e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 5, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.54

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506083 |

Prediction of Cardiovascular Disease Using Machine Learning Library (MILB)

MOSES K, PAVAN KUMAR M, SAKTHI KUMARAN S, Dr.V.P GLADIS PUSHPARATHI

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

U.G Scholar, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

Professor, Department of CSE, Velammal Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT: In human life, healthcare is an unavoidable and important task to be done. Cardiovascular Diseases are a group of diseases that affects heart and blood vessels. The earlier methods of estimating the uncertainty levels of cardiovascular diseases helped in taking decisions to reduce the risk in high-risk patients. This project proposes a prediction model to predict whether a person has a heart disease or not and to provide awareness or diagnosis on the risk to the patient. Our goal is to enhance the performance of the model by adding significant attributes to this model which will increase the efficiency of this application and useful for the classification task. the main focus of the system is to make use data analytics to predict the presence of the disease and level of disease among patients, we have used over 10 Million data of various people with and without heart disease from year 2010 to 2022 to built an efficient model to predict the heart disease and we have used different algorithms to facilitate this process namely Support Vector Machine, Random Forest, Ada Boost and Gradient Boosting within these model we choose more precise one to deliver accurate data with an accuracy of max 76% and average of 74%.

I. INTRODUCTION

Healthcare means the maintenance or advancement of health through the prevention and diagnosis of people. Nowadays, healthcare is increasing day by day due to lifestyle and hereditary. Cardiovascular disease has become the deadliest enemy. A person with cardiovascular disease cannot be cured simply. So, diagnosing patients at the correct time is the toughest work in the medical industry and needs to be diagnosed at initial stages to reduce the risk on the patient in the future. Every human body possesses different numbers for blood pressure, cholesterol, and pulse rate. But the normal values would be, blood pressure is 120/80, cholesterol is 200 mg/dl and pulse rate is 72. So combining these machine learning algorithms with medical data sources is useful. We have used over 10 Million data of various people with and without heart disease from year 2010 to 2022 to built an efficient model to predict the heart disease and we have used different algorithms to facilitate this process namely Support Vector Machine, Random Forest, Ada Boost and Gradient Boosting within these model we choose more precise one to deliver accurate data with an accuracy of max 76% and average of 74%.

II. LITERATURE SURVEY

In 2010, Khosla et al. [11] have implemented the Cox proportional hazards model using data mining methods for the prediction of stroke diseases. They performed implementation using the CHS (Cardiovascular Health Study) dataset. They concluded that the support vector machine classifier attained the highest evaluation using the ROC curve.
In 2015, Dewan, Ankita Sharma, Meghna et al. proposed a hybrid technique with the ability to solve complex skepticism which is indispensable for the prognosis of cardiac disease which may aid doctors to diagnose the condition. The proposed hybrid technique is based on a dataset with 13 attributes that were taken from the UCI repository. The evaluation matrix used is Accuracy and sensitivity.

In 2019, Shamsollahi, M Badiee, A Ghazanfari et al. built a model using an amalgam of descriptive and predictive analytics of KDD (Knowledge Discovery in Databases). The authors are determined the number of clusters using clustering indices. After that, the authors have applied some decision tree methods and artificial neural networks for all clusters. They used the original dataset for building a model that is gathered from a heart clinic database. The results have shown that the CART decision tree model attained the best of all methods.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506083 |

$$Performance\ of\ the\ stump\ =\ \frac{1}{2}\log_e(\frac{1-Total\ Error}{Total\ Error})$$

$$\alpha\ =\ \frac{1}{2}\log_e\left(\frac{1-\frac{1}{5}}{\frac{1}{5}}\right)$$

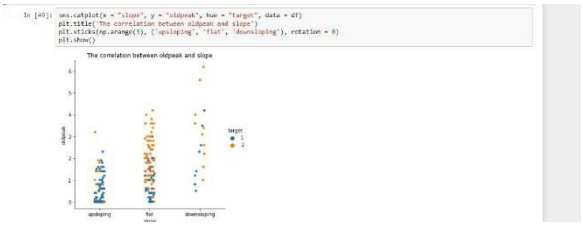
$$\alpha\ =\ \frac{1}{2}\log_e\left(\frac{0.8}{0.2}\right)$$

$$\alpha\ =\ \frac{1}{2}\log_e(4)\ =\ \frac{1}{2}*(1.38)$$

$$\alpha\ =\ 0.69$$

In 2019, Makumba, Dominic Obwogi Cheruiyot, Wilson Ogada, Kennedy et al. developed a model using data mining on heart disease prognosis that can support the making a decision. There are used such decision trees, naive bayes, knearest neighbors, and also Waikato Environment for Knowledge Analysis application programming interface. Dataset for the proposed model has been accessed from UCI with 13 attributes. The confusion matrix has been used for the evaluation of the models. The researchers concluded that the proposed framework provides a better prediction to prognostic of the coronary artery disease from the present-day blueprint.
In 2020, Ahmed et al. [14] created a system using the Spark ML library which is accompanied by the big data platform Apache Spark. To build the prediction model for diabetes patients, they have used classification methods that included are support vector machine algorithm, decision tree algorithm, logistic regression algorithm, naive bayes algorithm, and random forest algorithm. Following that, evaluated all models using some matrices such as accuracy, recall, and precision then they found that the logistic regression model accomplished the best percentage score of Accuracy (82%), Recall (92%), and Precision (82%).

In 2020, Kaur, H. et al. [13] developed a predictive model that uses different machine learning algorithms which are KNN, Linear SVM algorithm, RBF-kernel SVM algorithm, and ANN algorithm. Pima Indian diabetes dataset has been trained and validated for predicting diabetic and nondiabetic patients. The results have shown the Linear SVM model accomplished a good accuracy with 89% out of all models.



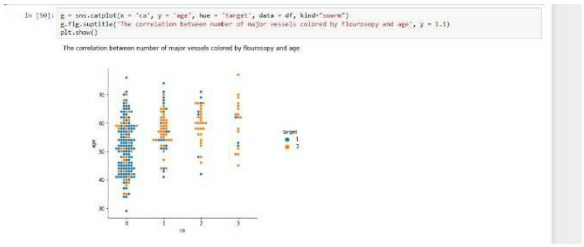
In 2020, Muktevi Srivenkatesh [12] developed a prediction model for cardiovascular disease and the dataset is taken from Kaggle.com for creating a proposed model. There are used various classification models that are consist of logistic regression classifier, support vector machine classifier, random forest classifier, and naive bayes classifier. They have using metrics accuracy, sensitivity, and specificity to attain the evaluation of the diagnosis model. After evaluating and comparing the models, the authors obtained a better accuracy with 77.06 % of the logistic regression model.

In 2020, Komal Kumar et. al [7] developed a proposed system for heart disease. They have used a heart disease dataset that was gathered from the UC Irvine(UCI) Repository, which had 10 attributes. The classification techniques random forest, decision tree, logistic regression, knearest neighbors, and support vector machine had used for building a prediction model of cardiovascular ailment. After comparing the performance of models, the researchers have shown that the random forest machine learning model obtained the best accuracy with 85.71% and Area Under ROC of 0.8675. Therefore, they used a random forest machine learning model that aids in building a system to classify the patients who are affected with Cardiovascular sickness.

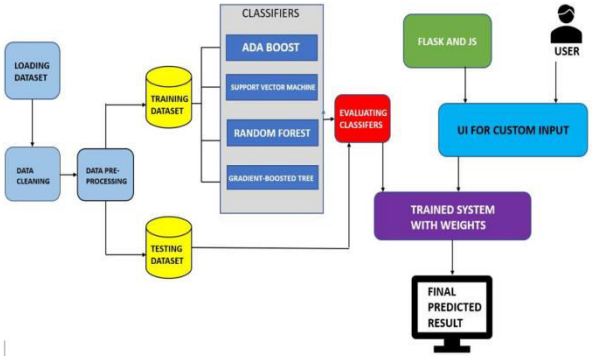
| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506083 |



In 2022 Surbhi Kumari Student of Department of Computer Science, in Banasthali Vidyapith had developed heart disease prediction application Performance Evaluation of Distributed Machine Learning for Cardiovascular Disease Prediction in Spark This research work aims to build a prediction model to predict whether individuals have cardiovascular disease or not, using machine learning classification techniques which include logistic regression, decision tree, support vector machine, random forest, and gradient-boosting tree classifier and also applied hyperparameter tuning and cross-validation with 5-fold to improve the performance of models. They compared the evaluation of all applied machine learning models and the results observed that the Gradient-Boosting Tree Classifier achieved better Accuracy (73.20%) and Area Under ROC value (0.8002)



III. PROPOSED SYSTEM

In previous studies, they have discussed predicting the significant features of heart disease prediction by using different machine learning. We proposed machine learning Library (MLlib) techniques such as Support Vector Machine, Random Forest, AdaBoost and Gradient Boosting for heart disease prediction of significant features. ML process starts from a pre-processing data phase followed by feature selection based on data cleaning, classification of modeling, performance evaluation, and the results with improved accuracy. We have splitted

training data and testing data with 80% and 20% in that 10 Million data to have a better test result and increase the accuracy of this application, unlike the previous model we have also included much more additional parameters to facilitate the prediction and classification process the following attributes are used to predtict the possibility of heart attack for a person age, sex, blood pressure, cholestrol, fasting blood sugar, ECG, Thalessemia, Chest pain type, heart rate, ST depression, excercise induced angina, Slope of the Peak Exercise ST Segment and Number of Vessels Colored by Flourosopy. These attributes complex the application and also increases the precision thus we can get much more accurate results on the possibility of heart disease for an individual.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506083 |

SVC

classification_report :

	precision	recall	f1-score	support
1	0.71	0.90	0.79	30
2	0.81	0.54	0.65	24
accuracy			0.74	54
macro avg	0.76	0.72	0.72	54
weighted avg	0.76	0.74	0.73	54

confusion_matrix :

```
[[27  3]
 [11 13]]
```

IV. RESULT

In this work, we have addressed the challenge of precision in predicting heart disease in previous base model, our model increases the accuracy of predicting Possibility of Cardiac arrest using multiple machine learning algorithm to advocate our predictions accuracy with Machine Learning algorithm such as Support Vector Machine Algorithm, Random Forest Algorithm, AdaBoost Algorithm and Gradient Boosting Algorithm also we have used medical terminologies to facilitate the heart disease prediction, and our model can be used to assist novice doctor.

RandomForestClassifier

classification_report :

	precision	recall	f1-score	support
1	0.76	0.73	0.75	30
2	0.68	0.71	0.69	24
accuracy			0.72	54
macro avg	0.72	0.72	0.72	54
weighted avg	0.72	0.72	0.72	54

confusion_matrix :

```
[[22  8]
 [ 7 17]]
```

V. CONCLUSION AND FUTURE DIRECTIONS

From this research work, we have built a predictive model for predicting Cardiovascular Disease. We tried to apply supervised algorithm that included using such as Support Vector Machine Algorithm, Random Forest Algorithm, AdaBoost Algorithm and Gradient Boosting Algorithm tree classifier and for evaluating the performance of the models, have used metrics as Area Under ROC Curve and Accuracy. Included some stages for making a proposed framework such as loading cardiovascular diseases dataset, cleaning and preprocessing of the dataset, classifiers, cross-validation & hyperparameter tuning, and at the end evaluating classifiers .The gradient boosting tree achieved the best results, and then we performed tuning with the ParamGridBuilder and the CrossValidator that help to improve the model performance. The results showed that the gradient-boosting tree classifier had accomplished the highest score of Area Under ROC and Accuracy at 80.02% and 76.20%. There will be a thorough study of huge datasets. with more attributes to attain the highest accuracy in future work. Another future work is; researchers can use deep learning techniques to improve the performance of the model.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54|

| Volume 5, Issue 6, June 2022 |

| DOI:10.15680/IJMRSET.2022.0506083 |

$$w(x_i, y_i) = \frac{1}{N},\ i = 1,2,\ldots.n \frac{1}{2}\log\frac{1 - Total\ Error}{Total\ Error}$$

REFERENCES

[1] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the World Congress on Engineering and computer

[2] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.

[3] Rajmohan, K., Paramasivam, I., & SathyaNarayan, S. (2014, February). Prediction and Diagnosis of Cardio Vascular Disease--A Critical Survey. In 2014 World Congress on Computing and Communication Technologies (pp. 246-251). IEEE.

[4] Sitar-WăXW $ =GUHQJKHD ' 3RS ' 6LWDU-WăXW ' (2009). Using machine learning algorithms in cardiovascular disease risk evaluation. Age, 1(4), 4.

[5] 0DU\ 1 .KDQ % ,VKIDT 4 .KDQ 0 = (PSLULFDO Study of Intelligence Techniques for Cardio Vascular Disease.

[6] Maini, E.,Venkateswarlu, B., & Gupta, A. (2018, August). Applying machine learning algorithms to develop a universal cardiovascular disease prediction system
In International Conference on Intelligent Data Communication Technologies and Internet of Things (pp. 627-632). Springer, Cham.

[7] Kumar, N. K., Sindhu, G. S., Prashanthi, D. K., & Sulthana, A. S. (2020, March). Analysis and prediction of cardiovascular disease using machine learning classifiers. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 15-21). IEEE.

[8] Dewan, A., & Sharma, M. (2015, March). Prediction of heart disease using a hybrid technique in data mining classification. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 704-706). IEEE.

[9] Shamsollahi, M., Badiee, A., & Ghazanfari, M. (2019). Using combined descriptive and predictive methods of data mining for coronary artery disease prediction: a case study approach. Journal of AI and Data Mining, 7(1), 47-58.

[10] Makumba, D. O., Cheruiyot, W., & Ogada, K. (2019). A model for coronary heart disease prediction using data mining classification techniques. Asian Journal of Research in Computer Science, 1-19.

[11] Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010, July). An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 183-192).

[12] Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-7). IEEE

[13] Kaur, H., & Kumari, V. (2020). Predictive modeling and analytics for diabetes using a machine learning approach. Applied computing and informatics.

[14] Ahmed, H., Younis, E. M., & Ali, A. A. (2020, February). Predicting Diabetes using Distributed Machine Learning based on Apache Spark. In 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE) (pp. 44-49). IEEE.

[15] Ali, A. A. (2019). Stroke Prediction using Distributed Machine Learning Based on Apache Spark. Stroke, 28(15), 89-97.

[16] Suma, V., & Hills, S. M. (2020). Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics. Journal of Soft Computing Paradigm (JSCP), 2(03), 153-159.

[17] Kumar, T. S. (2020). Data Mining Based Marketing Decision Support System Using Hybrid Machine Learning Algorithm. Journal of Artificial Intelligence, 2(03), 185- 193.

[18] Anand, J. V. "A Methodology of Atmospheric Deterioration Forecasting and Evaluation through Data Mining and Business Intelligence." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 02 (2020): 79-87.

[19] In 2022 Surbhi Kumari Student of Department of Computer Science, in Banasthali Vidyapith had developed heart disease prediction application Performance Evaluation of Distributed Machine Learning for Cardiovascular Disease Prediction in Spark.

Impact Factor
7.54

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com