Heart Failure Prediction using PySpark

Pham Thuy Dung[1][20521214]

[1] University of Information Technology – UIT
20521214@gm.uit.edu.vn

Abstract. The heart failure prediction study aims to develop a machine learning model using PySpark to predict the occurrence of heart failure in patients based on relevant medical data and risk factors. The study utilizes PySpark's SQL, PySpark's MLlib library for scalable and distributed machine learning on big data. It involves data preprocessing, including data cleaning, feature selection, and transformation. Various machine learning algorithms, such as Logistic Regression, Support Vector Machines, Random Forest Classifier, Multilayer Perceptron Classifier, and Decision Tree , are employed to train predictive models. The performance of the predictive models is evaluated using metrics like accuracy, F1-score, and Hyperparameter Tuning with Cross-Validation. Based on the experimentation and evaluation, the heart failure prediction study using PySpark concludes that the developed machine learning models show promising predictive capabilities. The chosen algorithms demonstrate varying levels of accuracy and sensitivity in identifying potential heart failure patients.

Keywords: Heart Failure, Machine Learning, Big Data, PySpark, Logistic Regression, Support Vector Machines, Random Forest Classifier, Multiple Perception Classifier, Decision Tree.

1 Introduction

1.1 About Heart Failure

Heart failure is a medical condition characterized by the inability of the heart to pump enough blood to meet the body's needs. It is a serious and chronic condition that affects millions of people worldwide[1]. And it also is one of the abnormal medical conditions caused by cardiovascular disease or heart disease. Heart failure can be chronic or acute and may affect people of all ages. Heart failure can result from various underlying causes, such as coronary artery disease, high blood pressure, or heart valve disorders. Early prediction of heart failure is of great significance for better patient care and outcomes. By identifying individuals at high risk of developing heart failure, healthcare providers can implement proactive interventions and personalized treatment plans. Early detection allows for timely management, lifestyle modifications, and targeted therapies, which can potentially slow down the progression of the disease and improve patient outcomes.

1.2 About Dataset

In recent years, the field of healthcare has witnessed a surge in the availability of large-scale medical data. This dataset [2] was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. Handling and analyzing such big data requires advanced technologies that can efficiently process and extract meaningful insights.

1.3 About PySpark

PySpark, a Python library built on top of Apache Spark, has emerged as a powerful tool for handling big data and conducting large-scale machine learning and predictive analytics. PySpark provides a distributed computing framework that enables parallel processing of data across multiple nodes, making it well-suited for analyzing large healthcare datasets. It offers a wide range of machine learning algorithms and tools through its PySpark's SQL and Pyspark's MLlib library, allowing researchers and data scientists to develop predictive models for various healthcare applications, including heart failure prediction.By leveraging PySpark's capabilities, researchers and data scientists can preprocess and analyze large volumes of medical data, extract relevant features, and train machine learning models to predict the occurrence of heart failure. The use of PySpark in heart failure prediction studies enables scalability, efficiency, and the ability to handle complex and diverse datasets.

In summary, early prediction of heart failure using PySpark offers the potential to improve patient care and outcomes. By harnessing the power of big data and machine learning, PySpark enables researchers to develop accurate and scalable predictive models. This can aid in identifying individuals at high risk of heart failure, facilitating timely interventions, and ultimately improving patient outcomes in the field of cardiovascular health

2 Related Work

Heart failure is a complex medical condition characterized by the heart's inability to pump blood effectively, leading to reduced oxygen supply to the body's tissues. Machine Learning (ML) has emerged as a valuable tool in the healthcare sector, offering various applications in understanding and managing heart failure. ML techniques can be used for early detection and prediction of heart failure, identifying risk factors and patterns associated with its development. By detecting early signs of heart failure, healthcare providers can intervene proactively and implement preventive measures. In [3], the researchers designed a robust system that can predict the possibility of heart failure accurately. The study utilized many algorithms such as SVM, Naive Bayes,

Logistic Regression, Decision Tree, and KNN to predict the possibility of heart failure accurately. In [4] the authors discussed the integration of Apache Kafka with Apache Spark to identify heart disease from patients' social posts. In [5], the paper referenced the heart disease identification from patients' social posts using machine learning solution on Spark. In [6] the researchers compared the performances of various machine learning techniques for risk prediction of cardiovascular disease. It referenced the heart disease identification from patients' social posts using machine learning solution on Spark.

3 Detail Dataset

In this dataset collected from Kaggle[2], it includes 5 heart datasets with combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

- Cleveland dataset: 303 observations
- Hungarian dataset: 294 observations
- Switzerland dataset: 123 observations
- Long Beach VA dataset: 200 observations
- Stalog (Heart) dataset: 270 observations
- The total number of observations in the combined dataset is 1190.

Additionally, there are 272 duplicated observations within the combined dataset, meaning that some data points may have been present in multiple individual datasets before the combination.

This combined dataset with 1190 observations and 272 duplicated observations can be a valuable resource for heart disease research, providing a larger and more diverse dataset for analysis and model development.

Features	Description
Age	age
Sex	Sex(M, F)
ChestPainType	Chest pain type (ASY, NAP, Other)
RestingBP	Resting blood pressure
Cholesterol	Serum cholesrol
FastingBS	Fasting blood sugar
RestingECG	Resting electrocardiogram results (Normal, LVH, Others)
MaxHR	Maximum heart rate achieved
ExerciseAngina	Exercise induced angina
Oldpeak	Oldpeak = ST
ST_Slope	The slope of the peak exercise ST segment

HeartDisease	Target to predict

Table 1: Description of Dataset

First, the data was acquired from the Kaggle repository dataset. It had data on the following indices:

- ✓ Age
- ✓ Sex
- ✓ ChestPaintType
- ✓ RestingBP
- ✓ Cholesterol
- ✓ FastingBS
- ✓ RestingECG
- ✓ MaxHR
- ✓ ExeciseAngina
- ✓ Oldpeak
- ✓ ST_Slope
- ✓ HeartDisease

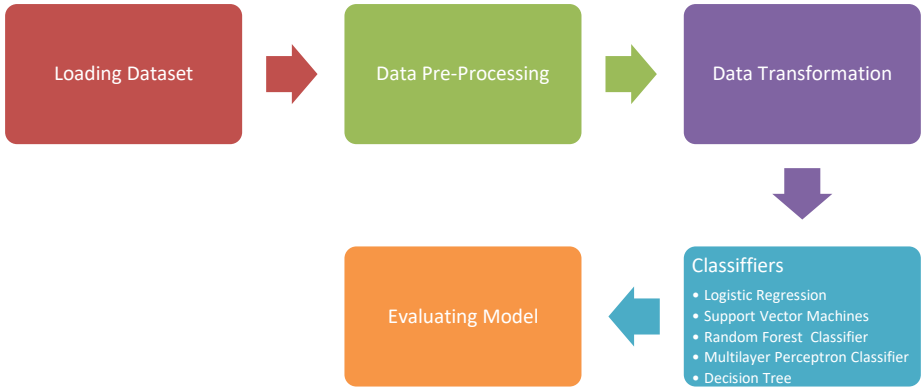

Figure 1: Process Flow Diagram

Process of heart failure prediction has four stages:

- Data collection and Analysis
- Data pre-processing
- Training and Testing the ML Models
- Evaluating the ML Models

4 PySpark Overview

PySpark is a Python library built on top of Apache Spark, which is a unified analytics engine for large-scale data processing. PySpark provides a distributed computing framework that enables parallel processing of data across multiple nodes, making it well-suited for analyzing large healthcare datasets. It offers a wide range of machine learning algorithms and tools through its MLlib library, allowing researchers and data scientists to develop predictive models for various healthcare applications, including heart failure prediction.

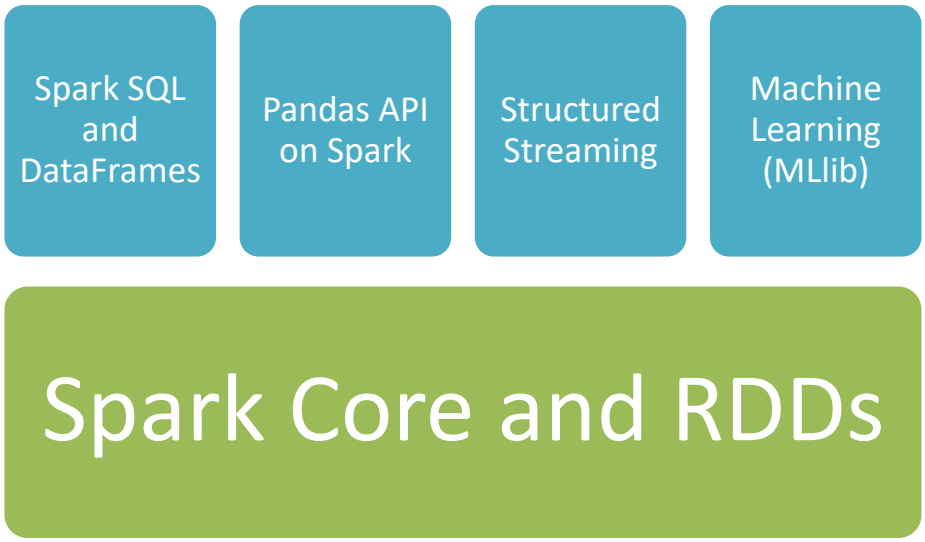

Figure 2: Overview of PySpark

PySpark combines Python's learnability and ease of use with the power of Apache Spark to enable processing and analysis of data at any size for everyone familiar with Python. And it supports all of Spark's features such as Spark SQL, DataFrames, Structured Streaming, Machine Learning (MLlib) and Spark Core.

Part of PySpark cover:

- PySpark SQL - contains commands for data processing and manipulation.
- PySpark MLlib - includes a variety of models, model training and related commands.

5 Machine learning

5.1 Logistic Regression

Logistic regression is a statistical method used to estimate the relationship between one or more independent variables and a binary (dichotomous) outcome variable. It is a type of multivariable analysis used with increasing frequency in the health sciences because of its ability to model dichotomous outcomes[7]. Logistic regression is widely used in healthcare analytics to reduce the number of patient readmissions in hospitals.

Here is logistic function:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Where μ is a location parapmeter (the midpoint of the curve, where $p(\mu) = \frac{1}{2}$ and s is a scale parameter

And here is graph for logistic regression:

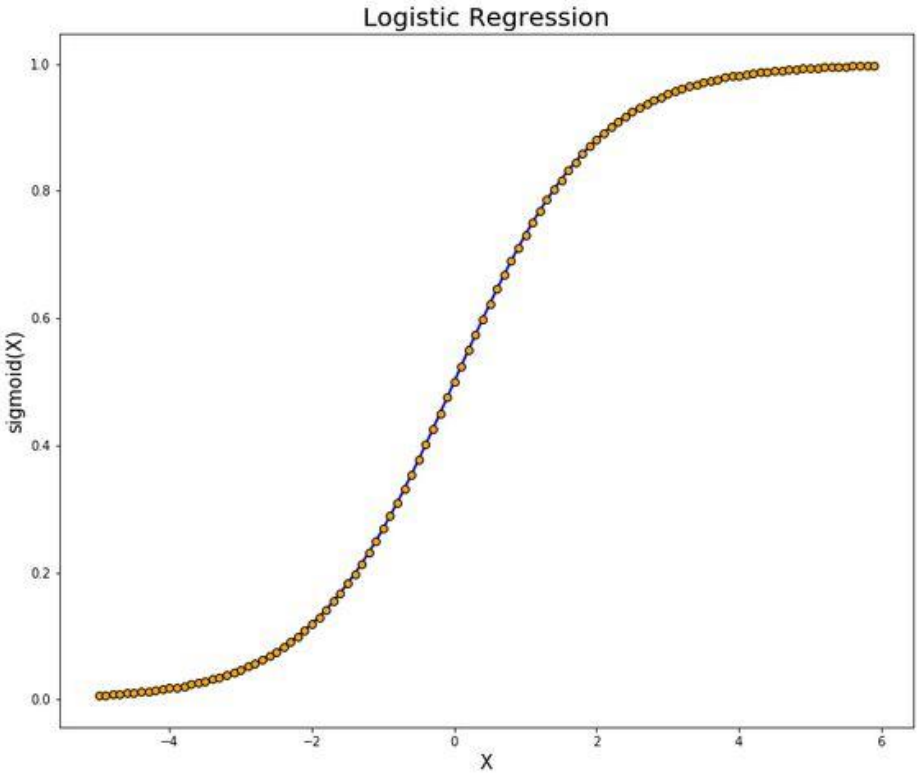

Figure 3: Graph for logistic regression

5.2 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a supervised machine learning algorithm used for both classification and regression tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships

The main aim of the SVM classifier is to find the hyper plane in an n-dimensional space:

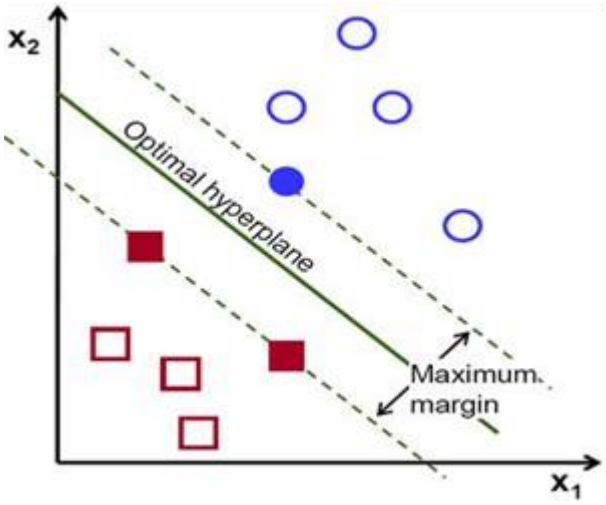

Figure 4: Graph for SVM classifiers

In SVM classifier the main aim to determine the plane with the maximum margin between the two data classes.

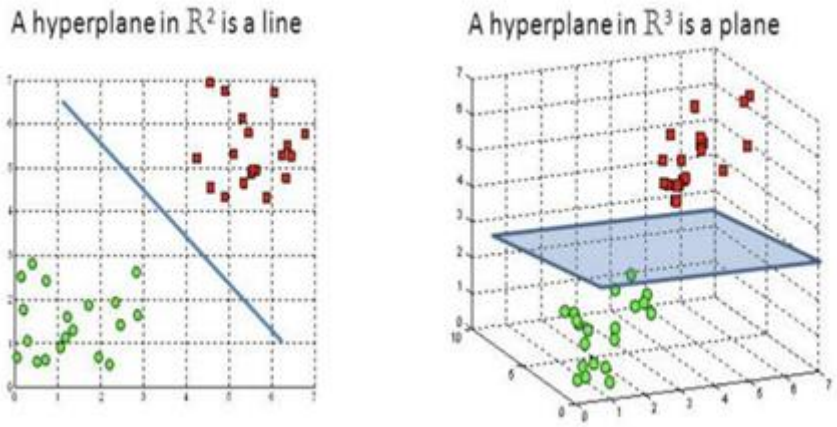

Figure 5: Hyper-lane in 2D and 3D

5.3 Decision Tree

A decision tree is a decision support hierarchical model that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is a versatile supervised machine-learning algorithm that is used for both classification and regression problems. A decision tree is a flowchart-like tree

structure where each internal node denotes the feature, branches denote the rules, and the leaf nodes denote the result of the algorithm.

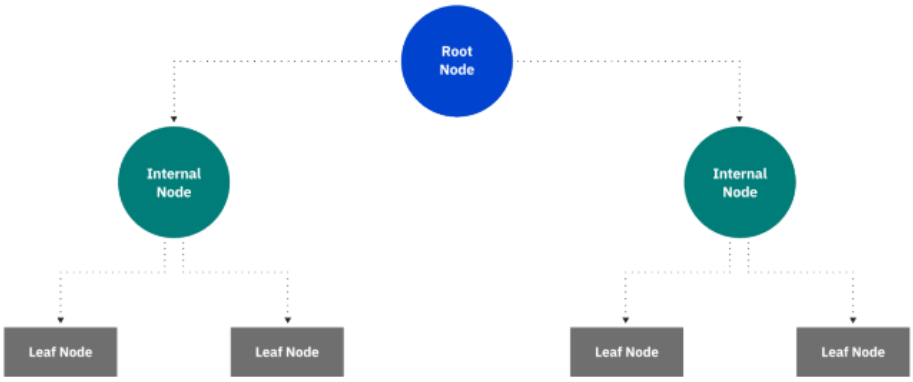

Figure 6: Decision Tree Architecture

5.4 Random Forest

Random forest is a machine learning algorithm that is used for classification, regression, and other tasks. It is an ensemble learning method that constructs a multitude of decision trees at training time.

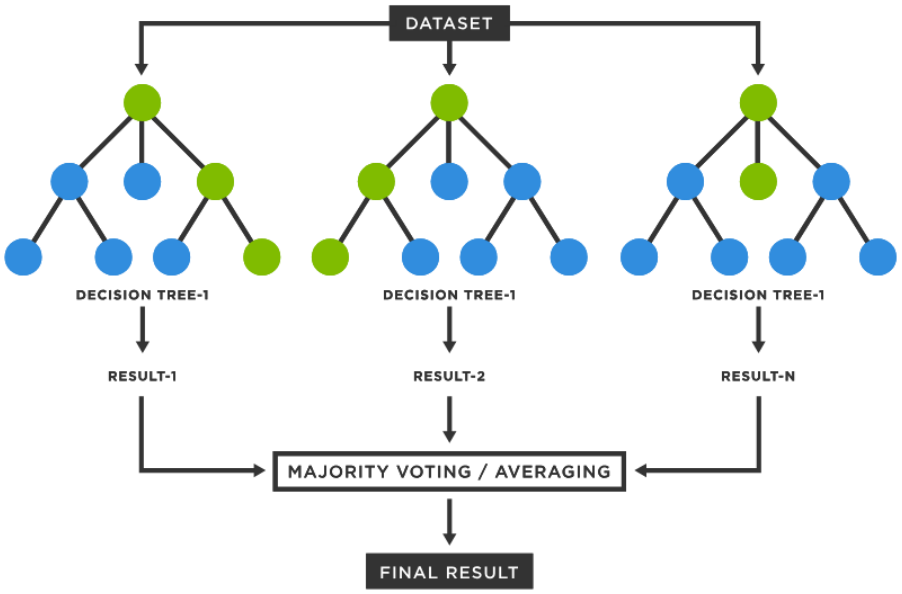

Figure 6: Random forest processing

5.5 Multilayer Perceptron (MLP)

Multilayer Perceptron (MLP) is a type of artificial neural network (ANN) that is used for supervised learning. It consists of three types of layers: input layer, output layer, and hidden layer(s) . MLP is a fully connected class of feedforward ANN, which means that each neuron in one layer is connected to every neuron in the next layer. MLP is a neural network where the mapping between inputs and output is non-linear. MLP uses backpropagation algorithm to train the network. During training, the weights of the connections between neurons are adjusted to minimize the error between the predicted output and the actual output.

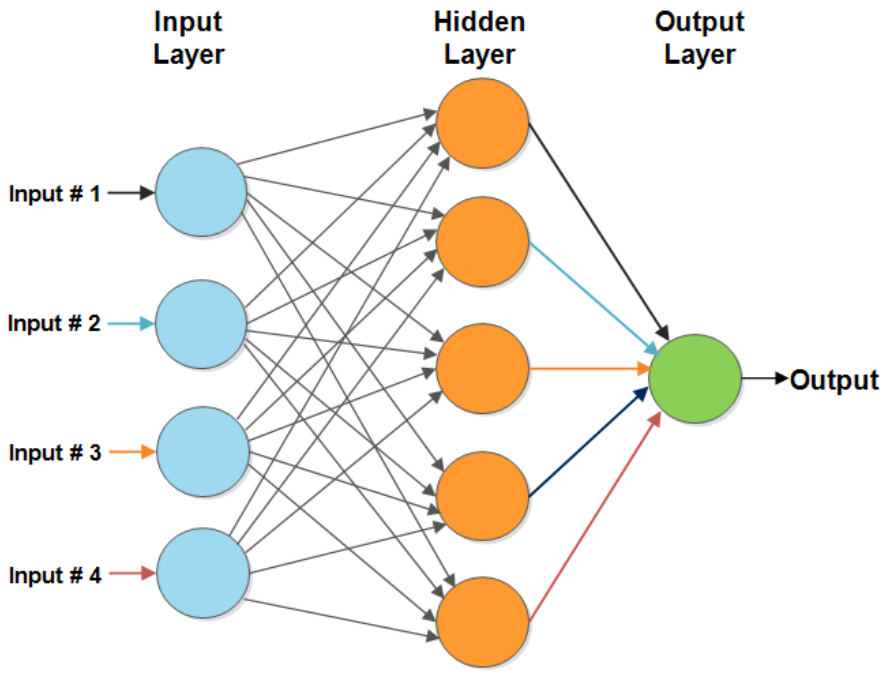

Figure 7: Multi-layer processing

6 Experiemental Setup

6.1 Install Jupyter using docker

The machine utilised to conduct this project has a hard disc drive with a capacity of 20 GB, 4 GB of RAM, and Ubuntu used in Virtual machines .

6.2 Splitting strategy for training and testing datasets

When splitting a dataset for training and testing in heart failure prediction using decision trees, I use an 8:2 split, where 80% of the data is used for training and 20% is used for testing

7 Model Evaluation and Result

7.1 Model Evaluation

Here we are going to fit the data to each of the model and finding the accuracy. Whatever the model is going to give you good accuracy we are going to consider it for our problem statement.

Accuracy = Total Number of Correct Predictions / Total Number of Observations

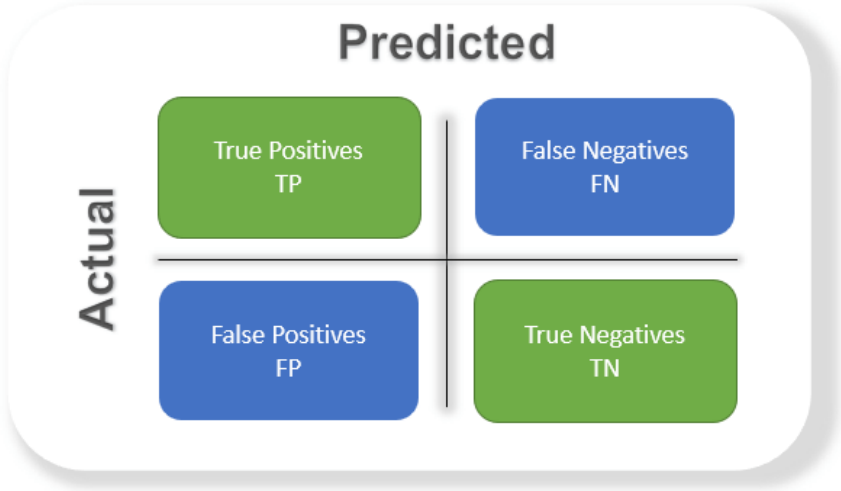

Figure: Confusion Matrix

7.2 Result of Accuracy

Model	Train Accuracy Score	Average Accuracy Score
Logistic Regression Model	84.45	91.35
Support Vector Machines Model	84.46	85.19
Random Forest Classifier Model	83.78	85.14
Multilayer Perceptron Classifier Model	81.08	79.48
Decision Tree Model	81.08	81.16

8 Conclusion

We have seen accuracy on various machine learning algorithms and got the accuracies accordingly. It is clear that Logistics have given the best accuracy with 84,45% in train accuracy score and 91.35% in average accuracy score, here ML plays a key role to analyze the heart disease. The results demonstrate the significance of using machine learning techniques to analyze and predict heart disease. By leveraging the power of data and pattern recognition, machine learning models like Logistic Regression can assist in identifying potential heart disease cases, enabling early detection and intervention, which can be crucial for patient outcomes.

References

1. Ahmed, H., Younis, E. M. G., Hendawi, A., & Ali, A. A., Heart disease identification from patients' social posts, machine learning solution on Spark, October 2019, Future Generation Computer Systems 111
2. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
3. Sahoo, Prasanta Kumar Jeripothula, Pravalika, Heart Failure Prediction Using Machine Learning Techniques, Sreenidhi Institute of Science and Technology (SNIST), January 29, 2021
4. Hager Ahmed , Eman M.G. Younis , Abdeltawab Hendawi , Abdelmgeid A. Ali, Heart disease identification from patients' social posts, machine learning solution on Spark, Future Generation Computer Systems (IF 7.5) Pub Date: 2019-10-05
5. Ibrahim M. El-Hasnony, Omar M. Elzeki, Ali Alshehri, and Hanaa Salem, Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction, Sensors (Basel). 2022 Feb; 22(3): 1184,Published online 2022 Feb 4.
6. Madhumita Pal, Smita Parija, Ganapati Panda , Kuldeep Dhama and Ranjan K. Mohapatra, Risk prediction of cardiovascular disease using machine learning classifiers, Journal: Open Medicine, Publication Date: June 17, 2022
7. Steven C Bagley, Halbert White, Beatrice A Golomb, Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain, Volume 54, Issue 10, October 2001, Pages 979-985