# *Prediction of Heart Disease Using Machine Learning in Apache Spark Approach*

1st Md. Solaimanur Rahman
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
solaimanur15-2951@diu.edu.bd
ID: 193-15-2951

2nd Md. Al Amin Miah
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
alamin15-2965@diu.edu.bd
ID: 193-15-2965

3rd Mahabub Hossan Emon
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
mahabub15-2950@diu.edu.bd
ID: 193-15-2950

4th Md. Abu Hana Mostafa Zaman Tushar
Dept. of Computer Science & Engineering
Daffodil International University
Dhaka, Bangladesh
abu15-3005@diu.edu.bd
ID: 193-15-3005

## Project Proposal :

Our project is already done that's why we write the full project……

## Type Of Project

We decided to work on a research base project, We choose a project titled which is about heart disease because It is one of the most common problems in the world so we decided to predict the problem as soon as possible so that we can save people's life. In the past, many authors work on that topic. In their paper, the highest accuracy which was 91 % was good enough but we make that 93%. For heart disease prediction if the accuracy is better then we can predict the problem better than the low accuracy. We use the data set from online because it is a medical type of data and in our country, we can't get the data from medical for the illegal issue. We used some algorithms for our project which are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, and Random Forest in apache spark. We choose those algorithms because by using those algorithms in apache spark we will get better results.

Here is our Full Project,,,,

***Abstract— One of the most frequent diseases is heart disease. This disease is quite widespread these days, so we used many factors that are related to cardiac disorders to come up with a better method of prediction, as well as algorithms. This paper uses the data set from online and has 13 important attributes. This work is implemented using some classifiers such as Support Vector Machine(SVM), K nearest neighbor(KNN), Logistic regression& Random Forest in apache spark. It is found that Random Forest gave the best result with an accuracy of up to 93%. A comparative statement of all the algorithms is also presented in the implementation part of the paper. This research also uses the model validation technique to design the best suitable model fitting in the current scenario.***

***Keywords—component; formatting; style; styling; insert (Heart Disease Prediction; machine Learning; SVM; KNN; Naïve Bayes; Random Forest; Apache Spark)***

## I. Introduction

In this modern era, people are very busy and working hard to satisfy their materialistic needs and are not able to spend time for themselves, which leads to physical stress and mental disorder. There are also reports that the heart suffers because of the global pandemic coronavirus. Inflammation of the heart muscle can be caused by Covid-19. Thus, heart disease is very common now a day's particularly in urban areas because of excess mental stress due to Covid-19. As a result, heart disease has become one of the most important factors in the death of men and women in the so-called material world. It has emerged as the top killer that has affected both the urban and rural populations. CAD (coronary artery disease) is one of the most common types of heart disease. In the medical field predicting heart disease has become a very complicated and challenging task, requiring a patient's previous health records and in some cases, they even need Genetic information as well. So, in this contemporary lifestyle, there is an urgent need for a system that will predict accurately the possibility of getting heart disease. Predicting a heart disease at an early stage will save many people's life. There were many heart disease prediction systems available at present, the Authors have researched well and proposed different Classification and prediction algorithms but each one has its limitations. The main objective of this paper is to overcome the limitations and to design a robust system that works efficiently and will be able to predict the possibility of heart failure accurately.

## II. LITERATURE REVIEW

There are several papers, studies, and research articles on cerning life expectancy that have already been completed by a variety of authors. Here are some work reviews are provided below to correspond with our work

Dilip et al. [1] use data analytics to detect and predict disease patients. First of all, they used the most relevant features by the correlation matrix then apply three data analytics techniques to the datasets. his accuracy is 90%.

Abhay et al. [2] use a multi-layer perceptron (MLP)-based approach to predict 30-day HF readmission or death. The MLPbased approach produced the highest AUC (0.62) and AUPRC (0.46) with 48% sensitivity and 70% specificity.

Prasanta et al. [3] Prediction of Heart Disease Using Machine Learning Algorithms. For classification, they use the ID3 algorithm, and to get accuracy they use Logistic regressionand K means.

Santhi et al. [4] used machine learning algorithms called Backpropagation Algorithms and later they used optimization algorithms. They have found a decent result based on 14 attributes of data for prediction.

Saqib et al. [5] presented a comparison of used machine learning algorithms over different evaluation metrics. They used different types of classifiers. Such as kNN, Artificial neural network, Decision Tree, Random Forest, Support vector, etc. All of them have an accuracy between 80%-85%.

Oladosu et al. [6] annual deaths of approximately 17 million people worldwide. They collect a data set from a hospital there are 105 women and 194 men with the age bracket of 40- 95 years. After the data prepossessing, they choose Random forest, Logistic regression, SVM, and K-nearest neighbor algorithms. There is different accuracy in different algorithms. In random forests the accuracy is 0.7947%, In Naive Bayes, the Accuracy is 07456% the last one is KNN which was 0.7504%.it can be said that serum creating, ejection fraction, smoking status, and age are the major determinants in predicting heart failure survival.

Jamil et al. [7] used the NA¨IVE BAYES Algorithm, Logistic Regression Algorithm, and Neural Network Algorithm this algorithm was used to evaluate the proposed system performance. So, their proposed method produced an accuracy of 91.26% which is more than the previous accuracy of 87%. LSTM methods.

Sulabha et al. [8] showed a Neural Network algorithm to make this Heart Disease Prediction system. Confusion matrices are included here. In the end, the experimental result shows that by using neural networks the system predicts Heart disease with nearly 91% accuracy.

Fatma et al. [9] designed and developed by using MATLAB's GUI feature with the implementation of a Backpropagation Neural Network. The Backpropagation Neural Network used in this study is a multi-layered Feed Forward Neural Network, which is trained by a supervised Delta Learning Rule. The dataset used in this study is the signs, symptoms, and results of the physical evaluation of a patient. The proposed system achieved an accuracy of 90%.

Abinaya et al. [10] used the data set consisting of various clinical attributes like patient Age, Sex, Chest pain, Fbs and etc. They divide the data set into two sets, the Training set is 70% and the testing set is 30%. This research work is carried out by implementing the data set over five different algorithms(SVM confusion matrix, Logistic regression, Logistic Regression, Decision Tree, KNN) and results are compared. By using the Support vector machine this model could be able to predict with an accuracy of about 85.2% which is the highest as compared to other algorithms. Thomas et al. [11] authors are using some algorithms like Decision tree, K-Nearest Neighbor, Support Vector Machines and, Random Forest of which Decision Tree gives an accuracy of 79%, K-Nearest Neighbor gives an accuracy of 87%, Support vector Machines give an accuracy of 83%, and Random Forest Gives an accuracy of 84%. Virender,

Rohila et al. [12] were using Machine Learning and Data Analytics Approach. By using different types of data mining and machine learning techniques to predict the occurrence of heart disease have summarized. In this paper they propose many techniques such as an Efficient Heart Disease Prediction System using a Decision Tree, Prediction and Diagnosis of Heart Disease by Data Mining Techniques, Heart Disease Diagnosis using an Artificial Neural Network, and Prediction of Heart Disease using the WEKA tool. And it provides 86.3% accuracy in the testing phase and 87.3% in the training phase.

Shobana et al. [13] suggested a prediction of high-risk heart disease using a Logistic regression algorithm. Also, they use some techniques such as neural networks, KNN algorithms, etc. They found that with age and gender the accuracy level of the prediction was 40.3% and when two additional attributes such as smoking and previous heart disease were added the accuracy level of the prediction increased to 80.6%.

## III. RESEARCH QUESTION

Here I will talk about my Research Question.

1: Why does machine learning predict heart disease?

2: How is heart disease predicted?

3: How can we get a better result?

## IV. RESEARCH OBJECTIVE

From this section, we will talk about why we do the project

1. To get a more accurate method of prediction.

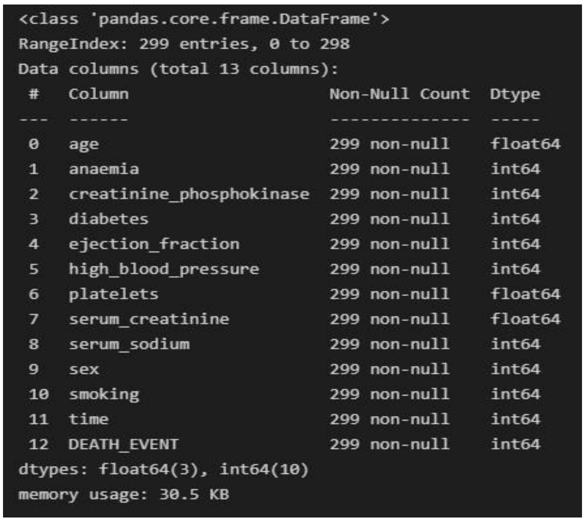
2. To compare with SVM, KNN, Logistic regression, and Random Forest in Apache Spark to find the best result.

3. To overcome the limitations.

4. To design a robust system that works efficiently and will be able to predict the possibility of heart failure accurately.

## V. DATASET

We analyzed a dataset containing the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All 299 patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of the New York Heart Association (NYHA) classification of the stages of heart failure. The dataset contains 13 features, which report clinical, body, and lifestyle information, that we briefly describe here. Some features are binary: anemia, high blood pressure, diabetes, sex, and smoking. The hospital physician considered a patient having anemia if hematocrit levels were lower than 36%.



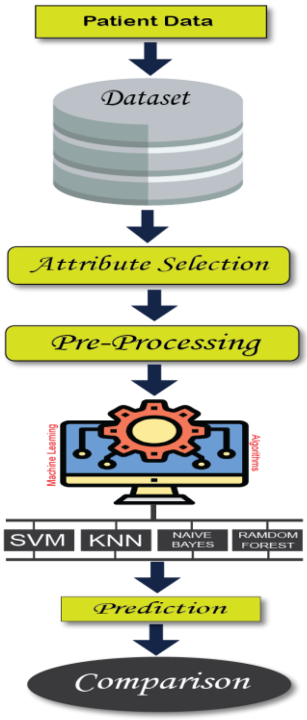
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   diabetes                  299 non-null    int64
 4   ejection_fraction         299 non-null    int64
 5   high_blood_pressure       299 non-null    int64
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64
 9   sex                       299 non-null    int64
 10  smoking                   299 non-null    int64
 11  time                      299 non-null    int64
 12  DEATH_EVENT               299 non-null    int64
dtypes: float64(3), int64(10)
memory usage: 30.5 KB
```

*Figure: Dataset Information*

## VI. METHODOLOGY

After we explore the dataset, we selected all of the attributes because there are no null values. Then we process the dataset through the heatmap. After that, we did some graphical visualization and we applied machine learning algorithms. We used SVM (Support Vector Machine), KNN, Naïve Bayes & random forest. We have found different kinds of accuracy by using those classifiers. In the end, we have done a comparison between them.

- Processing the Dataset.
- Applying machine learning Algorithms.
- Prediction.
- Comparison.



*VI: ANALYSIS*

### A. Heat Map:

The time of the patient's follow-up visit for the disease is crucial as initial diagnosis with cardiovascular issues and treatment reduces the chances of any fatality. It holds an inverse relation. The ejection fraction is the second most important feature. It is quite expected as it is basically the efficiency of the heart. The age of the patient is the third most correlated feature. Clearly as the heart's functioning declines with aging

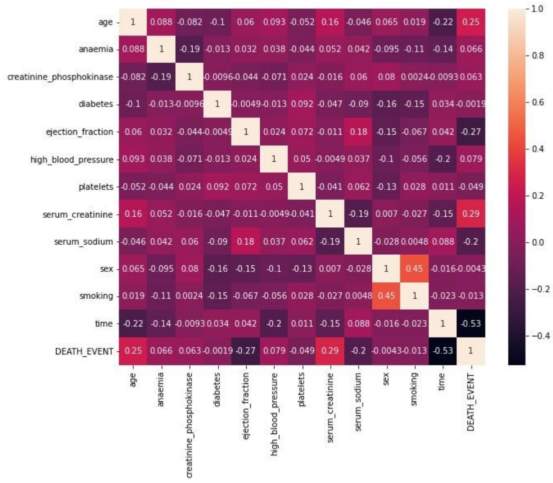
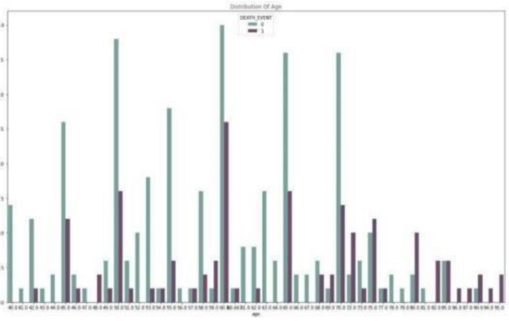


Fig: Heat map

### B. *Death Event:*

Using the attribute Age, we have graphed death events from this dataset. We can observe that people around the age of 60 have a higher chance of dying.



### C. *Machine Learning Algorithms:*

**Logistic regression–** It is a simple learning algorithm that uses Bayes' rule in combination with a strong assumption that the features are conditionally independent, given classes. Although this independence assumption is often violated in practice, Logistic regressionoften provides competitive classification accuracy.The naïve bayes formula is,

$$P(A|B) = P(B|A) * P(A) / P(B)$$

*Figure: Naïve Bayes*

**KNN -** K nearest neighbor algorithms fall under the category of supervised learning and are used for classification (at most) and regression. It is a versatile algorithm that is also used for missing values and resampling datasets. The KNN formula is,

$$dist(x,z)=(d\sum r=1|xr-zr|p)1/p.$$

*Figure: KNN*

**SVM (Support Vector Machine)** - A support vector machine, or SVM, is a linear model that can be used to solve classification and regression problems. It can solve both linear and nonlinear problems and is useful for a wide range of applications. SVM is a basic concept: the method divides the data into classes by drawing a line or hyperplane. Support Vector Machine (SVM) is a type of linear classifier that works on the concept of margin maximization. They use structured risk minimization to increase the complexity of the classification to achieve outstanding generalization performance.

*Figure: SVM*

**Random Forest -** As a result, in a random forest, the technique for splitting a node only considers a random subset of the features. Instead of searching for the greatest possible thresholds, you may make trees even more random by employing random thresholds for each feature (like a normal decision tree does). Random forest classification uses an ensemble methodology to achieve the desired result. Various decision trees are trained using the training data. This dataset contains observations and features that will be chosen at random when nodes are split. Various decision trees are used in a rainforest system. The Random Forest formula is,

$$MSE = \frac{1}{N} \sum_{i=1}^{N}(fi - yi)^2$$

*Figure: Random Forest Classification*

*Figure: RF Classification Confusion Matrix*

## VII. RESULT & COMPARISON

After Evaluate the Random Forest Classification Confusion Matrix we found TN = 43, FN = 4, FP = 0, TP = 13. The accuracy of SVM is 88%. We have got 77% accuracy for Naïve Bayes Classification. We have 80% accuracy in KNN But When we used Random Forest Classification, we get the highest accuracy of 93%. So, in this scenario, Random Forest Classification gives the best accuracy.

Comparison

## VIII. CONCLUSION & FUTURE WORK

The major goal of this research is to provide insight into applying machine learning techniques to detect heart disease risk rates. Many papers describe various data mining approaches and classifiers that are utilized for efficient and effective heart disease diagnosis. According to the analysis mode, many authors utilize a variety of technologies and a different number of attributes in their research. As a result, depending on a variety of factors, different methods provide varying degrees of precision. The prediction of heart disease was recognized using SVM, Logistic regression, and Random Forest Classifier, and the accuracy level was also reported for a variety of parameters. In our paper, there are some limitations for the dataset we took the dataset from online, and there are maximum true values and some are false they are not equal in number. Because of that equivalent, we can't get a result more than that. If we get an equivalent value then the accuracy will increase more. Other algorithms could be used in the future to reduce the number of attributes while increasing accuracy. There are other changes that might be investigated to improve the predictability and scalability of this system. Due to a lack of time, the following research/work will have to be completed in the future. Testing different discretization strategies, numerous classifiers voting techniques, and other decision tree kinds, such as information gain and gain ratio, is something I'd like to do. Willing to investigate various rules, including association rules and grouping methods.

## REFERENCES

[1] Dhai Eddine Salhi, Abdelkamel Tari & M-Tahar Kechadi Using Machine Learning for Heart Disease Prediction. "

[2] Mohammed Bennamoun and Saqib Ejaz: " Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics " ESC Heart Failure (2019).

[3] Rajesh N, T Maneesha, Shaik Hafeez and Hari Krishna: " Prediction of Heart Disease Using Machine Learning Algorithms " International Journal of Engineering & Technology, 7 (2.32) (2018) 363-366.

[4] V. Srinivas1, K.Aditya, G. Prasanth, and R.G. Babukarthik: "A Novel Approach for Prediction of Heart Disease: Machine Learning Techniques" International Journal of Engineering Technology, 7 (2.32) (2018) 108-110

[5] Virendar Ranga and D. Rohila: " Parametric Analysis of Heart Attack Prediction Using Machine Learning Techniques " International Journal of Grid and Distributed Computing Vol. 11, No. 4 (2018), pp.37- 48.

[6] Oladosu Oyebisi Oladimeji, Olayanju Oladimeji: " Predicting Survival of Heart Failure Patients Using Classification Algorithms " JITCE (Journal of Information Technology and Computer Engineering) SSN (Online) 2599-1663.

[7] Md. Jamil-Ur Rahman, Rafi Ibn Sultan, Firoz Mahmud: " EN SEMBLE OF MULTIPLE MODELS FOR ROBUST INTELLIGENT HEART DISEASE PREDICTION SYSTEM "

[8] ] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte: " A DATA MINING APPROACH FOR PREDICTION OF HEART DISEASE USING NEURAL NETWORKS" International Journal of Computer Engineering and Technology (IJCET), ISSN 0976 – 6367(Print), ISSN 0976 – 6375(Online) Volume 3, Issue 3, October-December

[9] Fatma Zahra Abdeldjouad, Menaouer Brahami, Nada Matta: " A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques" M. Jmaiel et al. (Eds.): ICOST 2020, LNCS 12157, pp. 299–306, 2020.

[10] Prasanta Kumar Sahoo, Pravalika Jeripothula: " HEART FAILURE PREDICTION USING MACHINE LEARNING TECHNIQUES."

[11] Keshav Srivastava, Dilip Kumar Choubey: " Heart Disease Prediction using Machine Learning and Data Mining " International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume 9 Issue-1, May 2020.

[12] M. Marimuthu, M. Abinaya ,K. S. Hariesh, K. Madhankumar, V. Pavithra: " A Review on Heart Disease Prediction using Machine Learn ing and Data Analytics Approach " International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 18, September 2018.

[13] Theresa Princy. R and J. Thomas: " Human Heart Disease Prediction System using Data Mining Techniques." 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT