death. For that, prediction of heart disease in early stage is important to reduce the death rate. ML algorithms play a vital role in healthcare for predict and early diagnosis. There are many studies and project proposed to predict the heart disease earlier. Hybrid Random Forest with a Linear Model (HRFLM) used and obtained result with accuracy of 88.7%. Here 11.3% of inaccuracy is big number and cause large mortality in present world. Ensemble learning technique (Max voting) to predict the heart disease which is used in this project to provide more accuracy in result and overcome that drawback. Main aim of this project to provide a more accuracy in real time.

1. INTRODUCTION

Mortality rate due to heart disease is increasing day by day in the world. Most of them are low and middle countries.17 million people are under the 70 age in affected people. There are many types in the heart disease. They are Coronary Artery Disease (CAD), heart arrhythmias, heart failure, heart valve disease, endocarditis, rheumatic heart disease, pericardial disease, cardiomyopathy, congenital heart disease. Major causes of the heart disease are smoking, obesity, blood pressure, genetic, cholesterol, unhealthy food and life style. The risk of lives is high if heart disease stage crossed the initial stage Today healthcare industry has become a big money-making business. Money saving is very important for middle class and lower middle-class people. This industry produces large amount of data. This data is used for effective and best possible treatment for patient’s health. By machine learning, heart disease can predict easily using various medical parameters such as blood pressure, Body Mass Index (BMI), obesity, early attack and patient medical history. Early prediction of disease will reduce lives risk and cost of treatment. The major challenge is to extract the correct information by selecting the correct features because of presents of huge amount of data. For that, various feature selection will apply and choose the best one. Accuracy of prediction of disease is more important to save lives. Therefore, main objective of this project is to provide result with best accuracy. In this project, logistic regression, decision tree, random forest, k-nearest neighbour to improve the accuracy of the result.

CHAPTER 2

LITERATURE REVIEW

In [1], aims at finding significant features and improving the accuracy in the prediction of cardiovascular disease. The 88.7% of enhanced accuracy is obtained [1]. The [2] proposed a ML based system that predicts presence of disease. The dataset comprises of 133 columns of 132 varied symptoms experienced by patients suffering from a range of alignments. Total of 40 disease are present in that dataset. They got accuracy more the 90% and used K-Nearest Neighbor [K-NN], Logistic Regression [LR], Decision Tree [DT], Naive Bayes [NB], Linear Discremental Analysis [LDA], Random Forest [RF] algorithms for prediction.[6] in the model UCI ML depository used for analysis NB, DT, Support Vector Machine [SVM], RF, Convolutional Neural Networks [CNN]. They gave 84.5%, 11.9 seconds accuracy on common disease,86.89% accuracy in heart diseases, above 90% accuracy in breast disease, nearly 85% for parkinson’s disease. LDA and PCA to select essential features from the dataset [7]. K-NN, SVM, DT, RF, NB, Ensemble algorithms are used for application to predict the disease. Data split, 75% for training data and 25% foe test data, gives best performance with 98.7% recall and 98% Area Under Curve [AUC] and 98% precision, while the worst performance achieved by Navies Bayes [NB]: 83 % of accuracy, 88% of recall, 81.9 % of precision, 85% of F-measuring and 92% of Area Under Curve. They used data from [8]. The researcher in [3] contributes ML based model using Logistic Regression [LR], Random Forest [RF] with accuracy of 80% with precision 0.96. In [4] used K-Nearest Neighbour and Naive Bayes for developing web-based model with accuracy of nearly 90% of accuracy. Logistic Regression, K-Nearest Neighbour, Random Forest classifiers with accuracy of 87.5% are used in [5].[9] research work aims to design a framework for heart disease prediction by using major risk factors based on different classifier algorithms such as Naïve Bayes (NB), Bayesian Optimized Support Vector Machine (BO-SVM), K-

Nearest Neighbours (KNN), and Salp Swarm Optimized Neural Network (SSA-NN). This research is carried out for the effective diagnosis of heart disease using the heart disease dataset available on the UCI Machine Repository. The highest performance was obtained using BO-SVM (accuracy of 93.3%, precision = 100%, sensitivity = 80%) followed by SSA-NN with (accuracy = 86.7%, precision = 100%, sensitivity = 60%) respectively.[10] analyses the supervised learning models of Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbours, Decision Tree, Random Forest and the ensemble technique of XG Boost to present a comparative study for the most efficient algorithm and found that Random Forest provides most accuracy with 86.89% in comparison to other algorithms.

CHAPTER 3

PROPOSED MODEL

The Proposed Model of Predicting Heart Disease:

The objective of the proposed model is to improve the accuracy and performance by using large data set and efficient ML algorithms. It consists of four stages such as Data collection, Data Processing, Model Development and Evaluate Model.

Data Collection:

The heart disease dataset is utilized for training and evaluating models. It consists of 31994 records,17 features and one target column. The target column consists of two class:1 indicates person suffering from heart disease and 0 indicates normal person without heart disease.

Dataset description:

The dataset contains 18 variables (9 Booleans, 5 strings and 4 decimals). In machine learning projects, "Heart Disease" can be used as the explanatory variable

Features/attributes	**Description**	**values**
Heart Disease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)	True and False
BMI	Body Mass Index (BMI)	Float values, range 12 to 95

Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]	True and False
Alcohol Drinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week.	True and False
Stroke	(Ever told) (you had) a stroke?	True and False
Physical Health	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30	Discrete values range from 1 to 30
Mental Health	Thinking about your mental health, for how many days during the past 30 days was your mental health not good?	Discrete values range from 1 to 30
Difficult Walking	Do you have serious difficulty walking or climbing stairs?	True and False
Sex	Are you male or female?	Male or female
Age Category	Fourteen-level age category	Fourteen-level age category

		Range from 18 or greater
Race	Imputed race/ethnicity value	Any thing
Diabetic	(Ever told) (you had) diabetes?	Yes and No
Physical Activity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job	Yes and No
General Health	Would you say that in general your health is……….	Good, fair, bad, very good, better
Sleep Time	On average, how many hours of sleep do you get in a 24-hour period?	Values range from 1 to 24
Asthma	(Ever told) (you had) asthma?	True and False
Kidney Disease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?	Yes and No
Skin Cancer	(Ever told) (you had) skin cancer?	Yes and No

Data Preprocessing:

The features are scaled to be in the interval [0,1]. It is worth noting that missing values are deleted from the dataset or substitute mean value of feature.

SYSTEM DESIGN:

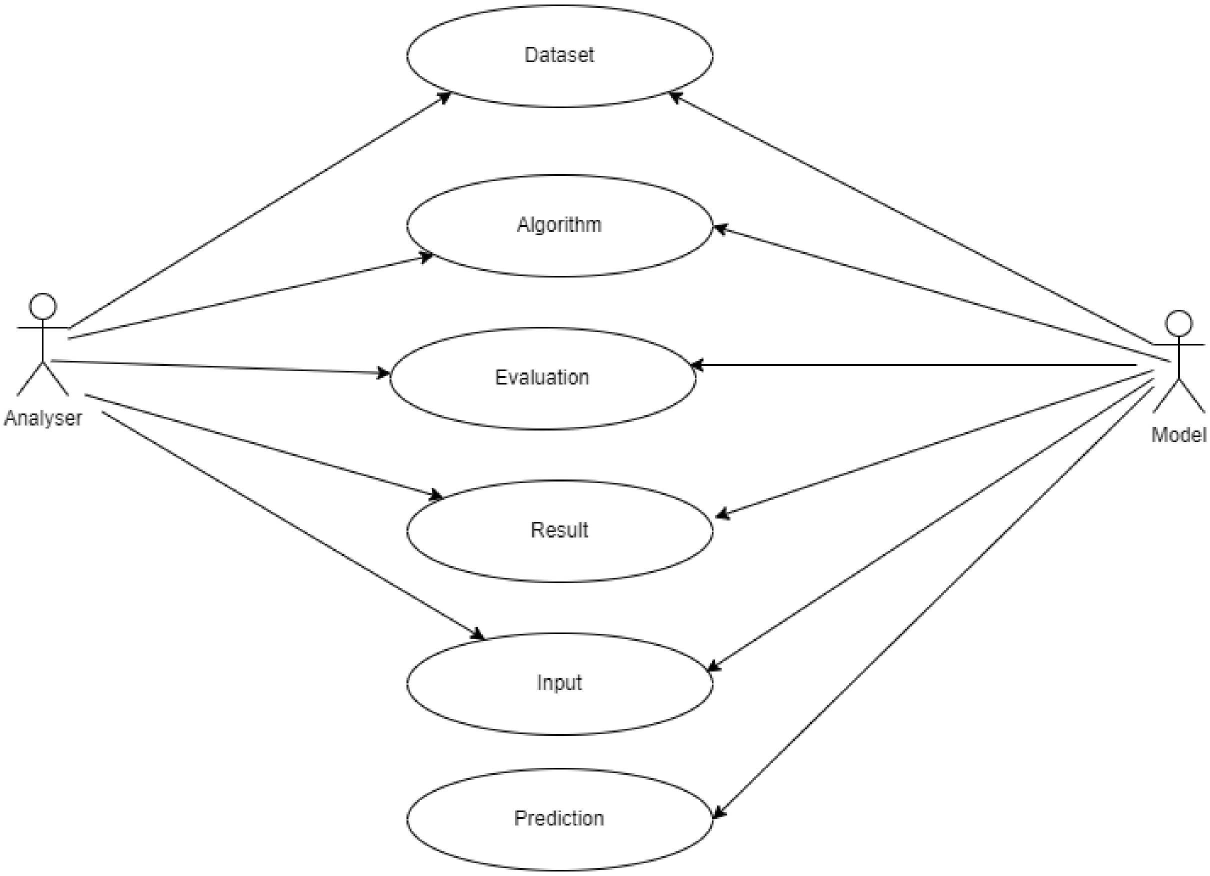
Systems design is the process of defining the architecture, UML diagrams, modules, data flow diagram, Entity Diagram and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to disease analysis.

UML DIAGRAM:

UML (Unified Modelling Language) is a standard language for specifying, visualizing, constructing, and documenting the artifacts of software systems. UML was created by the Object Management Group (OMG) and UML 1.0 specification draft was proposed to the OMG in January 1997. It was initially started to capture the behaviour of complex software and non-software system and now it has become an OMG standard. This tutorial gives a complete understanding on UML.

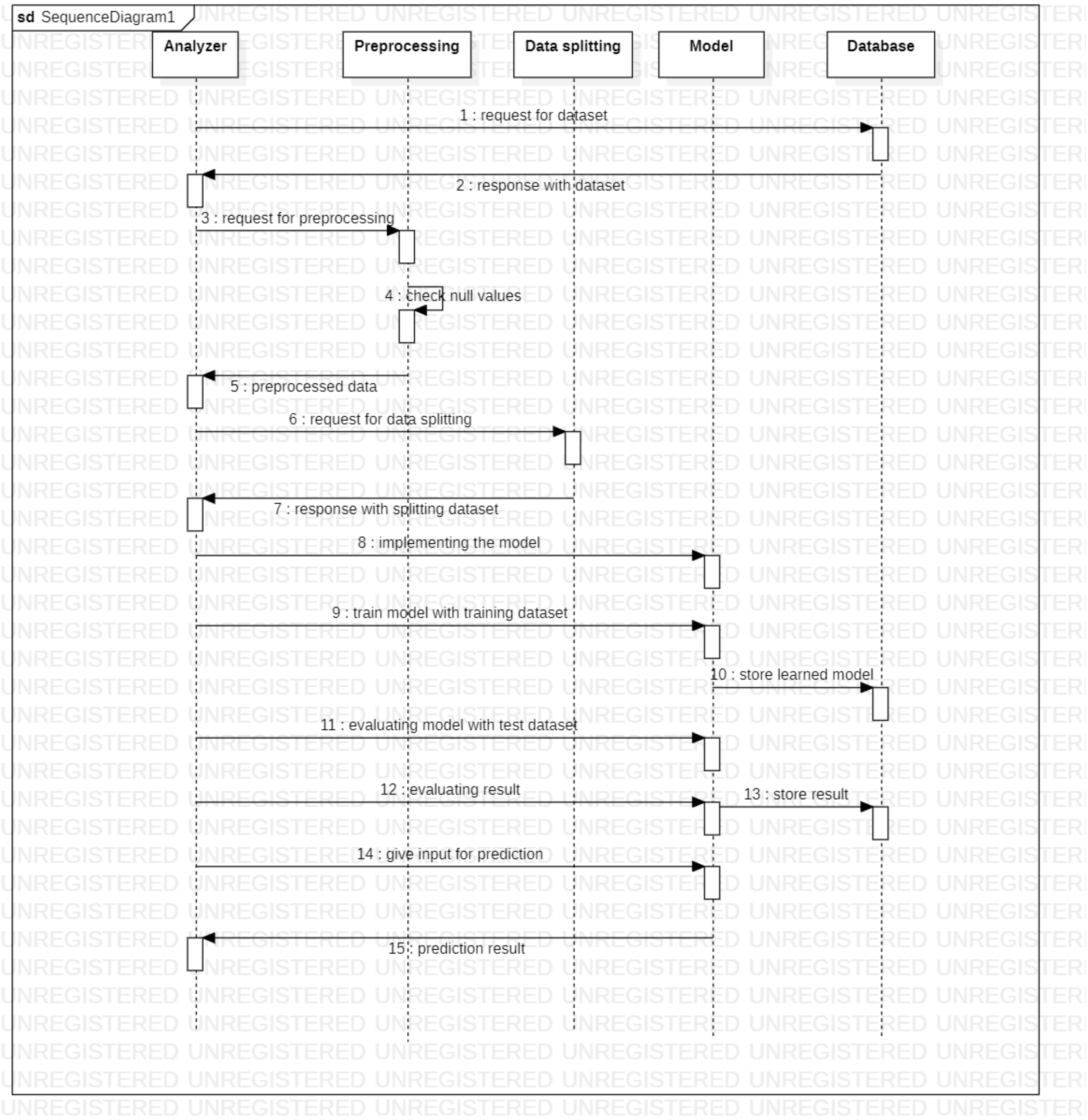
USE CASE DIAGRAM:

A use case diagram is used to represent the dynamic behaviour of a system. It encapsulates the system's functionality by incorporating use cases, actors, and their relationships. It models the tasks, services, and functions required by a system/subsystem of an application. It depicts the high-level functionality of a system and also tells how the user handles a system. The main purpose of a use case diagram is to portray the dynamic aspect of a system. It accumulates the system's requirement, which includes both internal as well as external influences. It invokes persons, use cases, and several things that invoke the actors and elements accountable for the implementation of use case diagrams. It represents how an entity from the external environment can interact with a part of the system.



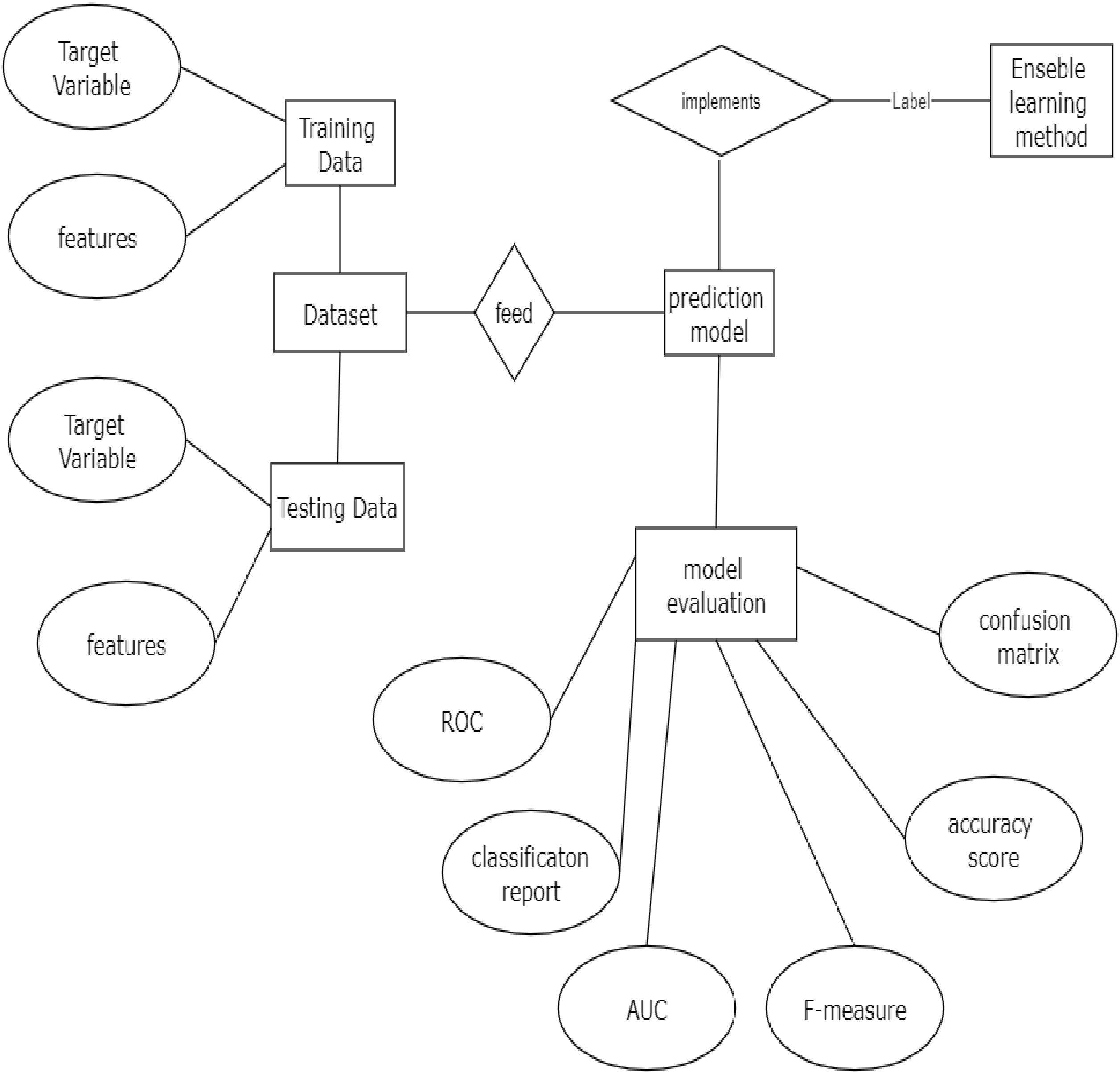
SEQUENCE DIAGRAM:

A sequence diagram is a type of interaction diagram because it describes how—and in what order—a group of objects works together. These diagrams are used by software developers and business professionals to understand requirements for a new system or to document an existing process. Sequence diagrams are sometimes known as event diagrams or event scenarios.


sd SequenceDiagram1
Analyzer
Preprocessing
Data splitting
Model
Database
1 : request for dataset
2 : response with dataset
3 : request for preprocessing
4 : check null values
5 : preprocessed data
6 : request for data splitting
7 : response with splitting dataset
8 : implementing the model
9 : train model with training dataset
10 : store learned model
11 : evaluating model with test dataset
12 : evaluating result
13 : store result
14 : give input for prediction
15 : prediction result

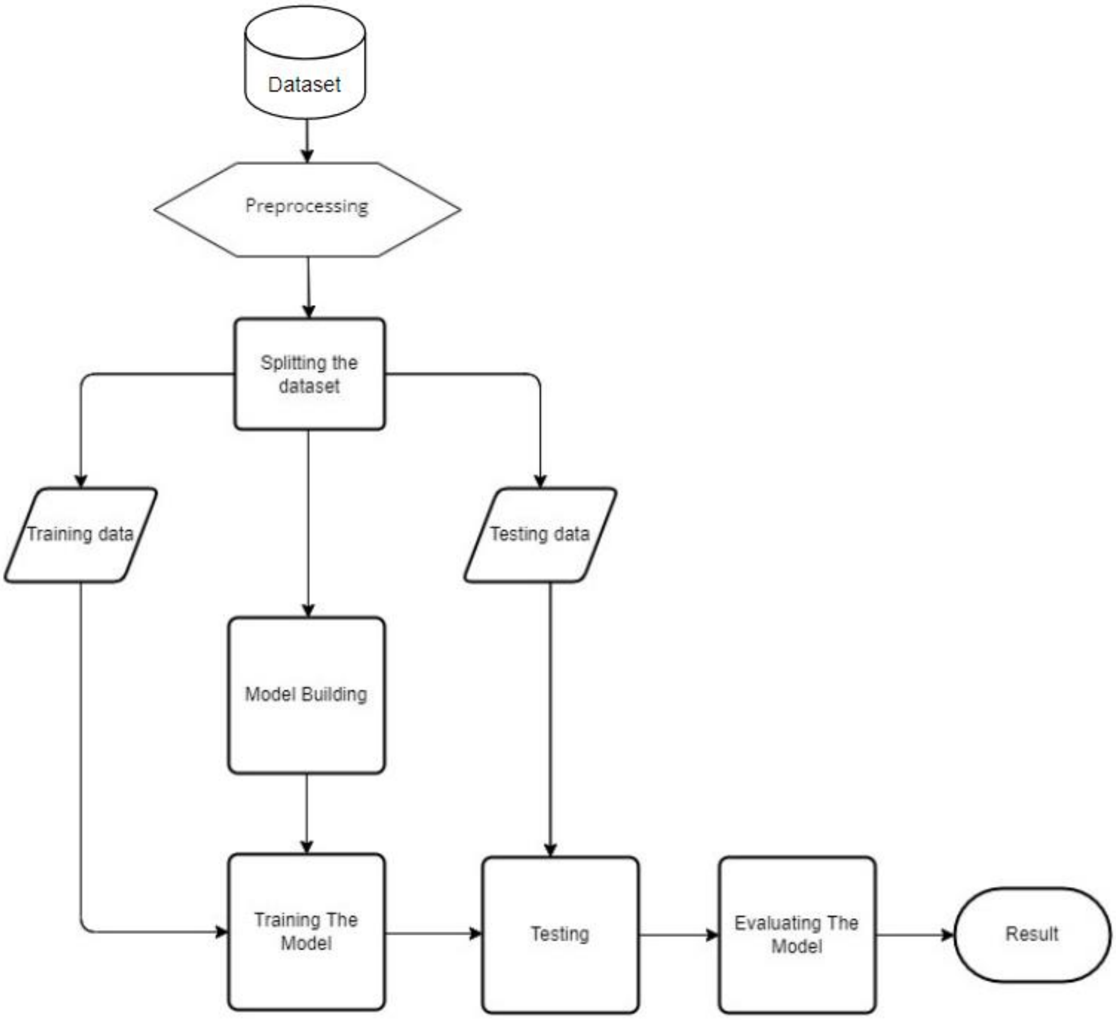
ER DIAGRAM:

ER diagram is a high-level conceptual data model diagram. ER model helps to systematically analyse data requirements to produce a well-designed database. The ER Model represents real-world entities and the relationships between them.ER diagram of the model shows relationship between the entities. Target variable and features are two attributes of the training data and testing data. Both training and testing data has strong relationship with dataset entity. Dataset has feed relationship with the prediction model. Ensemble learning technique is one of best technique to improve the performance of prediction systems which have strong implement relation with prediction model developed. Evaluation measures are used to evaluate the accuracy, performance of developed model or system such as ROC, confusion matrix, classification report, accuracy, F-measure and AUC. Model Evaluation shows how does model developed well and their performance which has strong relationship to the prediction model.



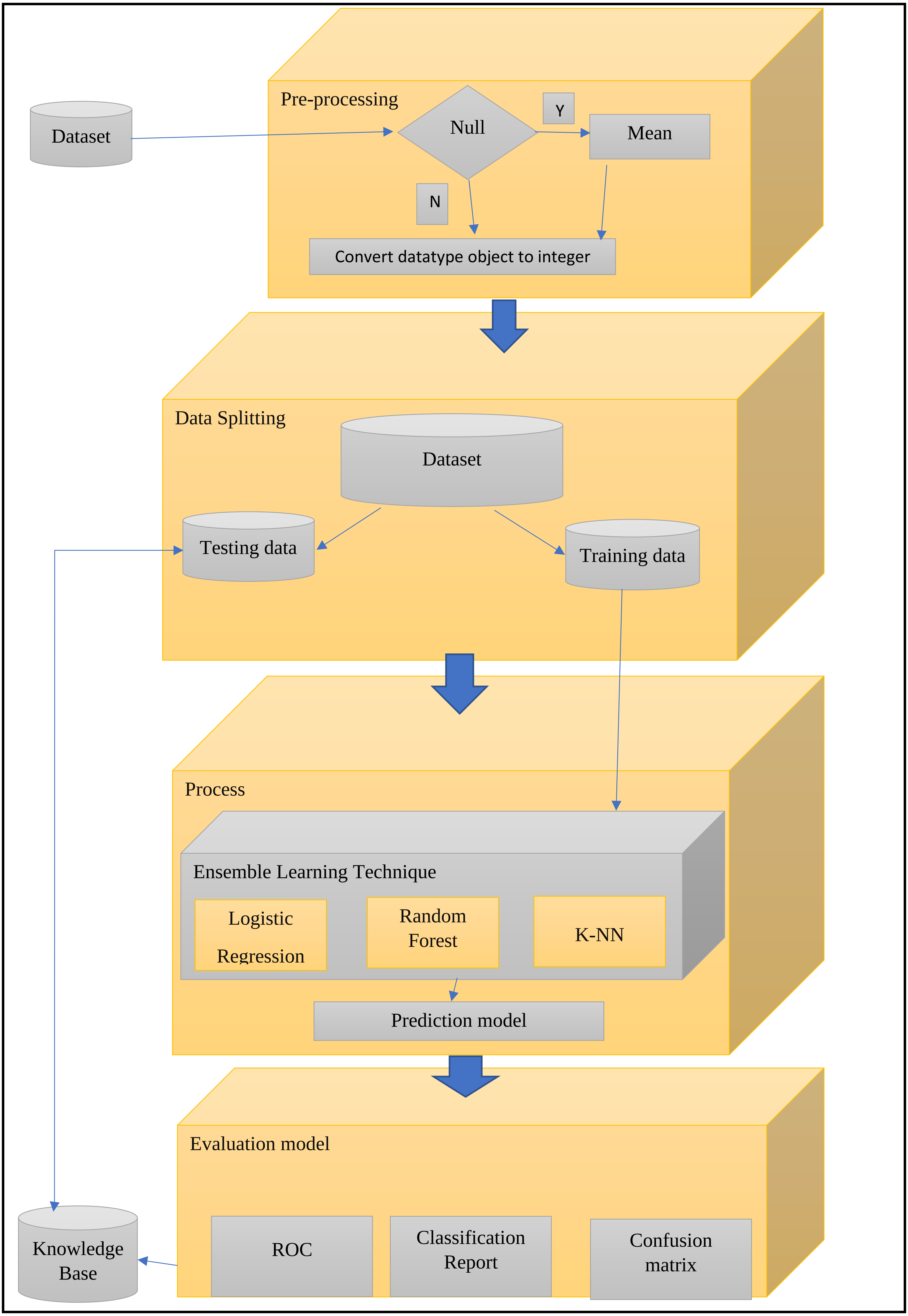
DATAFLOW DIAGRAM:

The Data Flow Diagram represents flow of data through proposed model. In proposed system, dataset is loaded from the database or spreadsheet. The loaded dataset is preprocessed for eliminate null values if null values are present and convert the data type of features into integer or float, the eliminate noise variable. After the data preprocessing, dataset is split into training and testing data. Using training data, the prediction model is developed which build by implementing ensemble learning technique, then test the accuracy of trained model using testing data. After testing, model is evaluated by various statistical measures such as accuracy, recall, precision, F-measure, ROC and AUC.



3.2 SYSTEM ARCHITECTURE DESIGN:

An architectural diagram is a visual representation that maps out the physical implementation for components of a software system. It shows the general structure of the software system. In the proposed system, the dataset is loaded from the database and undergoes pre-processing. After the pre-processing technique, dataset is split into testing data and training data. The training data is used for model development. Then the model is stored in a knowledge base and used later for evaluation.


Dataset
Pre-processing
Null
Y
Mean
N
Convert datatype object to integer
Data Splitting
Dataset
Testing data
Training data
Process
Ensemble Learning Technique
Logistic Regression
Random Forest
K-NN
Prediction model
Evaluation model
Knowledge Base
ROC
Classification Report
Confusion matrix

Ensemble learning technique:

Ensemble learning technique incorporates various algorithms to provide better accuracy**.** Different algorithms are used in the ensemble learning technique are

Logistic Regression:

Logistic regression is a regression analysis to study when the dependent variable is binary in nature. Like all regression analyses, logistic regression is also a predictive analysis. Logistic Regression is used when a particular variable or target is categorical in its type. It uses maximum estimation as a method of approximation.

The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between 0 and 1. The equation of the sigmoid function is,

$$p(x) = \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 x)}}$$

K-Nearest Neighbor:

K-NN is a non-parametric and also a lazy learning algorithm. Its purpose is to use a particular database in which the data will point to several classes to predict the classification of a new sample. When we say one technique is nonparametric, we mean that it does not make any assumptions beforehand on the underlying data. In other words, the structure is determined from the data only.

Euclidean distance:

This is the most commonly used distance measure, and it is limited to real-valued vectors. Using the below formula, it measures a straight line between the query point and the other point being measured.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

Random Forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Decision tree is a supervised learning algorithm which has a predefined target variable that is used in classification problems. It works for concepts like categorical and continuous input and output variables for the model. In this technique, we will split the population of sample into some homogeneous sets based on most significant input variables. In a decision tree, the output is mostly “yes” or “no”.

Entropy:

Entropy is nothing but the uncertainty in our dataset or measure of disorder. The formula for Entropy is,

$$E(S) = -p_{(+)}\log p_{(+)} - p_{(-)}\log p_{(-)}$$

Random decision tree is a type of ensemble learning method for classification. They are used for correction for the habit of overfitting of the training set. The Ensemble Learning methods are used to provide accurate result.

3.3 MODULES AND DESCRPTION

Dataset Preprocessing Module:

Data pre-processing in Machine Learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.

Steps involves,

i. Acquire the dataset
ii. Import all the crucial libraries
iii. Import the dataset
iv. Identifying and handling the missing values
v. Encoding the categorical data
vi. Splitting the dataset
vii. Feature scaling.

Need for pre-processing:

Data pre-processing is the first step marking the initiation of the process. Typically, real-world data is incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends. This is where data pre-processing enters the scenario – it helps to clean, format, and organize the raw data, thereby making it ready-to-go for Machine Learning models. Let’s explore various steps of data pre-processing in machine learning.

Prediction Module:

Prediction Module is used to learn the algorithms and predict the future occurrences of events or things which implements various Machine Learning algorithms such as Logistic regression, Random Forest, K- Nearest Neighbor, Navies Bayes and Ensemble Learning Algorithm with Max voting. Prediction Model is built and develop using the above-mentioned algorithms separately. The developed models then trained by training dataset which is preprocessed. The trained prediction model is tested by testing data, then undergoes various evaluation process such as confusion matrix, recall, precision, F-measure and Receiver Operating Curve -Area Under Curve

Evaluation Models:

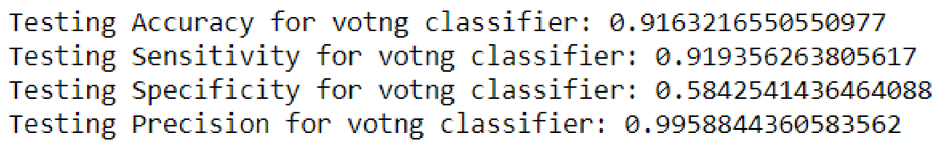
1. Accuracy
2. AUC-ROC
3. Classification Report
   i. Precision
   ii. Recall
   iii. F-measure
4. Confusion matrix

1.Accuracy:

Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

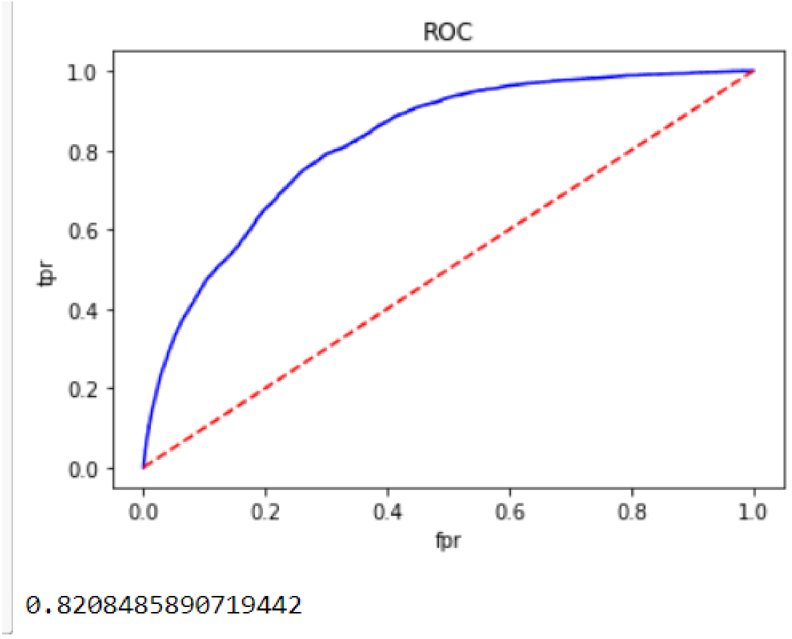
In Logistic Regression prediction model , accuracy of 91.5 % is got. In K-Nearest Neighbor prediction, accuracy of 90.3 % is got. Random Forest gave 91.56% accuracy , Navies Bayes gave 80.2% accuracy and final proposed model the ensemble learning technique gave the 91.63 % of accuracy



```
Testing Accuracy for votng classifier: 0.9163216550550977
Testing Sensitivity for votng classifier: 0.919356263805617
Testing Specificity for votng classifier: 0.5842541436464088
Testing Precision for votng classifier: 0.9958844360583562
```

2.AUC-ROC:

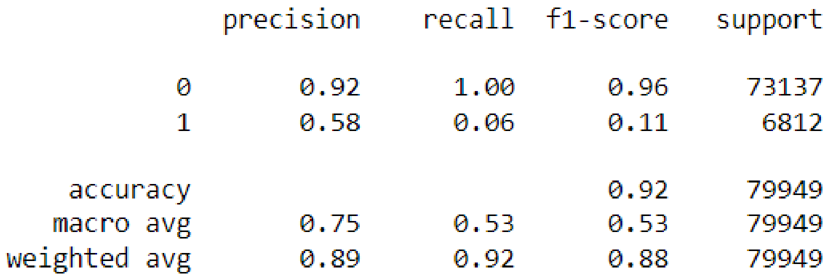
AUC-ROC is the valued metric used for evaluating the performance in classification models. The AUC-ROC metric clearly helps determine and tell us about the capability of a model in distinguishing the classes. The judging criteria being - Higher the AUC, better the model. AUC-ROC curves are frequently used to depict in a graphical way the connection and trade-off between sensitivity and specificity for every possible cut-off for a test being performed or a combination of tests being performed. The area under the ROC curve gives an idea about the benefit of using the test for the underlying question. AUC - ROC curves are also a performance measurement for the classification problems at various threshold settings.



```
0.8208485890719442
```

3.Classification report:

The classification report visualizer displays the precision, recall, F1, and support scores for the model. In order to support easier interpretation and problem detection, the report integrates numerical scores with a color-coded heatmap. All heatmaps are in the range (0.0,1.0) to facilitate easy comparison of classification models across different classification reports. The classification report shows a representation of the main classification metrics on a per-class basis. This gives a deeper intuition of the classifier behaviour over global accuracy which can mask functional weaknesses in one class of a multiclass problem.



```
              precision    recall  f1-score   support

           0       0.92      1.00      0.96     73137
           1       0.58      0.06      0.11      6812

    accuracy                           0.92     79949
   macro avg       0.75      0.53      0.53     79949
weighted avg       0.89      0.92      0.88     79949
```

i. Precision

- Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive.
- Precision = TP/ (TP + FP)

ii. Recall

- Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives.
- Recall = TP/(TP+FN)

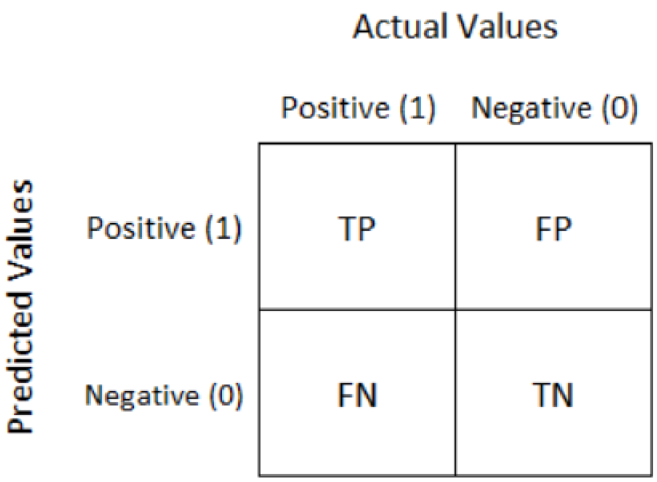
iii. F-measure

- The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

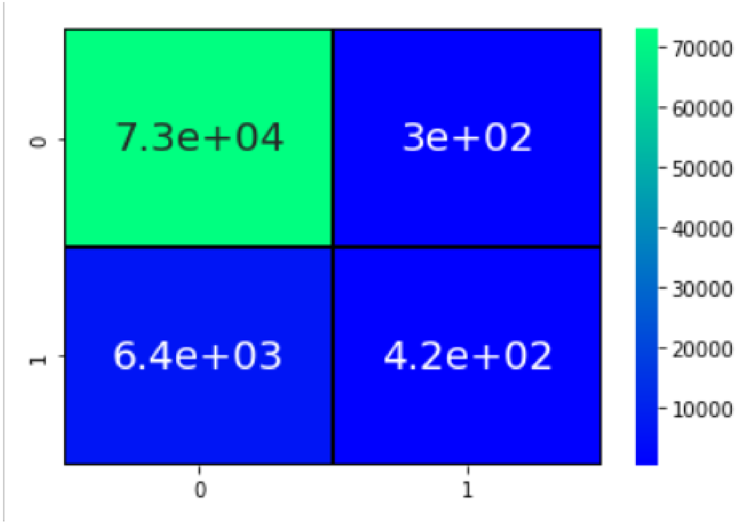
- F1 Score = 2*(Recall * Precision) / (Recall + Precision)

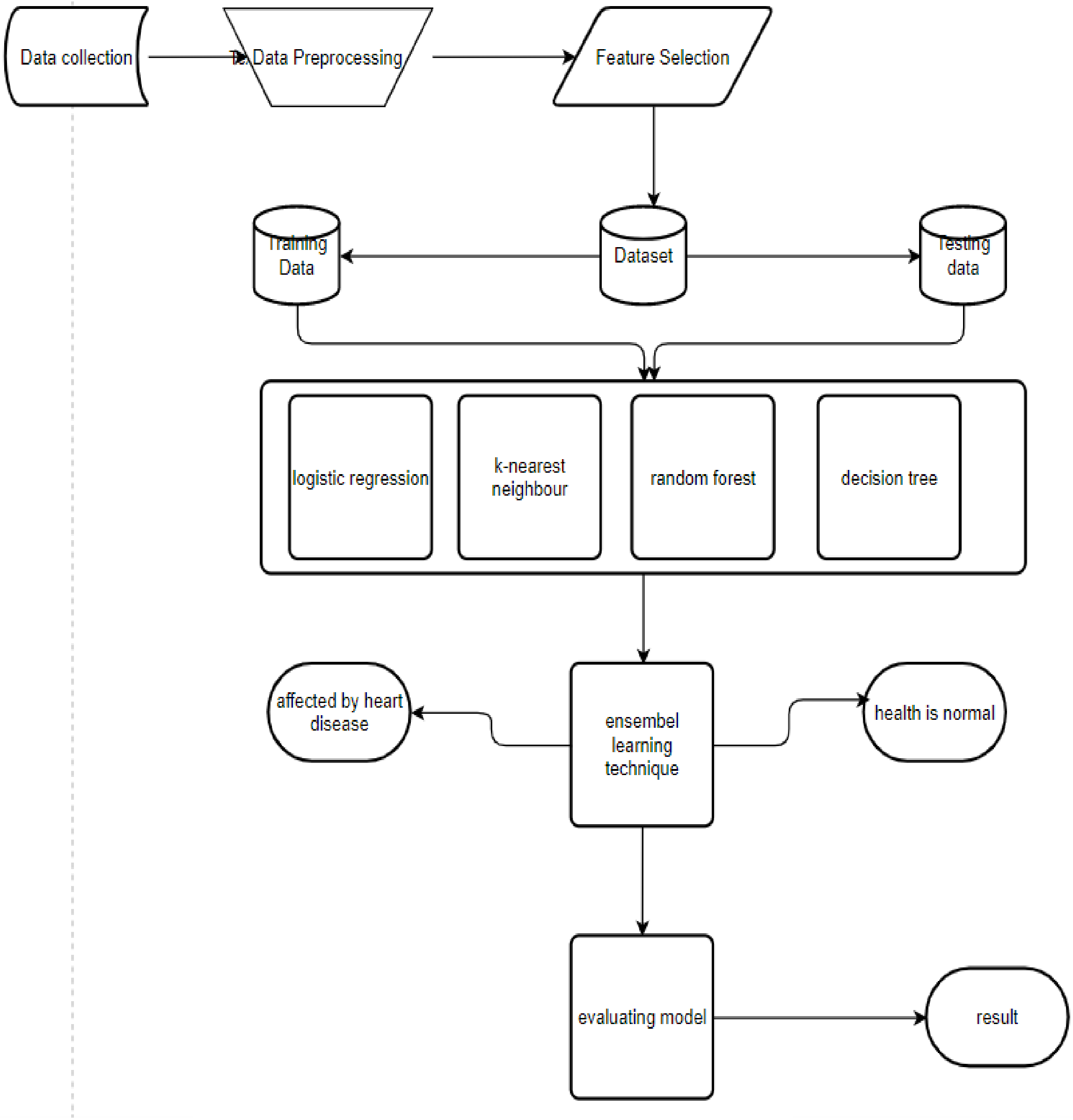
4.Confusion matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN




Data collection
Data Preprocessing
Feature Selection
Training Data
Dataset
Testing data
logistic regression
k-nearest neighbour
random forest
decision tree
affected by heart disease
ensembel learning technique
health is normal
evaluating model
result

CHAPTER 4

RESULTS AND DISCUSSION

The Prediction Model build using Logistic Regression algorithm obtained

```
Testing Accuracy for Logistic Regression: 0.9105
Testing Sensitivity for Logistic Regression: 0.923
Testing Specificity for Logistic Regression: 0.533
Testing Precision for Logistic Regression: 0.9921
and          precision    recall f1-score   support

          0       0.92      0.99      0.96     73137
          1       0.53      0.10      0.16      6812
```

The Prediction Model that implements Navies Bayes algorithm obtained

```
Testing Accuracy for Navies Bayes: 0.8171584385045466
Testing Sensitivity for Navies Bayes: 0.950916179938048
Testing Specificity for Navies Bayes: 0.24083665338645419
Testing Precision for Navies Bayes: 0.8436769350670659
```

The Prediction Model that implements Random Forest algorithm got

```
Testing Accuracy for Random Forest: 0.905
Testing Sensitivity for Random Forest: 0.923
Testing Specificity for Random Forest: 0.360
Testing Precision for Random Forest: 0.977
and           precision    recall f1-score   support

          0       0.92      1.00      0.96     73137
          1       0.61      0.05      0.10      6812
```

The Prediction Model using K-Nearest Neighbour obtained

```
Testing Accuracy for knn 0.9059150208257765
Testing Sensitivity for knn 0.9207977938818701
Testing Specificity for knn 0.32106854838709675
Testing Precision for knn 0.9815825095368965
  and      precision    recall   f1-score   support

          0       0.92      0.98      0.95     73137
          1       0.32      0.09      0.14      6812
```

The Prediction Model that implements Ensemble Learning Technique obtained

```
Testing Accuracy for Ensemble Learning Technique: 0.9163216550550977
Testing Sensitivity for Ensemble Learning Technique: 0.919356263805617
Testing Specificity for Ensemble Learning Technique: 0.5842541436464088
Testing Precision for Ensemble Learning Technique: 0.9958844360583562

and   precision      recall     f1-score   support

           0       0.92      1.00      0.96     73137
           1       0.58      0.06      0.11      6812
```

CONCLUSION AND FUTURE WORK:

According to above discussed result, Ensemble Learning Technique obtained 91.63 % accuracy, recall 100%, precision 92% and f1-score 96 % which show that developed model gives good accuracy and better performance. The future scope is to develop front end and back end as software or web application for the Heart Disease Prediction Model for public use.

REFERENCE:

1. SenthilKumar Mohan, Chandrasekar Thirumalai and Gautham Srivastava proposed, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”, publisher IEEE ,03 July 2019
2. Kriti Gandhi, Mansi Mittal, Neha Gupta, Shafali Dhall “Disease Prediction using Machine Learning”, publisher IJRASET Volume 8 issue VI June 2020.
3. Arnab Das, A. Udith Sai, P. Asha, “Disease Prediction Application Using Machine Learning”, publisher IJRASET, volume 10 issue III March 2022.
4. Harish Rajora, Narinder Singh Punn, Sanjay Kumar Sonbhadra and Sonali Agarwal, “Machine learning equipped web-based disease prediction and recommender system”, publisher IJRASET in 2021.
5. Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath,” Heart Disease prediction using machine learning algorithms”, published in IOP Conference Series: Materials Science and Engineering in 2021.
6. Marouane Fethi Ferjani, “Disease Prediction Using Machine Learning”, publisher ASM in December 2020.
7. Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan and Eman M.Anwar, “Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method”, publisher Wiley Volume 2021 10 pages, China 2021.
8. https://www.kaggle.com/johnsmith88/heart-disease-datase.
9. S.P.Patro et al “Heart disease prediction by using novel optimization algorithm: A supervised learning prospective” Published by Elsevier Ltd. accepted 7 August 2021 and available online 11 August 2021.
10. Pooja Anbuselvan, Student at Bangalore Institute of Technology “Heart Disease Prediction using Machine Learning Techniques” Published by: www.ijert.org Vol. 9 Issue 11, November-2020.