JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 26 Issue 02, 2023 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

HEART FAILURE PATIENTS ARE CLASSIFIED USING THE RANDOM FOREST AND NAÏVE BAYES ALGORITHMS

Badiaa Rahman Khalil
University of Suleimani
badiaa.khalil@univsul.edu.iq

Dr. Sozan Sabir Haider
University of Suleimani
sozan.haider@univsul.edu.iq

Dr. Mohammad Mahmood Faqe
University of Suleimani
mohammad.faqe@univsul.edu.iq

Abstract

Heart failure is the heart's inability to pump blood efficiently, causing symptoms. Annual deaths due to this condition reach hundreds of thousands globally, impacting millions. This study presents the utilization of two machine learning models, namely Random Forest and Naïve Bayes, to classify a dataset of 299 heart failure patients' data obtained from the UCI repository in 2015 based on their survival outcomes during follow-up. Employing distinct classification techniques, we thoroughly evaluate their performance through various valuation metrics. Our methodology involves training multiple decision trees on diverse subsets of data by ratio (80%), followed by the aggregation of their predictions to establish patient categories using the Random Forest technique. In parallel, the Naïve Bayes algorithm computes probabilities for each category based on patient attributes, assigning probabilities such as 0.68 for patients likely to pass away and 0.32 for those likely to survive. Various training-test ratios, including (60-40%), (70-30%) and (80-20%) are explored using the random forest approach in conjunction with Naïve Bayes. We demonstrate that the Random Forest classifier exhibits superior accuracy and predictive capability when compared to the Naïve Bayes classifier. With an 80% dataset training and 20% testing split, the Random Forest model achieves an accuracy rate of 85%, showcasing its robustness in categorizing patients effectively. Remarkably, the (80-20%) ratio consistently yields the highest accuracy, reaffirming the significance of optimal data partitioning for accurate patient classification. This study highlights the successful application of Random Forest and Naïve Bayes models to classify heart failure patients' survival outcomes. The Random Forest model outperforms the Naïve Bayes model in accuracy and predictive capability. The study emphasizes the importance of proper data partitioning and demonstrates the potential of machine learning techniques in medical research.

Keywords: Machine Learning, Random Forest, Naïve Bayes, Heart Failure, Survival Outcomes, Classification.

Submitted: 23/09/2023 **Accepted**: 13/10/2023

1.1 Introduction

Heart failure, a prevalent and intricate cardiovascular condition, significantly impacts global health by impairing the heart's ability to adequately pump blood, leading to insufficient oxygen and nutrient delivery to organs and tissues. Precise classification and prognostication of heart failure patients are imperative for tailoring effective treatment approaches and enhancing patient well-being. In recent years, the convergence of medical informatics and machine learning has paved the way for transformative advancements in healthcare. Machine learning algorithms offer substantial promise in analyzing extensive and intricate datasets. Random Forest and Naïve Bayes stand out for their efficacy in various medical applications, including disease classification (Smith et al., Brown et al) [3]. Random Forest, an ensemble learning technique, harnesses the strength of multiple decision trees to enhance classification accuracy and robustness. By constructing numerous decision trees during training and consolidating their outcomes, Random Forest can make precise classifications (Breiman, 2001) [4]. On the other hand, Naïve Bayes, a probabilistic algorithm founded on Bayes' theorem, simplifies the modeling process by assuming conditional independence among features, achieving satisfactory outcomes across diverse real-world scenarios (Russell & Norvig, 2016; Domingos & Pazzani, 1997)[7][13].Given the intricate nature of heart failure and the potential benefits of machine learning, the adoption of Random Forest and Naïve Bayes algorithms for heart failure patient classification holds significant promise. Accurate classification of heart failure patients into distinct subgroups based on clinical attributes, diagnostic parameters, and other pertinent features is pivotal for tailoring treatment strategies, predicting patient outcomes, and optimizing healthcare resource allocation.

This study endeavors to assess the viability and efficacy of employing Random Forest and Naïve Bayes algorithms to classify heart failure patients. Drawing on a comprehensive dataset encompassing patient profiles, clinical variables, and diagnostic indicators, this research seeks to explore the algorithms' capacity to differentiate between diverse stages. The findings of this investigation could furnish clinicians and healthcare practitioners with valuable insights into patient stratification, enabling personalized and precisely targeted medical interventions. In essence, the integration of machine learning algorithms, particularly Random Forest and Naïve Bayes, bears the potential to revolutionize the classification of heart failure patients. As the prevalence of heart failure escalates, innovative approaches to patient categorization assume paramount importance, holding the key to enhancing clinical decision-making and ultimately elevating patient care standards and outcomes.

1.2 Related Works

There have been several studies given in this section. Various scientific comparisons between the use of machine learning methods like Random Forest and Naïve Bayes for categorization in various applications were undertaken in the study. The majority of these studies used data from the medical field.

Using the information, Pal, M., et al. (2021) employed a random forest method to forecast a patient's likelihood of developing CVD. The dataset, which consists of 303 samples and uses

14 attributes to describe itself, is taken from the Kaggle website. The machine learning algorithm Random Forest is used to classify the datasets. Accuracy, sensitivity, and specificity are used to describe the dataset's findings. We found that the forest algorithm with randomization has an accuracy of 86.9%, sensitivity of 90.6%, and specificity of 82.7% for predicting CVD. Using random forests, the detection rate for CVD prediction is 93.3%. The random forest method has established itself as the most effective algorithm for classifying CVDs [11].

Lemons, K. (2020), Compares two machine learning techniques for identifying breast cancer. We employ two alternative machine learning techniques, Naïve Bayes and Random Forest, to evaluate the diagnostic precision. Using information from 569 patients and 31 attributes, the two machine learning classifiers indicated above are used. According to the findings, the Random Forest classifier fared better than the Naïve Bayes method, with an accuracy rate of 97.82% [10].

Using the idea of Heart Rate Variability (HRV), Shashikant et al (2019). suggested a method for the prediction of Cardiac Arrest in Smokers. A non-invasive method to evaluate the control of heartbeat is HRV. Correct data points must be collected at the ideal time and location. Comparisons were made between Decision Tree, Logistic Regression, and Random Forest results. To evaluate the effectiveness of all categorization algorithms, the 10-fold validation method is utilized. Results indicated that Random Forest had an accuracy of 93.61%, Decision Tree had an accuracy of 92.59%, and Logistic Regression had an accuracy of 89.7%. The best results were obtained using Random Forest. among various techniques [16].

In order to appropriately forecast CVD and overcome the missing value in the medical dataset, Zhou et al (2019). proposed a learning technique. To forecast the CVD, the algorithms Naïve Bayes, SVM, Decision Tree, Logistic Regression, RBF, and Random Forest were used. The findings demonstrate that RF, with its 88% sensitivity, 87.6% specificity, and 88% accuracy, outperformed other approaches even when values were missing [14].

According to Sarica, A. et al. (2017), the Random Forest (RF) method has been effectively used to decrease high-dimensional and multi-source data in many different scientific fields. Our goal was to investigate the state of the art in the use of RF to analyze single and multimodal neuroimaging data for the diagnosis of Alzheimer's disease. After a quantitative and qualitative screening, twelve papers from the years 2007 to 2017 were included in this systematic review. The takeaways from these studies point to RF as having one of the best accuracy levels to date for predicting the conversion of moderate cognitive impairment (MCI) to Alzheimer's disease (AD) [15].

1.3 Random Forest (RF)

Random forest is one of the classifier methods that is classified as supervised learning. This approach is utilized for both classification and regression issues, but it is frequently employed in classification issues. The random forest contains numerous decision trees as well as the output class, and individual trees serve as the class output mode **[1].** Random Forest is a decision tree-based algorithm that is used to solve classification problems. The formula is based on the

ensemble of decision trees. The algorithm works by creating a large number of decision trees and then combining their predictions to make the final prediction. The parameters of a random forest are the variables and thresholds used to split each node learned during training). The construction of the decision tree is done by selecting a collection of random variables (features). Finally, such a collection of random trees is called a Random Forest. RF is considered as one of the most accurate classification algorithms, due to the higher classification accuracy [5][6]. Another characteristic of RF is its significance for unbalanced and missing data compared to other alternative techniques [11].

Random forest algorithm procedure is as follows.

Step1: choose the random samples from the dataset.

Step2: Create a decision tree for each sample. From every decision tree will be produced the prediction result.

Step 3: Voting will be conducted on each expected outcome.

Step 4: Choose the predicted outcome that received the most votes.

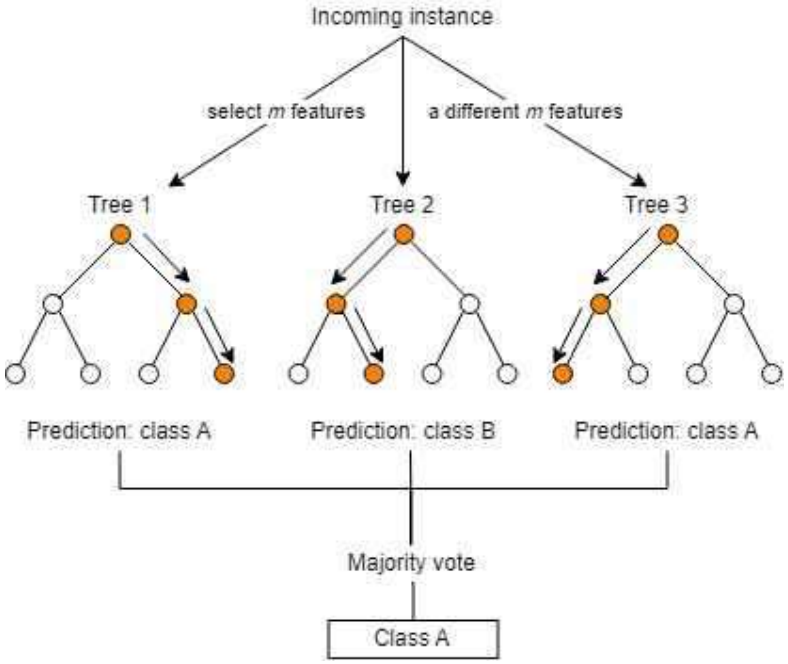

Figure (1) Random Forest method`s structure [17]

1.4. Bayesian Theory

Bayesian learning algorithm is the most practical learning approach for most learning problems and is based on evaluating explicit probabilities for hypotheses. Bayes learning classifier is extremely competitive with other learning algorithms and in many cases outperforms them. Bayesian learning algorithms are extremely important in machine learning since they provide unique perspective for understanding many learning algorithms that do not explicitly manipulate probabilities [8].

Bayes theorem states that:

Assuming that $|A| \neq 0$ and $|B| \neq 0$, we can state the following:

$$(A\backslash B) = \frac{|A\cap B|}{|B|} = \frac{\frac{|A\cap B|}{|\Omega|}}{\frac{|B|}{|\Omega|}} = \frac{P(A\cap B)}{P(B)} \quad (1)$$

$$(B\backslash A) = \frac{|B\cap A|}{|A|} = \frac{\frac{|B\cap A|}{|\Omega|}}{\frac{A}{|\Omega|}} = \frac{P(B\cap A)}{P(A)} \quad (2)$$

From Eq. 1 and Eq. 2, it is immediately obvious that:

$$P(A \cap B) = P(A|B)\, P(B) = P(B|A)\, P(A) \quad (3)$$

and therefore

$$P(A|B) = \frac{P(B\backslash A)P(A)}{P(B)} \quad (4)$$

which is the simplest (and perhaps the most memorable) formulation of Bayes' theorem. If the sample space **Ω** can be divided into finitely many mutually exclusive events *A1, A2, ...An,* and if B is an event with P(B) > 0, which is a subset of the union of all Ai, then for each Ai, the generalized Bayes' formula is

$$P(A_i\backslash B) = \frac{P(B\backslash Ai)P(Ai)}{\Sigma_{j=1}^{n} P(B\backslash Aj)P(Aj)} \quad (5)$$

which can be rewritten as

$$P(A\backslash B) = \frac{P(B\backslash A)P(A)}{P(B\backslash A)P(A)+P(B\backslash A^C)P(A^C)} \quad (6)$$

Both Eq. 6 and Eq. 5 follow from Eq. 4 because of the total probability theorem. Bayes' theorem can be used to derive the posterior probability of a hypothesis given observed data:

$$\boldsymbol{P\,(hypothesis\,|data)} = \frac{\boldsymbol{P(data|hypothesis)P(hypothesis)}}{\boldsymbol{P(data)}} \quad (7)$$

$$P(h|D) = \frac{P(D\backslash h)}{P(D)}$$

Were,

- P(h): Prior probability of hypothesis h- Prior
- P(D): Prior probability of training data D- Evidence
- P(D|h): Probability of D given h- Likelihood
- P(h|D): Probability of h given D- Posterior probability

In the general case, we have k mutually exclusive and exhaustive classes; hi, i = 1, ...n; $\boldsymbol{P(D\backslash h_i)}$is the probability of seeing D as the input when it is known to belong to class hi. The posterior probability of class hi can be calculated as-

$$\boldsymbol{P(h_i\backslash D) = \frac{P(D\backslash h_i)P(h_i)}{\Sigma_{i=1}^{n} P(D\backslash Ai)P(Ai)}} \quad (8)$$

In order to choose the best hypotheses from amongst the set of generated hypotheses the maximally probable hypothesis MAP is selected and is known as maximum a posteriori (MAP) hypothesis and if we assume that P(h) is same for all the hypothesis then the maximally probable hypothesis reduces to maximum likelihood hypothesis [14].

1.5 Naïve Bayes Classifier

Naïve Bayes is a probabilistic algorithm that is used to solve classification problems. It works by calculating the probability of a data point belonging to a particular class based on the probability of the features of the data point. We assume that a data set contains n instances (or

cases) $xi, = i..1$ which consist of p attributes, i.e., $xi = xi1, xi2, .., xip)$ Each instance is assumed to belong to one (and only one) classy $\in i$ {y1, y2, ..., yc}. Most predictive models in machine learning generate a numeric score s for each instance xi . This score quantifies the degree of class membership of that case in class yi . If the data set contains only positive and negative instances, y ∈ {0, 1}, then a predictive model can either be used as a ranker or as a classifier. The ranker uses the scores to order the instances from the most to the least likely to be positive. By setting a threshold t on the ranking score, s(x), such that {s(x) ≥ t} = 1, the ranker becomes a (crisp) classifier [12]. Naïve Bayes learning refers to the construction of a Bayesian probabilistic model that assigns a posterior class probability to an instance: $P(Y = yi|X = xi)$ P (Y = yi|X = xi). The simple Naïve Bayes classifier uses these probabilities to assign an instance to a class. Applying Bayes' theorem (Eq. 4), and simplifying the notation a little, we obtain

$$P(yi\backslash xi) = \frac{P(xi\backslash yi)}{P(xi)} \quad (9)$$

Note that the numerator in Eq. 9 is the joint probability of xi and yj (Eq. 3). The numerator can therefore be rewritten as follows; here, we will just use x, omitting the index i for simplicity:

$$P = (x|yi)P(yj) = P(x, yj) = P(x_1, x_2 \dots, x_p, y_j) \quad (10)$$

$$= P(x_1|x_2 \dots, x_p, y_j)\,(x_2, x_3 \dots, x_p, y_j) \quad \text{because P (a, b) = P(a|b) P(b)}$$

$$= P(x_1|x_{2,}x_{3,} \dots, x_p, y_j)\,P(x_2|x_{3,}x_{4,} \dots \dots, x_p, y_j)\text{P}\,(x_{3,}x_{4,} \dots, x_{p,}y_{j,})$$

$$= P(x_1|x_{2,}x_{3,} \dots, x_p, y_j)\,P(x_2|x_{3,}x_{4,} \dots, x_p, y_j)\dots\text{P}\,(x_{p,}\backslash y_{j,})P(yj)$$

Let us assume that the individual x_i are independent from each other. This is a strong assumption, which is clearly violated in most practical applications and is therefore Naïve—hence the name.

This assumption implies that $P = (x_{2|,}x_{3,}x_{4,}, \dots x_p, y_j) = P(x_1, y_j)$ for example. Thus, the joint probability of x and y_j is

$$P = (x|yi)P(yj) = P((x_1|yj).(x_2|yj) \dots P(x_p\backslash yj)P(yj) \quad (11)$$

$$\prod_{K=1}^{P} P = \left(x_{k,}\middle|y_j\right)\text{P}\,(y_j)$$

which we can plug into Eq. 9 and we obtain

$$P\left(y_j\backslash x\right) = \frac{\prod_{K=1}^{P} P=(x_{k,}|yi)\text{P}(y_j)}{P(x)} \quad (12)$$

Note that the denominator, P(x), does not depend on the class—for example, it is the same for class y_j and y_l . P(x) acts as a scaling factor and ensures that the posterior probability $P = (x|yi)$ is properly scale (i.e., a number between 0 and 1). When we are interested in a crisp classification rule, that is, a rule that assigns each instance to exactly one class, then we can simply calculate the value of the numerator for each class and select that class for which this value is maximal.

This rule is called the maximum posterior rule (Eq. 13). The resulting "winning" class is also known as the maximum a posteriori (MAP) class, and it is calculated as ˆy for the instance x as follows:

$$\hat{Y} = argmax \prod_{K=1}^{P} P = \left(x_{k,}\middle|y_i\right)\text{P}\,(y_j) \quad (13)$$

A model that implements Eq. 11 is called a (simple) Naïve Bayes classifier.

A crisp classification, however, is often not desirable. For example, in ranking tasks involving a positive and a negative class, we are often more interested in how well a model ranks the cases of one class in relation to the cases of the other class [10].

The estimated class posterior probabilities are natural ranking scores. Applying again the total probability theorem (Eq. 3), we can rewrite Eq. 12 as

$$P(y_{j,}|x) = \frac{\prod_{K=1}^{P} P=(x_{k,}|y_i)\mathrm{P}(y_j)}{\prod_{K=1}^{P} P=(x_{k,}|yi)\mathrm{P}(yj)+\prod_{K=1}^{P} P\left(x_{k,}|y_j^c\right)\mathrm{P}(y_j^c)} \quad (14)$$

2.Materials and Description of the dataset:

The Cleveland Database, a UCI source, provided the dataset [18]. The patient profiles in the dataset, which includes 299 heart failure patients who were followed up on, each contain 13 clinical characteristics. Each row has one patient record. One of the 13 qualities of the record is a predictive trait called Y, whose value shows the kind of heart failure (whether the patient died during the follow-up time or the patient did not die during the follow-up period). The final 12 attributes are used in the algorithm's prediction stage. There are 13 distinct traits in total. The table below displays the dataset that was used in this investigation.

Table (1): Dataset Description

No	Attribute Name	Attribute Description	Values
1	Age	age of the patient (years)	No particular range
2	Sex	woman or man (binary)	Female = 0 Male = 1
3	Anemia	decrease of red blood cells or hemoglobin (boolean)	
4	High blood pressure	if the patient has hypertension (boolean)	No particular range
5	Creatinine phosphokinase (CPK)	level of the CPK enzyme in the blood (mcg/L)	Fasting blood sugar > 120 mg/dl True =1 and False = 0
6	Diabetes	if the patient has diabetes (boolean)	Normal = 0 Abnormal = 1
7	Ejection fraction	percentage of blood leaving the heart at each contraction (percentage)	No = 0 Yes = 1
8	Platelets	platelets in the blood (kiloplatelets/mL)	No = 0 Yes = 1
9	Serum creatinine	level of serum creatinine in the blood (mg/dL)	No = 0 Yes = 1
10	Serum sodium	level of serum sodium in the blood (mEq/L)	No = 0 Yes = 1

11	Smoking	if the patient smokes or not (boolean)	No = 0 Yes = 1
12	Time	Follow -up period (dayes)	Nominal value
13	[target] death event Y = Type	, Yi = +1 A , Yi = -1 B	if the patient deceased during the follow-up period (boolean) = A if the patient not deceased during the follow-up period (boolean) = B

3 Results and Discussion

3.1 Measures for Performance Evaluation

In this section, we go over the machine learning-based heart failure predictive diagnosis process. The process is broken down into numerous steps, as depicted in Figure 1.

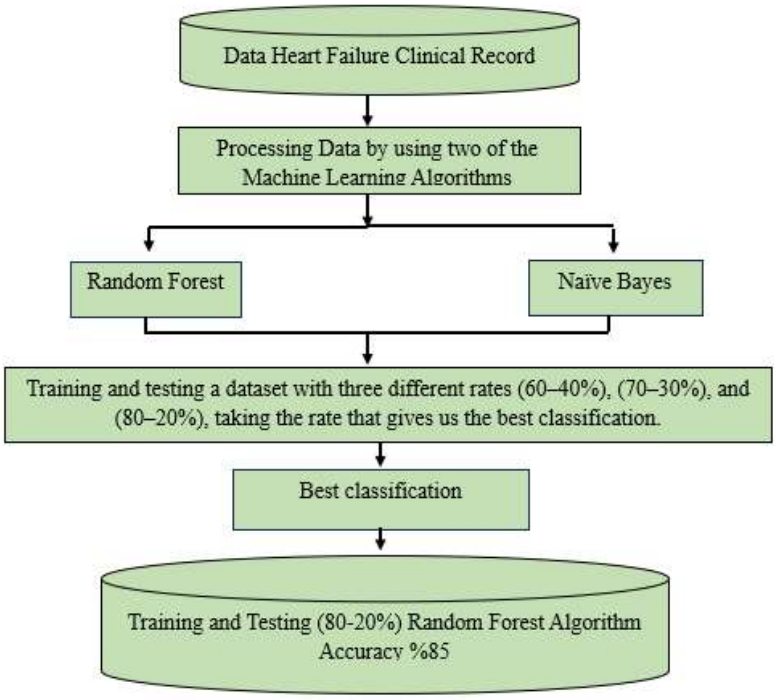

Figure 1: Discusses the classifying process in detail

3.2 Confusion Matrix

The confusion matrix is used for identifying the mislabeling or error in prediction. It matches the actual and predicted values with four elements (True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)). Type-I and Type-II errors are seeded by False

Positive and False Negative values. The confusion matrix is very expedient to calculate Precision, Recall, F1-score and Accuracy [2][19].

Table (2): Confusion matrix

Actual Class	Predicted Class	
	patient not deceased	patient deceased
patient not deceased	True Positive (TP)	False Negative (FN)
patient deceased	False Positive (FP)	True Negative (TN)

This paper tests the efficiency of the technique proposed using precision, specificity, sensitivity and geometric mean. The right prediction in proportion to the total number of predictions made by a classifier decides its accuracy which is formulated as [9]:

$$Accuracy(MI, HF) = \frac{(TP+TN)}{(TP+FP+FN+TN)} * 100\% \quad (14)$$

Were,
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

3.3 Analysis of data by using Random forests and Nave Bayes

In this paper, we have performed a number of steps in order to best classify our data through the program Weka using both machine learning models that are random forest and Naïve bayes. The definition of core predictive characteristics is used to categorize heart failure symptoms (if the patient died while being followed up or a patient who survived the time of follow-up without dying).

Table (3): Accuracy random forest and Naïve bayes for different ratio training and testing for data set

Accuracy Random Forest and Naïve bayes			
Model	Training 60% & Testing 40%	Training 70% & Testing 30%	Training 80% & Testing 20%
Random Forest	80%	82%	85%
Confusion Matrix RF	CM=$\begin{bmatrix}27 & 19\\ 5 & 69\end{bmatrix}$	CM=$\begin{bmatrix}22 & 12\\ 4 & 52\end{bmatrix}$	CM=$\begin{bmatrix}16 & 6\\ 3 & 35\end{bmatrix}$
Naïve Bayes	74.2%	74.4%	80%
Confusion Matrix NB	CM=$\begin{bmatrix}20 & 26\\ 5 & 69\end{bmatrix}$	CM=$\begin{bmatrix}15 & 19\\ 4 & 52\end{bmatrix}$	CM=$\begin{bmatrix}12 & 10\\ 2 & 36\end{bmatrix}$

This above table illustrates the varying accuracy rates of the Random Forest and Naïve Bayes models across different testing proportions, highlighting the Random Forest model's superiority in accuracy. The included confusion matrices provide insights into the models' true positive, true negative, false positive, and false negative predictions under each testing condition.

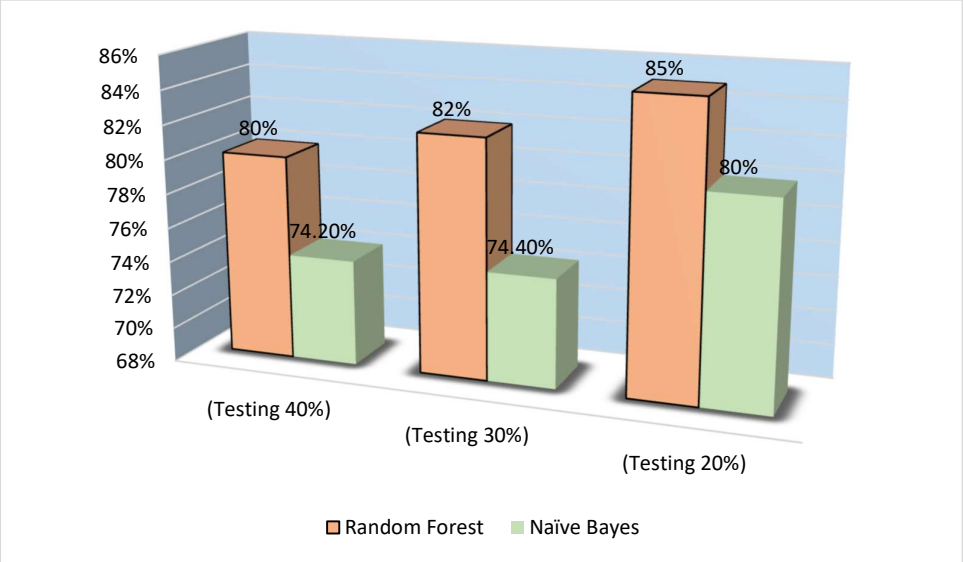

Figure (2): Accuracy Random Forest and Naïve Bayes

Table (4)**:** Confusion Matrix Random Forest for Training 80% and Testing 20% of the dataset

	Predicted Class		
Actual Class	Patient not deceased	patient deceased	Sum
patient not deceased	16 (TP)	3 (FN)	19
patient deceased	6 (FP)	35 (TN)	41
Sum	22	38	60

The above confusion matrix provides multiple different details about the results obtained from the evaluated method. As we mentioned earlier, 80% of the preprocessed dataset has been used as training set and the remaining 20% as testing. There are 60 testing instances or information from 60 patients has been used to test the evaluated method. 22 patients were actually diagnosed where patients did not die during the follow-up period and the remaining 38 patients were diagnosed due to patients dying during the follow-up period. The RF classifier was able to accurately classify the disease for 51 of the patients (TP and TN) and failed to correctly classify the status of 9 patients (FP and FN).

Table (5): Confusion Matrix Naïve Bayes for Training 80% and Testing 20% of the dataset

	Predicted Class		
Actual Class	Patient not deceased	patient deceased	Sum
patient not deceased	12 (TP)	2 (FN)	14
patient deceased	10(FP)	36 (TN)	46
Sum	22	38	60

The above confusion matrix provides multiple different details about the results obtained from the evaluated method. As we mentioned earlier, 80% of the preprocessed dataset has been used as training set and the remaining 20% as testing. There are 60 testing instances or information from 60 patients has been used to test the evaluated method. 22 patients were actually diagnosed where patients did not die during the follow-up period and the remaining 38 patients were diagnosed due to patients dying during the follow-up period. The RF classifier was able to accurately classify the disease for 48 of the patients (TP and TN) and failed to correctly classify the status of 12 patients (FP and FN).

Table (6): Detailed Accuracy by Class Random Forest for Training 80% and Testing 20%

Class	TP Rate	FP Rate	Precision	Recall (Sensitivity)	F-Measure	ROC Area	PRC Area
Patient not deceased	0.727	0.079	0.842	0.727	0.780	0.964	0.937
Patient deceased	0.921	0.273	0.854	0.921	0.886	0.964	0.981
Weighted Avg	0.850	0.202	0.849	0.850	0.847	0.964	0.965

The average weights of the criteria for both groups, taking into consideration their distribution, are shown in the above table. For instance, the weighted average precision is 0.849, the weighted average TP Rate is 0.850, and so forth. These measures give an all-encompassing picture of model performance, taking into account factors like sensitivity, specificity, accuracy, and trade-offs between them.

3.4 The criteria details used to evaluate the performance of our proposed model are as follows:

P= Total Number of patients not deceased =22, N= Total Number of patient deceased =38

- $TPR(\text{patients not deceased}) = \frac{TP}{P} = \frac{16}{22} = 0.727$
- $TNR(\text{ patient deceased}) = \frac{TN}{N} = \frac{35}{38} = 0.921$
- $FPR(\text{patients not deceased}) = 1 - TNR(HF) = 1 - 0.921 = 0.079$
- $FNR(\text{ patient deceased}) = 1 - TPR(MI) = 1 - 0.727 = 0.273$

- $precision(\text{patients not deceased}) = \frac{TP}{TP+FP(PD)} = \frac{16}{16+3} = 0.842$
- $precision(\text{ patient deceased}) = \frac{TN}{TN+F\ \ (PD)} = \frac{35}{35+6} = 0.854$
- $Sensitivity$ patients not deceased $(Recall) = \frac{TP}{TP+FP(PND)} = \frac{16}{16+6} = 0.727$
- $Sensitivity$ patient deceased$(Recall) = \frac{TN}{TN+F\ \ (PD)} = \frac{35}{35+3} = 0.921$
- $Weighted\ Avg(TPR) = \frac{(\text{TPR (PND)}*\text{P})+(\text{TNR(PD)}*\text{N})}{\text{P+N}}$
- $\text{Weighted Avg(TPR)} = \frac{(0.727*22)+(0.921*38)}{22+38} = 0.850$

⋮

- $\text{Weighted Avg(PRC Area)} = 0.965$

Table (7): Detailed Accuracy by class Naïve Bayes for Training 80% and Testing 20%

Class	TP Rate	FP Rate	Precision	Recall (Sensitivity)	F-Measure	ROC Area	PRC Area
Patient not deceased	0.545	0.053	0.857	0.545	0.667	0.829	0.820
Patient deceased	0.947	0.455	0.783	0.947	0.857	0.829	0.823
Weighted Avg	0.800	0.307	0.810	0.800	0.787	0.829	0.822

The average weights of the criteria for both groups, taking into consideration their distribution, are shown in the above table. As an illustration, the weighted average TP rate is 0.850, the weighted average precision is 0.849, etc. These measures give an all-encompassing picture of model performance, taking into account factors like sensitivity, specificity, accuracy, and trade-offs between them.

Table (8): The comparison between random forests and Naïve bayes for measuring training is 80% and testing is 20%.

Measure	Random forest		Naïve Bayes	
Correctly Classified Instances (Accuracy)	51	85%	48	80%
Incorrectly Classified Instances (Error rate)	9	15%	12	20%
Specificity		72.73%		78.26%

The data in the table above displays the classification accuracy and error rates for the "Random Forest" and "Nave Bayes" models. Compared to the "Naïve Bayes" model's (80%)

accuracy, the "Random Forest" model performed better (85%). The error rates also show that, as compared to the "Naïve Bayes" model, the "Random Forest" model had a lower rate of wrong classifications (15%). To fully comprehend the success of the models, it's crucial to take into account additional measures like precision, recall, and specificity. Because specificity is a model's capacity to forecast the actual negative of each class, the lower our specificity, the greater our accuracy.

- Accuracy (PND, PD) Random Forest = $\frac{TP+TN}{(TP+FP+FN+T\ )} * 100 = \frac{16+35}{16+6+3+35} * 100 =$ 85%
- Accuracy (PND, PD) Naïve Bayes = $\frac{12+36}{12+10+2+3} * 100 = 80\%$
- Error rate (PND, PD) Random Forest = $\frac{FP+F}{TP+FP+TN+FN} * 100 =$ $\frac{6+3}{60} = 15\%$
- Error rate (PND, PD) Naïve Bayes = $\frac{10+2}{60} * 100 = 20\%$
- Specificity (PND, PD) Random Forest =$\frac{TN}{(TN+FP)} * 100 = \frac{16}{16+6} =$ 72.73 %
- Specificity (PND, PD) Naïve Bayes =$\frac{36}{36+10} * 100 = 78.26\ \%$

Table (9): Random forests and Naïve bayes are compared for calculating training data (80%) and testing data (20%).

Measure	Random forest	Naïve Bayes
Kappa Statistic	0.6675	0.5337
Mean absolute error	0.2075	0.2577
Root mean squared error	0.301	0.3854
Relative absolute error	46.1466%	57.3135%
Root relative squared error	62.0467%	79.4449%

According to the data in the previous table (9), the random forest kappa statistic is higher than the kappa statistic of the Naïve Bayes statistic because, when the data are classified by the random forest model, the results of each measure (mean absolute error, root mean square error, relative absolute error, and square error) root ratio) is lower than the Naïve Bay model because the higher the accuracy, the higher the kappa statistic, and the lower these measures. Moreover, the opposite is true.

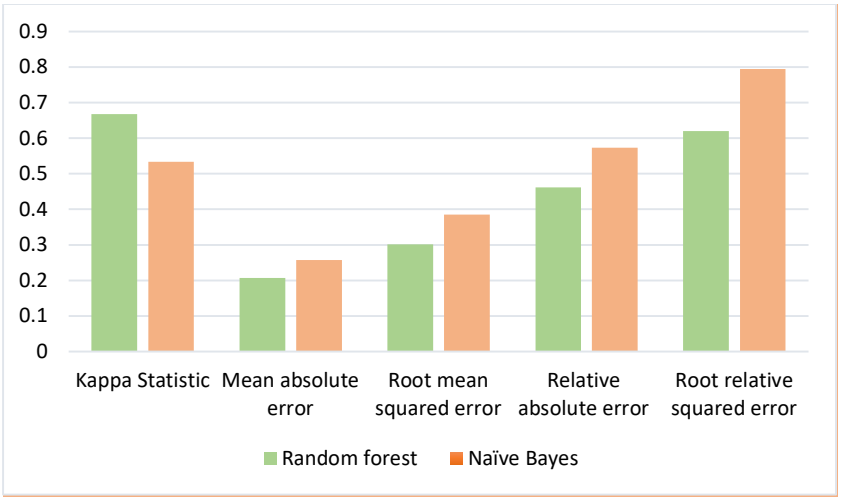

Figure (3): Measures: 80% for training and 20% for testing for Random Forest and Naïve Bayes

4. Conclusion

This study demonstrates the use of Random Forest and Nave Bayes models to categorize the survival outcomes of heart failure patients. With an 80-20% training-testing split, the Random Forest model regularly outperforms Naïve Bayes in terms of accuracy and predictive power. For accurate patient classification, the best possible data partitioning is essential. These results highlight the potential of machine learning in the field of medicine and highlight the need of data processing in producing accurate forecasts for heart failure patients.

References:

[1] Anitha, S., & Vanitha, M. (2022, February). Classification of VASA Dataset Using J48, Random Forest, and Naïve Bayes. In *Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)* (pp. 283-291). Singapore: Springer Nature Singapore.

بديعة رحمن خليل, أ. م. د. محمد محمود فقي, & أ. م. د. سوزان صابر حيدر.

[2] .(2020)

Classifying Patients with Myocardial Infarction and Heart Failure by Using SVM and KNN Learning Techniques .*Journal of Administration and Economics*.327-315 ,(126) ,

[3] Brown, L. K., Williams, R. S., & Garcia, M. L. Application of Random Forest and Naïve Bayes algorithms in heart failure patient classification. *Cardiovascular Computing and Applications*,

[4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[5] Breiman, L. (2001). Random forests. Machine Learning 45 5–32.

[6] Biau, G. and Scornet, E. (2016). A random forest guided tour. Test 25 197–227.

[7] Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 29(2-3), 103-130.

[8] E. Alpaydin, “An Introduction to Machine Learning” The MIT press, Cambridge, Massachusetts, London, England, 2004.
[9] Faqe, M. Mohammad, S. & Hassan, K. “Using Random Forest algorithm to classify Iron anemia unspecified and coagulation defect unspecified diseases”. A scientific Journal Issued by University of Sulaimani, Part (B- for Humanities) No. (68) (15 Feb.2022) .
[10] Lemons, K. (2020). A comparison between Naïve bayes and random forest to predict breast cancer. *International Journal of Undergraduate Research and Creative Activities*, *12*(1).
[11] Lemons, K. (2020). A comparison between Naïve bayes and random forest to predict breast cancer. *International Journal of Undergraduate Research and Creative Activities*, *12*(1).
[12] Pal, M., & Parija, S. (2021, March). Prediction of heart diseases using random forest. In *Journal of Physics: Conference Series* (Vol. 1817, No. 1, p. 012009). IOP Publishing.
[13] Rasul, Comparison between SVM and K-NN with application for diagnosis of heart disease patient. 2021.Master thesis, College of Administration & Economic, University of Sulaimani .
[14] Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson.
[15] Runchuan Li, Shengya Shen, Xingjin Zhang, Runzhi Li, Shuhong Wang, Bing Zhou and Zongmin Wang, “Cardiovascular Disease Risk Prediction Based on Random Forest”, Proceedings of the 2nd International Conference on Healthcare Science and Engineering, vol. 536, pp. 31-43, May 2019.
[16] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Frontiers in aging neuroscience*, *9*, 329.
[17] Shashikant R, Chetankumar P, “Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter”, Applied Computing and Informatics, June 2019.
[18] https://www.mararkcr.top/ProductDetail.aspx?iid=156897435&pr=41.88
[19] https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

خلاصة

فشل القلب هو عدم قدرة القلب على ضخ الدم بكفاءة، مما يسبب الأعراض. تصل الوفيات السنوية الناجمة عن هذه الحالة إلى مئات الآلاف على مستوى العالم، مما يؤثر على الملايين. تعرض هذه الدراسة استخدام نموذجين للتعلم الآلي، وهما Random Forest وNaïve Bayes، لتصنيف مجموعة بيانات مكونة من 299 بيانات لمرضى قصور القلب تم الحصول عليها من مستودع UCI في عام 2015 بناءً على نتائج بقائهم على قيد الحياة أثناء المتابعة. ومن خلال استخدام تقنيات تصنيف متميزة، نقوم بتقييم أدائها بدقة من خلال مقاييس التقييم المختلفة. تتضمن منهجيتنا تدريب أشجار قرارات متعددة على مجموعات فرعية متنوعة من البيانات بنسبة (80%)، يليها تجميع توقعاتهم لتحديد فئات المرضى باستخدام تقنية Random Forest. بالتوازي، تحسب خوارزمية Naïve Bayes الاحتمالات لكل فئة بناءً على سمات المريض، وتخصيص احتمالات مثل 0.68 للمرضى الذين يحتمل أن يموتوا و0.32 لأولئك الذين يحتمل أن يظلوا على قيد الحياة. تم استكشاف نسب اختبار التدريب المختلفة، بما في ذلك (60-40%)، (70-30%) و(80-20%) باستخدام نهج الغابة العشوائية بالاشتراك مع Naïve Bayes. لقد أثبتنا أن مصنف Random Forest يُظهر دقة فائقة وقدرة تنبؤية عند مقارنته بمصنف

.Naïve Bayesمن خلال التدريب على مجموعة البيانات بنسبة 80% وتقسيم الاختبار بنسبة 20%، يحقق نموذج Random Forestمعدل دقة يصل إلى 85%، مما يعرض قوته في تصنيف المرضى بشكل فعال. ومن اللافت للنظر أن النسبة (80-20%) تنتج دائمًا أعلى دقة، مما يؤكد من جديد أهمية التقسيم الأمثل للبيانات من أجل التصنيف الدقيق للمريض. تسلط هذه الدراسة الضوء على التطبيق الناجح لنماذج Random Forest و Naïve Bayesلتصنيف نتائج بقاء مرضى قصور القلب على قيد الحياة. يتفوق نموذج Random Forest على نموذج Naïve Bayes من حيث الدقة والقدرة التنبؤية. تؤكد الدراسة على أهمية تقسيم البيانات بشكل صحيح وتوضح إمكانات تقنيات التعلم الآلي في البحث الطبي.

الكلمات المفتاحية: التعلم الآلي، الغابة العشوائية، ساذجة بايز، فشل القلب، نتائج البقاء، التصنيف.