Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR444

Heart Failure Prediction using Machine Learning Algorithms

R. Renugadevi[1]
Assistant Professor (Sr. G),
M.E. Department of CSE,
KIT – Kalaignar Karunanidhi
Institute of Technology (Autonomous),
Coimbatore, TN, India

Nivethitha. A[2]
Student
M.E. Computer Science and Engineering,
KIT – Kalaignar Karunanidhi
Institute of Technology (Autonomous),
Coimbatore, TN, India

Abstract:- This day and age individuals are increasingly giving precedence to their material needs as opposed to self-care, leading to physical and mental strain. Cardiovascular diseases (CVDs) present a significant menace worldwide, causing about 17.9 million deaths annually which is roughly 32% of global mortality. Heart failure, which impacts over 550,000 individuals on a yearly basis, emerges as an urgent global health concern. The formulation of effective prediction techniques for heart failure proves to be imperative in lessening its repercussions. Linear and machine learning models are put into service to forecast heart failure utilizing a myriad of inputs, comprising clinical data. With the burgeoning population, the early detection and intervention for heart disease grow more complex. Heart disease prevalence has escalated to concerning levels, culminating in untimely deaths due to arterial plaque accumulation. The premature pinpointing of heart disease holds the potential to rescue many lives by upholding arterial wellness. Our research integrates supervised machine learning algorithms to predict heart disease presence, underscoring methods to enhance classifier efficacy. Null values within the dataset are managed through mean value imputation, whereas irrelevant attributes are expunged utilizing information-gain feature selection. By wielding breakthroughs in machine learning (ML), the key aim of this study is to design prognostic models for cardiovascular disease utilizing 12 clinical attributes. By capitalizing on a dataset offered by Davide Chicco and Giuseppe Jurman, encompassing 12 clinical features and 299 data points, the efficacy of three ML algorithms: Support Vector Machine (SVM), Random Forest, and Logistic Regression is evaluated. Our examination discloses that Logistic Regression showcases the most outstanding accuracy and likelihood in foretelling cardio vascular disease presence. This predictive model exhibits potential in aiding healthcare experts in curtailing heart disease-linked fatalities.

Keywords:- *Random Forest, Support Vector Machine, Logistic Regression, Machine Learning Model, Heart Failure Prediction, Disease Prediction, Accuracy.*

I. INTRODUCTION

Machine learning is a powerful method that's used for analyzing vast datasets, digging up patterns, and grasping relationships within the data. This involves the collection, cleaning, pre-processing, and analysis of data to extract insights that are valuable. Situated within the domain of Artificial Intelligence (AI), machine learning operates on the principle that machines can autonomously learn and improve from data without explicit programming. One of the strengths of machine learning is its capability to decipher patterns within datasets that are immense, which enables machines to learn from experiences and make predictions. Detecting coronary heart diseases is becoming increasingly challenging in emerging countries due to the lack of specialized medical professionals and outdated examination tools, which make the process slow, and complex, this is a significant issue worldwide. Several factors contribute to coronary heart disease, such as smoking, high blood pressure, high cholesterol, and diabetes. Smoking, in particular, has become exceedingly prevalent among both young as well as old individuals and is currently trending among youth. It constricts the coronary arteries, leading to irregular heartbeats and an increase in blood pressure. High blood pressure poses numerous risks, as it compels the heart to pump harder to circulate blood throughout the body, leading to thickening of the lower-left heart chamber; consequently, increasing the risk of heart failure. Elevated blood sugar levels can also potentially cause damage to nerves or blood vessels that regulate heart function. However, we must be aware of the importance of maintaining a healthy heart through proper diet and exercise. This capability is notably useful for handling complex tasks beyond human capacity, resulting in significant time and cost savings. Its applications range across various domains, including self-driving cars, cyber fraud detection, facial recognition, and personalized recommendations on platforms like Facebook. Notably, prominent companies such as Netflix and Amazon utilize machine learning models for analyzing user behavior and customizing product recommendations. Machine learning algorithms are trained using datasets to build models, which in turn make predictions when new data is presented. These predictions are evaluated for accuracy, and, if found satisfactory, the machine learning algorithm is deployed. In scenarios where accuracy is lacking, the algorithm undergoes iterative

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR444

training with augmented datasets to boost performance. Machine learning, being a subset of artificial intelligence, focuses on developing and studying statistical algorithms that are capable of executing tasks without explicit instructions. Key algorithms like Support Vector Machine (SVM), Random Forest, and Logistic Regression play essential roles in constructing machine learning models.

II. RELATED WORK

Numerous researchers have utilized machine learning techniques, including Support Vector Machine (SVM), to devise strategies for predicting heart disease. For example, in their study titled "Heart Failure Prediction Using Machine Learning Algorithm," presented at the 3rd International Conference on Computation, Automation and Knowledge Management (ICCAKM) in 2022, Pandey and Kaur developed a system aimed at reducing heart diseases by addressing the challenge of similar symptoms, resulting in an impressive accuracy of 94.56% in diagnosis. During the research process, Pandey and Kaur employed a Support Vector Machine (SVM) to analyze the data collected from patients. The machine learning algorithm they used had proven to be effective in identifying patterns and trends, hence why it was chosen for this study. The findings of their study show that with proper utilization of machine learning algorithms, healthcare professionals can potentially enhance their ability to diagnose heart diseases accurately and efficiently. Additionally, the development of such systems can lead to early detection and intervention in individuals at risk, ultimately improving patient outcomes. Although machine learning algorithms can be greatly beneficial, it is important to note that they are not without limitations. Researchers must continue to refine these techniques and algorithms to ensure their accuracy and reliability. In conclusion, the utilization of machine learning techniques, such as the Support Vector Machine (SVN), presents promising opportunities for the field of healthcare, particularly in the prediction and diagnosis of heart diseases. Further research and development in this area will undoubtedly contribute to improved patient care and outcomes.

Boukhatem, Youssef, and Nassif proudly presented their latest paper "Heart Disease Prediction Using Machine Learning" at the 2022 Advances in Science and Engineering Technology International Conferences (ASET) in Dubai, UAE!!! They showcased the implementation of four classification techniques like Multilayer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB) to build predictive models for heart problems. Ahead of pattern creation, they completed data preprocessing and carefully selected features. Surprisingly, the SVM model stood out as the frontrunner, achieving an excellent accuracy rate of 91.67.

Abbas, Imran, Al-Aloosy, Fahim, Alzahrani, and Muzaffar presented their research titled "Heart Failure Prediction Using Machine Learning Approaches" at the 2022 Mohammad Ali Jinnah University International Conference on Computing (MAJICC) in Karachi, Pakistan. They employed various machine learning algorithms to detect and predict human heart disease, utilizing a heart disease dataset. Performance evaluation of these algorithms was conducted using metrics such as classification accuracy; F-measure, sensitivity, and specificity!!! They wanted to ascertain the effectiveness of these approaches in predicting heart failure accurately. Furthermore, the team explored other possibilities and experimented with different models to enhance the predictive capabilities of the machine learning systems being utilized. Despite some challenges faced during the research, the team remained resilient and focused on achieving their goals. In summary, their research sheds light on the potential benefits of using machine learning in predicting heart failure, which can ultimately aid in early detection and proactive intervention strategies for individuals at risk.

The study conducted by Montu Saw, Tarun Saxena, Sanjana Kaithwas, Rahul Yadav, and Nidhi Lal, titled "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning," released in January 2020, reported achieving an 87% accuracy using the logistic regression technique. Their research highlighted that men are more prone to cardiovascular disease compared to women. Additionally, factors such as aging, daily cigarette consumption, and systolic blood pressure were identified as influencing heart disease risk. Interestingly, the study found that total cholesterol alone did not significantly alter the likelihood of coronary heart disease (CHD), suggesting that the level of HDL within the total cholesterol value might be a contributing factor. Furthermore, the effects of glucose on CHD risk were deemed insignificant.The researchers suggested that further data collection and the utilization of additional machine learning models could potentially improve the predictive performance of the model. This underscores the importance of continuous refinement and validation of predictive models in healthcare research, particularly for complex and multifactorial conditions such as heart disease.

The study "Heart Disease Prediction Using Random Forest Algorithm," conducted by Kompella Sri Charan and Kolluru S S N S Mahendranath, was released in March 2022. In this study, the researchers evaluated the accuracy scores of several machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), AdaBoost, and Gradient Boosting, for identifying heart diseases. The study found that the Random Forest algorithm outperformed other algorithms, achieving an impressive accuracy rating of 92.16% in forecasting heart disease. This suggests that Random Forest is the most effective machine learning method for the identification of heart illnesses among those evaluated. Despite the promising results, the study indicates that there is still room for improvement and further potential for enhancing the predictive model. This highlights the importance of ongoing research and development efforts in the field of machine learning for healthcare applications, particularly in improving the accuracy and reliability of predictive models for diagnosing and managing heart diseases.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR444

➢ *Existing System*

World Health Organization (WHO) has recently released some astonishing statistics concerning the global burden of heart diseases. Shockingly, an estimated 12 million deaths worldwide are yearly associated with heart diseases! Unbelievable, right? Heart disease supposedly represents around 25% of deaths among individuals aged 25 to 69 years - a truly startling figure. In urban areas, this percentage seems to jump to a whopping 32.8%, while in rural areas, the numbers remain high at 22.9%. Furthermore, it's mind-boggling to know that over 80% of deaths worldwide are said to be due to heart disease WHO seems to believe that by 2030, the number of deaths because of heart disease could potentially reach nearly 23.6 million individuals. These statistics clearly shine a light on the pressing need for effective strategies to tackle and alleviate the impact that heart diseases have on global health.

➢ *Proposed System*

This study utilized machine learning methodologies to predict occurences of heart disease, showcasing the extensive potential of machine learning in healthcare sectors for prediction and categorization tasks. Notably, within healthcare, data mining precision has notably improved with the advent of machine learning techniques.The heart disease prediction model leverages machine learning, a field poised to revolutionize healthcare. Machine learning has demonstrated significant promise in improving prediction and classification accuracy, particularly in healthcare applications. Among the diverse algorithms explored, logistic regression stands out for its high accuracy rate in predicting heart disease.The model employs a variety of techniques and algorithms, including logistic regression, support vector machine (SVM), and random forest, to determine whether an individual is affected by heart disease. Through the integration of these algorithms and tools, the system aims to accurately classify individuals based on the presence or absence of heart disease.

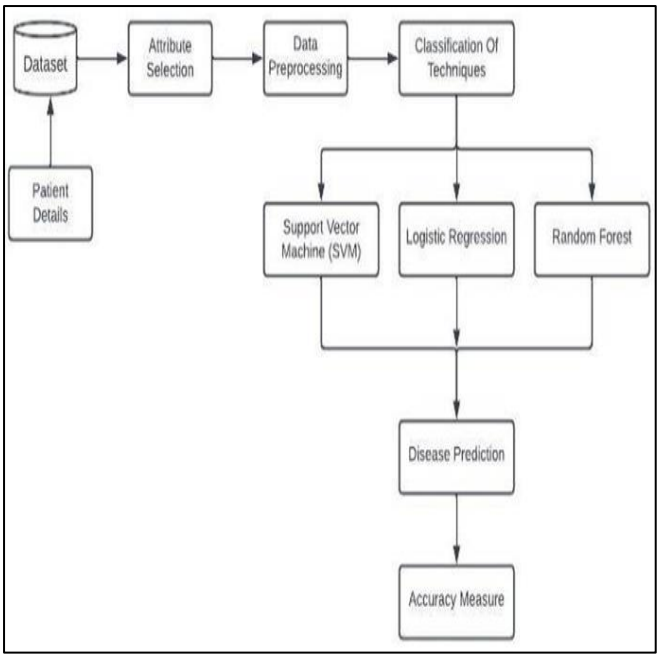


Fig 1 Architecture Diagram

➢ *Steps in Proposed System*

- *Data Collection:*

Starting with obtaining the dataset from Kaggle, you have patient details which is crucial for your project.

- *Attribute Selection (Feature Selection):*

This step involves identifying and selecting the most relevant features from your dataset that contribute the most to your model's predictive performance. It helps in reducing dimensionality and improving model interpretability.

- *Data Pre-Processing:*

Following attribute selection, you'll move on to data pre-processing. This step involves cleaning the data by handling missing, null, or corrupted values, and removing unnecessary features that were not selected in the attribute selection step.

- *Model Development:*

Once the data is pre-processed, you'll proceed to develop machine learning models. You've selected Support Vector Machine, Random Forest, and Logistic Regression due to their known accuracy and efficiency. This step involves training these models on your pre-processed data.

- *Model Evaluation and Selection:*

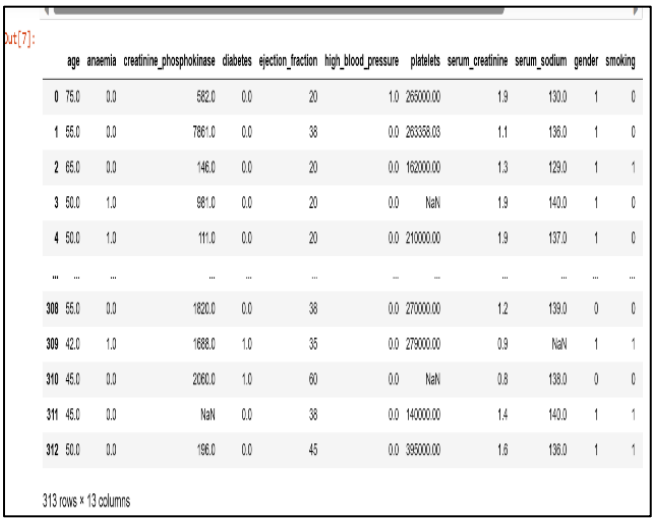
After training the models, you'll evaluate their performance using suitable evaluation metrics (such as accuracy, precision, recall, F1-score, etc.) on a separate test dataset. Then, you'll select the model with the highest accuracy among the three algorithms for deployment.

III. MODULE DESCRIPTION

➢ *Collection of Clinical Data*

The dataset used in this project originates from the research conducted by Davide Chicco and Giuseppe Jurman, titled "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," published in BMC Medical Informatics and Decision Making in 2020. This dataset comprises 13 features and 299 rows, specifically designed for predicting mortality due to heart failure. The attributes within the dataset are numeric, with one of the key features being the age of the patient. Age holds significant importance as it is considered the most critical risk factor in the development of heart diseases. Notably, the risk of heart disease doubles during adolescence, highlighting the relevance of age in predicting heart failure.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR444



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	gender	smoking
0	75.0	0.0	582.0	0.0	20	1.0	265000.00	1.9	130.0	1	0
1	55.0	0.0	7861.0	0.0	38	0.0	263358.03	1.1	136.0	1	0
2	65.0	0.0	146.0	0.0	20	0.0	162000.00	1.3	129.0	1	1
3	50.0	1.0	981.0	0.0	20	0.0	NaN	1.9	140.0	1	0
4	65.0	1.0	111.0	0.0	20	0.0	210000.00	1.9	137.0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
308	55.0	0.0	1820.0	0.0	38	0.0	270000.00	1.2	139.0	0	0
309	42.0	1.0	1688.0	1.0	35	0.0	279000.00	0.9	NaN	1	1
310	45.0	0.0	2060.0	1.0	60	0.0	NaN	0.8	138.0	0	0
311	45.0	0.0	NaN	0.0	38	0.0	140000.00	1.4	140.0	1	1
312	50.0	0.0	196.0	0.0	45	0.0	395000.00	1.6	136.0	1	1

313 rows × 13 columns

Fig 2 Overview of Dataset

➢ *Data – Pre-processing*

Data pre-processing is indeed a crucial step in data mining, as it involves preparing and transforming data into a suitable format for further analysis. This process serves to enhance the quality and effectiveness of data mining procedures. Data pre-processing encompasses several key tasks:

- *Data Cleaning:*

This involves identifying and rectifying errors or inconsistencies in the data, such as missing values, duplicate records, or inaccuracies.

- *Data Integration:*

In this step, data from multiple sources or formats are combined into a unified dataset. This ensures that all relevant data is available for analysis.

- *Data Reduction:*

Data reduction techniques are applied to reduce the size of the dataset while preserving its informational content. This may involve techniques such as sampling, aggregation, or dimensionality reduction.

- *Normalization:*

Normalization is performed to scale the numerical features of the dataset to a standard range, typically between 0 and 1. This ensures that all features contribute equally to the analysis and prevents biases due to differences in scale.

- *Outlier Detection and Removal:*

Outliers, or data points that deviate significantly from the rest of the dataset, can skew the results of data analysis. Hence, outlier detection techniques are used to identify and remove such outliers from the dataset.

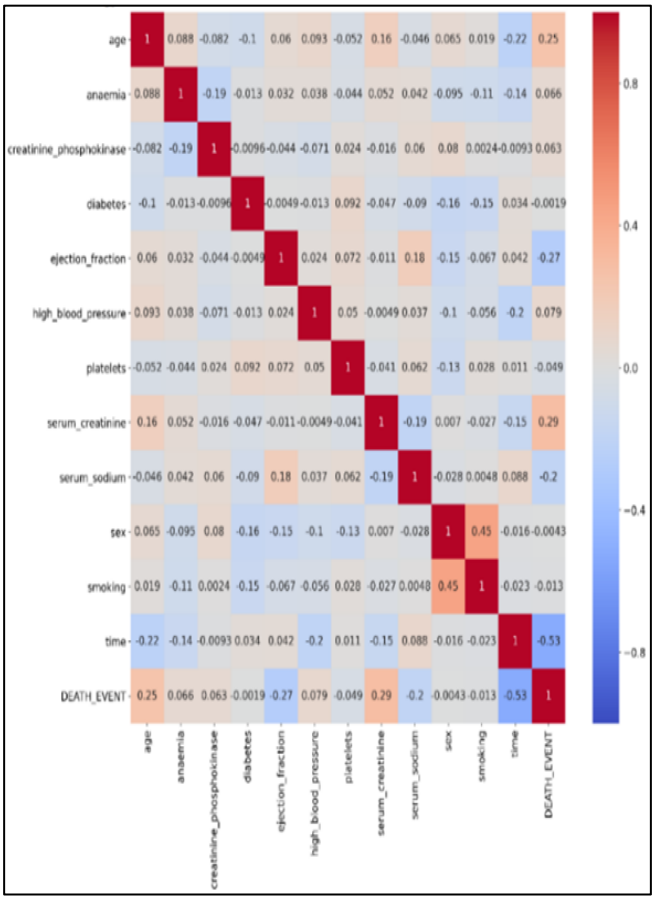


Fig 3 The Correlation between Different Columns

- *Feature Extraction:*

Feature extraction involves selecting or creating a subset of relevant features from the dataset. This helps reduce the dimensionality of the data and focuses on the most important aspects for analysis.

- *Feature Engineering:*

Feature engineering involves transforming or creating new features based on existing ones to improve the performance of machine learning algorithms.

Overall, data pre-processing plays a critical role in ensuring the quality, reliability, and effectiveness of data mining procedures, ultimately leading to more accurate and meaningful insights from the data.

IV. METHODOLOGIES

➢ *Support Vector Machine*

A Support Vector Machine (SVM) is a supervised learning algorithm utilized in machine learning to tackle intricate classification, regression, and outlier detection tasks. Indeed, SVM is a prominent method within supervised learning, extensively applied in statistical classification and regression analysis. It falls under the category of generalized linear classifiers, distinguished by their capacity to simultaneously minimize empirical error and maximize geometric edge regions. Hence, SVM is often referred to as a maximum edge region classifier.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR444

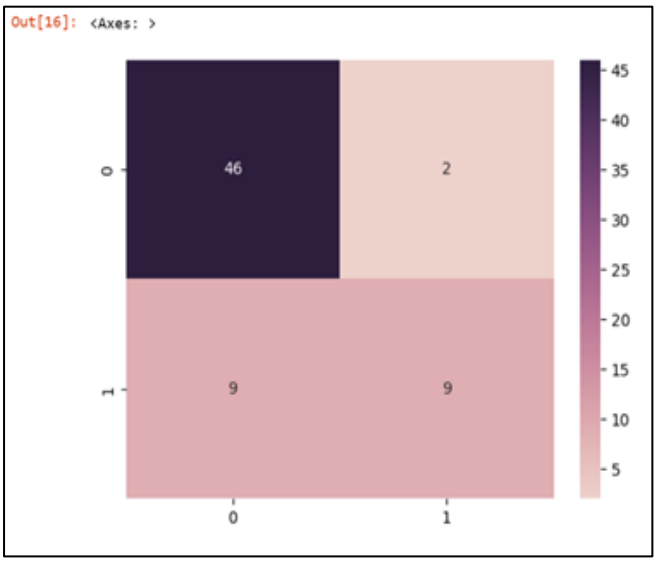


Fig 4 Confusion Matrix for SVM

➢ *Random Forest*

The Random Forest classifier is composed of numerous decision trees, and its output class is determined by the mode of the output classes of the individual trees. In mathematics, a random walk, also referred to as a drunkard's walk, is a stochastic process that depicts a trajectory comprising a series of random steps within a mathematical space. The process typically begins at a starting point, often denoted as 0, and at each subsequent step, it randomly moves either +1 or −1 with equal probability.During the training process, each decision tree in the Random Forest independently makes predictions, and then the final prediction is determined by aggregating the predictions of all the trees. In classification tasks, the most commonly occurring class among the predictions of all trees is chosen as the final prediction (mode). In regression tasks, the average of all predictions is taken as the final prediction. By averaging the predictions of multiple decision trees trained on different subsets of data, Random Forest can effectively reduce variance and produce more accurate and robust predictions compared to individual decision trees.

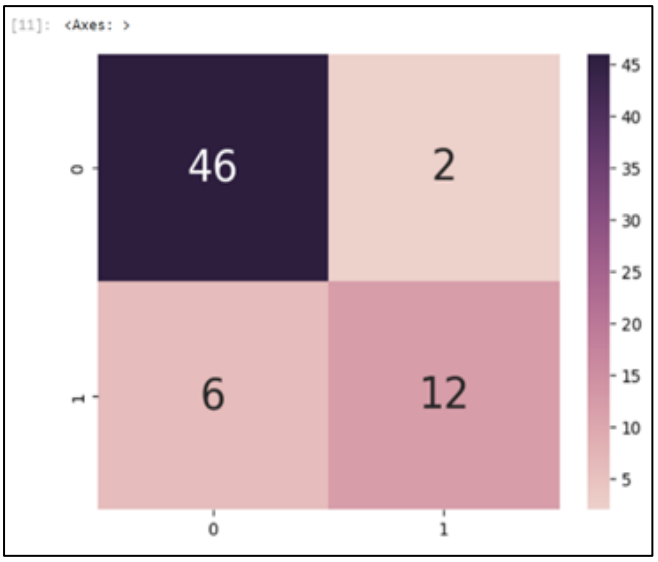


Fig 5 Confusion Matrix for Random Forest

➢ *Logistic Regression*

Logistic regression is a classification model frequently employed in binary scenarios. It serves as one of the fundamental algorithms in machine learning for addressing binary (0 or 1) problems, enabling the estimation of the likelihood of certain outcomes. Linear regression is utilized for solving regression problems, where the goal is to predict a continuous numerical outcome based on one or more input variables.

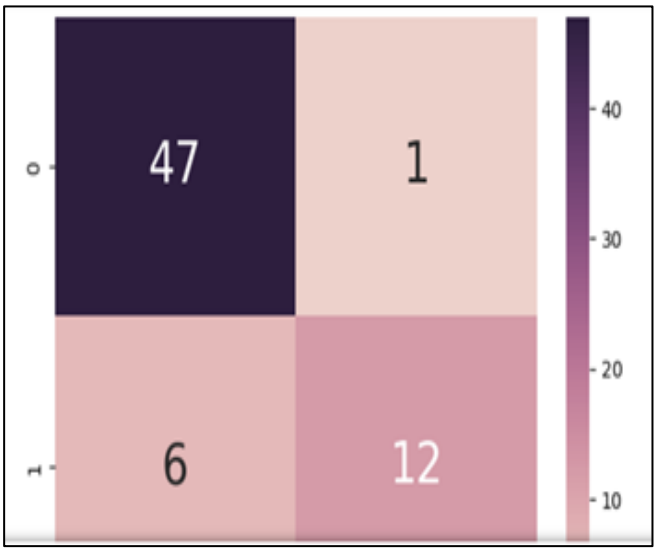


Fig 6 Confusion Matrix for Logistic Regression

V. CONFUSION MATRIX

The confusion matrix offers a detailed insight into model performance. Despite its name, you'll discover that the confusion matrix is a straightforward yet impactful concept. It comprises an N x N matrix utilized for assessing the performance of a classification model, with N representing the number of target groups. Within this matrix, the actual target values are compared against the predictions generated by the machine learning model. This holistic approach provides a comprehensive understanding of the model's performance and the nature of errors it may produce.

➢ *Accuracy:*

The accuracy metric measures the proportion of correct predictions relative to the total number of predictions made. It is computed by dividing the sum of true positives and true negatives by the total number of predictions.

➢ *Precision:*

Precision evaluates the accuracy of positive predictions by calculating the ratio of true positives to the sum of true positives and false positives.

➢ *Recall (Sensitivity):*

Recall assesses the model's ability to identify actual positives by determining the ratio of true positives to the sum of true positives and false negatives.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR444

➢ *F1 Score:*

The F1 score, also known as the F1 measure, offers a balanced assessment of precision and recall. It is computed as the harmonic mean of precision and recall, providing equal weight to both metrics. Actually, precision in the context of a confusion matrix refers to the proportion of correctly predicted positive instances among all instances predicted as positive by the model, regardless of whether they are true positive or false positive. It represents the accuracy of the model in identifying positive instances among all instances it predicted as positive. To summarize, precision measures the model's ability to correctly identify positive instances among all instances predicted as positive, regardless of whether they are true positive or false positive.

Table 1 Final Accuracy of Three Algorithms

Name of Classification Algorithm	Confusion Matrix	Accuracy
SVM	TP=46 FP=2 TN=9 FN=9	83.33%
LOGISTIC REGRESSION	TP=47 FP=1 TN=12 FN=6	89.39%
RANDOM FOREST	TP=46 FP=2 TN=12 FN=6	87.88%

VI. RESULTS AND DISCUSSIONS

Heart failure presents a significant global health challenge, impacting over 550,000 individuals annually. Enhancing prediction methods for this condition is pivotal in mitigating its effects. The results of the 13-feature classification models indicate the predictive performance of each model in determining the existence of certain outcomes, such as heart failure or heart disease. These models leverage a set of 13 features derived from clinical data or other relevant sources to make predictions. Leveraging advancements in machine learning (ML) within the healthcare sector, we can develop models capable of predicting heart diseases. This research employs a dataset comprising various clinical attributes such as patient age, sex, chest pain, and fasting blood sugar (Fbs), etc..,The dataset is then partitioned into two subsets: a training set comprising 70% of the data and a testing set comprising 30%. One critical aspect to consider in the analysis of these models is the presence of observations with null values. Null values, or missing data points, can significantly impact the reliability and accuracy of the models' predictions. It's essential to address these null values appropriately during the pre-processing stage to ensure the robustness of the models. Various techniques can be employed to handle null values, including mean imputation, median imputation, or removing observations with null values altogether. The choice of technique depends on the nature of the data and the specific requirements of the modelling task.The training set is utilized to construct the predictive model, while the testing set is employed to assess its accuracy. The research evaluates the dataset using three different algorithms, comparing their results. Through this methodology, the research achieves a prediction accuracy of 89.93% in determining whether a patient is suffering from heart disease (0 denoting absence, and 1 denoting presence). Notably, the Logistic Regression algorithm yields the highest accuracy of approximately 89.93% compared to other algorithms.

VII. CONCLUSIONS

In our study, we've highlighted the significance of managing null values and conducting feature selection to enhance the accuracy of classification models. Through comparisons of various classification methods, our aim was to identify the most effective classifier. By effectively managing null values, researchers can improve the overall quality of the dataset and enhance the reliability of the classification models' outcomes. This, in turn, contributes to more accurate predictions and better-informed decision-making in healthcare or related domains.We found that each classification method demonstrated commendable performance in handling observations with null values through mean imputation and employing the information gain feature selection strategy. However, it's worth noting that the dataset's lower-contributing features and presence of null values may have contributed to lower classification accuracy. Among the three classifiers assessed, the Logistic Regression Algorithm emerged as the most effective, achieving an overall accuracy of 89.93%, along with high recall scores. Moving forward, it's advisable to explore alternative classification algorithms that incorporate more robust feature selection methods for improved performance.

REFERENCES

[1]. C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Confserences (ASET), Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.

[2]. K. M. Hridoy et al., "Heart Disease Prediction Using Machine Learning Algorithms," 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, 2023, pp. 1-6, doi: 10.1109/IBDAP58581.2023.10271997.

[3]. S. Ibrahim, N. Salhab and A. E. Falou, "Heart disease Prediction using Machine Learning," 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), Jeddah, Saudi Arabia, 2023, pp. 1-6, doi: 10.1109/ICAISC56366.2023.10085522.

[4]. K. M. Hridoy et al., "Heart Disease Prediction Using Machine Learning Algorithms," 2023 4th International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, 2023, pp. 1-6, doi: 10.1109/IBDAP58581.2023.10271997.

[5]. Chicco, Davide, and Giuseppe Jurman. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC medical informatics and decision making 20.1 (2022): 16.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR444

[6]. B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin and X. Wei, "Predicting the Risk of Heart Failure with EHR Sequential Data Modeling," in IEEE Access, vol. 6, pp. 9256-9261,2018.
[7]. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.
[8]. Yilmaz R, Yagin FH. Early detection of coronary heart disease based on machine learning methods. International Medical Journal. 2022 Jan 1; 4(1): 1–6. doi: 10.37990/medr.1011924.
[9]. Riyaz L, Butt MA, Zaman M, Ayob O. Heart disease prediction using machine learning techniques: a quantitative review. International Conference on Innovative Computing and Communications, pp. 81–94, vol. 1394, Singapore: Springer; 2022.
[10]. S. Babu et al., "Heart disease diagnosis using data mining technique," 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2017, pp. 750-753, doi: 10.1109/ICECA.2017.8203643.