Research Article

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR HEART DISEASE PREDICTION

Ayat Bahaa ABDULHUSSEIN [†], Turgay Tugay BİLGİN [††]

[†] Bursa Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye

[††] Bursa Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye

[†]**Ayatbahaa21@gmail.com,** [††]**turgay.bilgin@btu.edu.tr**

Orcid.org/0000-0001-6511-8171, Orcid.org/0000-0002-9245-5728

Atıf/Citation: ABDULHUSSEIN, A.B., BİLGİN, T.T., (2024). Comparison of Machine Learning Algorithms for Heart Disease Prediction, Journal of Technology and Applied Sciences 7(1) s. 133-146,
DOI: 10.56809/icujtas.1433853

ABSTRACT

Machine learning, one of the most well-known applications of artificial intelligence, is altering the world of research. The aim of this study is to generate predictions for Heart Disease Prediction (HDP) by employing effective machine learning approaches and to predict whether an individual has heart disease. The primary objective is to evaluate the predictive accuracy of various machine learning algorithms in predicting the presence or absence of heart disease. The KNIME data analysis program has been selected, and overall accuracy is chosen as the primary indicator to assess the effectiveness of these strategies. Utilizing details such as chest pain, cholesterol levels, age, and other factors, along with different machine learning technologies such as K Nearest Neighbor (KNN), Naive Bayes, and Logistic Regression, a dataset of 319,796 patient records with 18 attributes was utilized. Naive Bayes, K Nearest Neighbor (KNN), and Logistic Regression were employed as machine learning techniques, and their prediction accuracies were compared. The application results indicate that the logistic regression approach outperforms the K Nearest Neighbor method and the Naive Bayes method in terms of predicting accuracy for heart disease. The prediction accuracy of K-NN is %90,77, Naive Bayes is %86,633, and logistic regression is %91,60. In conclusion, machine learning algorithms can accurately identify heart disease. The results suggest that these methods could assist doctors and heart surgeons in determining the likelihood of a heart attack in a patient.

Keywords: Naïve Bayes Algorithm, Logistic Regression, K Nearest Neighbor, Heart Disease.

KALP HASTALIĞI TAHMİNİNDE MAKİNE ÖĞRENİMİ ALGORİTMALARININ PERFORMANS KARŞILAŞTIRMASI

ÖZET

Makine öğrenimi, araştırma dünyasını değiştiren, yapay zekânın en bilinen uygulamalarından biridir. Bu araştırmanın hedefi, etkili makine öğrenimi yaklaşımlarını kullanarak Kalp Hastalığı Tahmini için tahminler üretmek ve kişinin kalp hastalığına sahip olup olmadığını tahmin etmektir. Temel amaç, çeşitli makine öğrenimi algoritmalarının kalp hastalığının varlığını veya yokluğunu tahmin etmedeki öngörü doğruluğunu değerlendirmektir. KNIME veri analizi programı genel doğruluk, bu stratejilerin etkinliğini değerlendirmek için temel gösterge olarak seçilmiştir. Göğüs ağrısı, kolesterol seviyeleri, bir kişinin yaşı ve diğer faktörler gibi detaylar kullanılarak ve K En Yakın Komşu (KNN), Naif Bayes ve Lojistik Regresyon gibi farklı makine öğrenimi teknolojileri kullanılarak, 319796 hasta kaydı ve 18 niteliğe sahip bir veri seti kullanılmıştır. Makine öğrenimi teknikleri olarak Naive Bayes, K En Yakın Komşu (KNN) ve Lojistik Regresyon kullanılmış ve tahmin doğrulukları karşılaştırılmıştır. Uygulama sonuçları, lojistik regresyon yaklaşımının kalp hastalığı için tahmin doğruluğu açısından K En Yakın Komşu yönteminden ve Naive Bayes yönteminden daha iyi olduğunu göstermektedir. K-NN'nin tahmin doğruluğu %90,77, Naive Bayes'in %86,633 ve lojistik regresyonun %91,60'dır. Sonuç olarak, makine öğrenimi algoritmalarının kalp hastalığını büyük oranda doğru bir şekilde tanımlayabileceği görülmüştür. Sonuçlar, bu yöntemlerin bir hastada kalp krizi olasılığını belirlemede doktorlara ve kalp cerrahlarına yardımcı olabileceğini göstermektedir.

Anahtar Kelimeler: Naïve Bayes Algoritması, Lojistik Regresyon, K En Yakın Komşu, kalp hastalığı.

Geliş/Received	:	**08.02.2024**
Gözden Geçirme/Revised	:	**29.03.2024**
Kabul/Accepted	:	**13.04.2024**

1. INTRODUCTION

Heart disorders are the leading cause of death worldwide. Heart disease is a term used to describe a condition that leads to the narrowing or blockage of the coronary arteries, potentially causing heart failure, chest pain, or stroke. The World Health Organization (WHO) predicts that in 2019, around 17.9 million individuals will succumb to fatalities caused by heart attacks. Pavan Kumar (2019). The objective of artificial intelligence (AI), a discipline within computer science, is to enhance the intelligence of computers. Intelligence relies on learning, making machine learning (ML) a crucial aspect of artificial intelligence (AI). Machine Learning (ML) is a rapidly advancing field within Artificial Intelligence (AI) that finds application in several domains, particularly in the healthcare sector (Ferdous et al., 2020). The healthcare industry notably benefits from machine learning (ML) due to its intelligent capabilities for information analysis, which is particularly valuable given the abundance of data in the medical sector. In recent years, the digital revolution has led to the capture and storage of vast quantities of data. Modern hospitals have easy access to monitoring and data collection equipment, which are often used and generate large amounts of data. Given the immense challenge that people face in extracting important insights from enormous amounts of information, machine learning is increasingly being utilized to analyze this data and detect problems in the healthcare sector (Bhardwaj et al., 2017). Vivekanandan (2017) was able to forecast heart disease and many other medical conditions using machine learning approaches. Thus facilitating the utilization of efficacious medications and thus saving numerous lives. The fundamental objectives of artificial intelligence and machine-learning algorithms employed in diagnosing cardiac disease are to achieve precise results and detect valuable trends. The initial stages of cardiac disease are often asymptomatic, with heart attack and brain stroke being the earliest manifestations. Heart disease is a concealed threat that claims more lives than cancer in many countries. Diabetes, elevated lipid levels, and increased bloodline weight are all risk factors for cardiovascular disease. Consequently, these variables ultimately cause damage to the heart. An overview of machine learning categorization is provided in this article, along with methods suggested to aid medical practitioners in diagnosing heart disease (Al-Janabi et al., 2018).
Many studies have recently been conducted in the era of artificial intelligence to detect whether a person has heart disease. The results of using machine learning algorithms to forecast heart disease were promising. According to reports, these algorithms are often successful. The Cleveland heart disease dataset was built by Kavitha et al. (2021) Hybrid machine learning models demonstrate efficacy in predicting cardiac disease by amalgamating various methods like decision trees, logistic regression, SVM, random forests, and neural networks. These models utilize the advantages of each method to enhance accuracy and resilience. The procedure entails gathering patient data, doing preprocessing, building the hybrid model, training, assessing, and fine-tuning it to achieve optimal performance. Hybrid models facilitate timely intervention and tailored treatment approaches for patients who are at risk of developing heart disease, but it is essential to validate them using varied datasets. To create and implement systems using Python. Yadav et al. (2020): This study uses machine learning techniques to diagnose cardiac problems by evaluating patient data and discovering patterns for precise prediction. Multiple algorithms, such as decision trees, logistic regression, support vector machines (SVM), random forests, and neural networks, are employed. Early detection and individualized therapy are facilitated by this technique, but ensuring reliability requires thorough validation using varied datasets. Averbuch et al. (2022) Artificial intelligence (AI) and machine learning (ML) are employed in several applications for heart failure in this study. These technologies aid in the early identification, anticipation of risks, tailored therapy, and prediction of outcomes. Artificial intelligence (AI) and machine learning (ML) algorithms examine patient data, including medical records, diagnostic tests, and data from wearable devices, in order to detect patterns and generate precise predictions. These apps possess the capacity to increase patient care, optimize treatment options, and improve overall results in the management of heart failure. Ramesh et al. (2022) in this study employ machine learning techniques to do predictive analysis of cardiac disorders. These methodologies scrutinize extensive volumes of patient data, including medical records, diagnostic examinations, and risk indicators, with the purpose of detecting trends and generating precise forecasts.

By leveraging machine learning algorithms, healthcare professionals can improve early detection, risk assessment, and personalized treatment strategies for individuals at risk of heart disease. This predictive analysis has the potential to enhance patient outcomes and contribute to more effective management of heart diseases. Sajja et al. (2021) The current project involves the utilization of machine learning techniques to categorize and forecast instances of heart disease. Machine learning algorithms can accurately anticipate outcomes by examining patient data, such as medical history and diagnostic testing, and identifying trends. This methodology has the capacity to augment the categorization of cardiovascular illness and boost prognostic models, resulting in superior diagnosis and individualized therapeutic approaches. Dwivedi (2018) assesses the performance of various machine learning algorithms in predicting cardiac disease. Researchers endeavor to determine the most precise and efficient models by comparing and evaluating different algorithms. The evaluation approach enhances diagnosis and risk prediction, empowering healthcare practitioners to make well-informed decisions and deliver superior care for patients suffering from heart disease. Nagavelli et al. (2022) employed four machine learning models to forecast

heart disease, aiming primarily to furnish clinicians with a tool to assist in the timely identification of cardiac problems. Consequently, effectively treating patients while mitigating severe repercussions will be far more manageable. The researchers are doing experiments with several decision tree classification algorithms, specifically XGBoost, in order to enhance the accuracy of diagnosing heart disease. Ali et al. (2021) This study uses supervised machine learning algorithms to predict cardiac disease and evaluates and compares their performance. Researchers endeavor to assess the efficacy of various algorithms in properly predicting cardiac disease. This study aids in identifying the most dependable and precise models for early identification and enhanced patient care. Ping Li et al. (2020) This study utilizes machine learning classification techniques in the field of e-healthcare to detect and diagnose cardiac problems. These techniques employ patient data and employ machine learning algorithms to categorize individuals as either having or not having cardiac disease. Through the utilization of this strategy, healthcare providers can optimize the detection of heart disease, resulting in prompt interventions and enhanced outcomes in e-healthcare environments. Tougui and Mhamdi (2020): This study uses data mining tools and machine learning approaches to classify cardiac disease. Through the examination of patient data, these methods detect trends and employ machine learning algorithms to categorize individuals into distinct heart disease groups. This approach improves the precision of cardiac disease categorization and assists in tailoring treatment regimens, thereby enhancing patient care. As contribution, the study will gather a large database encompassing various patient information, including demographic data, lifestyle habits, medical history, and numerous health markers. This dataset will provide an effective foundation for training and evaluating machine learning methods. The study will offer a performance comparison of various machine learning algorithms, such as Naïve Bayes, K-Nearest Neighbor (K-NN), and Logistic Regression. This comparison will offer an understanding of the advantages and disadvantages of each algorithm in the specific context of predicting heart disease. The study will utilize multiple metrics to evaluate the predictive accuracy of each algorithm including accuracy, precision, recall, and F1 score.. This review will provide a comprehensive assessment of the accuracy of each algorithm in predicting heart disease. The study will discover the primary determinants of heart disease, as indicated by machine learning algorithms. This information will assist healthcare practitioners in prioritizing the most crucial risk factors. The study will examine the practical effects of applying machine learning for the prediction of heart disease, including possible advantages such as timely identification and intervention as well as difficulties such as protecting data privacy and requiring additional validation as will see in section 3. The study will propose potential areas for future investigation, including the exploration of improved machine learning methods, the integration of different forms of data (such as genetic data), and the implementation of multi-center studies to improve the applicability of the results. The contributions will be as follows:

- Development of a comprehensive dataset for heart disease prediction.
- Comparison of the performance of multiple machine learning algorithms.
- Evaluation of predictive performance using various metrics.
- Identification of key predictors of heart disease.
- Discussion of the practical implications of using machine learning for heart disease prediction.
- Suggestion of directions for future research in this area.

The abstract gives a quick summary of the paper, covering what the study is about, how it was done, what was found, and what conclusions were drawn. The introduction section explains why predicting heart disease accurately matters and how machine learning fits into healthcare. It also outlines what the paper aims to achieve. Next, the literature review looks at what other studies have done in predicting heart disease with machine learning. The methodology section explains how the study was conducted, including the data used, how it was prepared, and which machine learning methods were tested. It also describes how these methods were evaluated. In the results and discussion section, the findings from testing different machine learning methods are presented with tables and graphs showing how well they performed. The discussion then analyzes these results, comparing the strengths and weaknesses of each method and discussing what they mean for predicting heart disease and healthcare. Finally, the conclusion summarizes the main points of the study and suggests areas for future research. The references section lists all the sources used in the paper.

2. MATERIAL AND METHOD

2.1. KNIME Platform

The open-source reporting and integration platform for data analytics is called KNIME, or Konstanz Information Miner. The Silicon Valley Software Company and the University of Konstanz jointly created it. With its modular data pipelining architecture, KNIME integrates multiple components for data mining and machine learning. The arrangement of nodes for modelling, data analysis, data visualization, and ETL (extraction, transformation, loading) of data is made possible through a graphical user interface. It is created in Java using Eclipse as a basis.

KNIME has been utilized in medicinal research since 2006 (Bernd Wiswedel, 2009). The open-source program KNIME tries to address these issues by offering a platform that may quickly be expanded. has newly integrated tools and a tightly typed data structure, enabling workflow authors to meticulously describe the workflow's moves. Old nodes are also deprecated in KNIME, which means that even after many years, using workflows developed with prior versions still produces the same results. i.e., Figure 1 shows the KNIME user interface (Berthold et al., 2009).

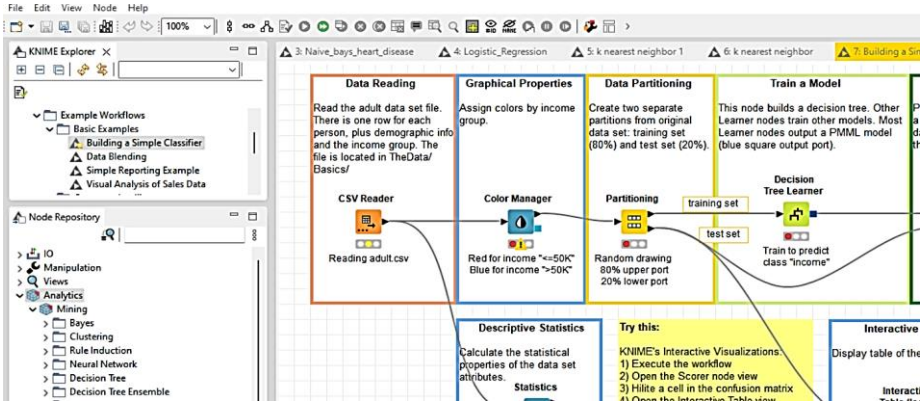

Figure 1: Interface of KNIME

KNIME's quantity of data analysis and machine learning nodes is one of its advantages. Although its default configuration already includes a wide range of algorithms for this purpose, the plug-in system is what permits outside developers to quickly put together their products and make them match with one another's production. Tool integrations are of particular importance to users in the data sciences (Fillbrunn et al., 2017).

2.2. Naïve Bayes

Naive Bayes is a machine learning method that relies on Bayesian formula-based probability models and is generally considered to be straightforward. Despite its simplicity, Naive Bayes often surpasses more intricate classification approaches. The basis of this algorithm is conditional probability. The approach relies on a probability table as its model, which is then updated using training data. The "probability table" derives its class probabilities for predicting a new observation based on the values of its features. The term "naive" is used to describe it due to its core assumption of conditional independence (Ahmed et al., 2023). Figure 2 explains the principle of Naïve Bayes.

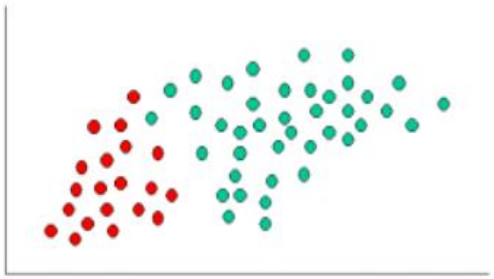

Figure 2: Present the Naive Bayes principle

Consider the graphic above as a visual representation of the concept of Naive Bayes classification. The objects can be classified as either red or green, as demonstrated. It is incumbent upon us to categorize incoming situations by ascertaining their appropriate class label, considering the existing items. Given the ratio of twice as many green items to red objects, it is plausible to infer that an undetected new example is more likely to belong to the green group. The assumption mentioned is commonly known as the prior probability in Bayesian analysis. Prior probabilities are commonly utilized to anticipate outcomes based on previous experience, specifically by considering the relative proportions of green and red objects. The naive Bayes algorithm employs a probabilistic approach to determine the most likely class for a given set of data by considering the judgments of many classes. The naive Bayes algorithm employs fictional probabilities in a deliberate manner to perform its calculations. The Bayes Theorem can be expressed using the following formula: (Yang, 2018).

$$P(Q|X) = \frac{P(X|Q) \cdot P(Q)}{P(X)} \quad (1)$$

$$P(Q|x) = P(x1|Q) \times P(x2|Q) \times \ldots \times P(xn|Q) \times P(X) \quad (2)$$

With details as i.e. Table 1:

Table 1: Definition of variables

ITEM	NEEDED
X	Data with an unclassified class
Q	The assumption *X* is a particular class
P(Q|X)	The probability of the Q assumption refers to *X*
P(Q)	Probability of the assumption Q (prior probability)
P(X|Q)	Probability *X* in the assumption Q
P(X)	Probability *X*

Patil (2013) This classification technique analyzes the relationship between each feature and the class for every occurrence. It calculates a conditional probability to determine the correlations between the feature values and the class. This text offers a comprehensive examination of machine learning categorization. The Naive Bayes algorithm utilizes the joint probabilities of features and classes to assess the probability of a document belonging to a certain class.

2.3. K-Nearest Neighbor (K-NN)

The majority of classification issues employ the supervised machine learning technique known as k-nearest neighbor (KNN). The utilization of this technique in forecasting illnesses has a long-standing historical background. The KNN algorithm which is a supervised learning method, utilizes the labels and properties of the training data to make predictions about the categorization of unlabeled data. The KNN classifier, a case-based machine learning approach, is employed to automatically classify or categorize textual data. The KNN classifier is built upon the Euclidean distance, which is used to measure the similarity between texts and the k training data (Rajeswari et al., 2017). The equation provided calculates the Euclidean distance, denoted as d (x, y), between two points x and y.

$$d(x|y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad (3)$$

The K-NN approach aims to categorize a given sample data point as a classification problem by utilizing a dataset consisting of data points arranged into multiple classes (Uddin et al., 2022). Mahesh (2020) says the K-nearest neighbors (KNN) approach generally utilizes the k nearest training data points. Which datasets are the most similar to the testing query when using a training model that corresponds to the testing query for classification? The category is determined by applying a majority selection rule. The KNN technique is widely recognized and extensively used for classification problems due to its very versatile and straightforward design. To explain the KNN algorithm in visual form, i.e., figure 3.

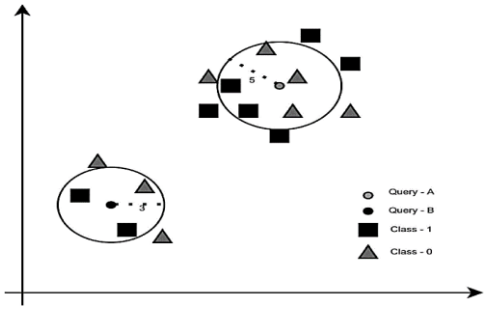

Figure 3 : Scenario of the KNN algorithm in visual form.

An object is provided with a class that includes its k-nearest neighbors. The K-NN algorithm categorizes a vector by utilizing the classes of its k-nearest neighbors in the new test feature, as seen in figure 4. (Medjahed et al., 2013).

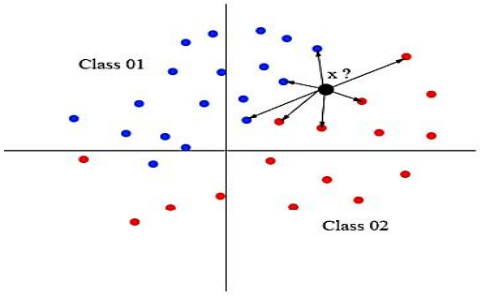

Figure 4: The K-nearest neighbors' method.

2.4. Logistic Regression

The logistic regression approach in supervised machine learning is used for binary classification tasks by estimating the probability of an action, occurrence, or observation. Logistic regression (LR) is the primary statistical and data mining technique employed by mathematicians and scientists to analyze and classify binary and relative response datasets (Haziemeh et al., 2023). Ferdous et al. (2020) Logistic regression, like the Naïve Bayes model, derives a set of weighted features from the input, transforms them into logarithmic values, and then combines them in a linear manner. The technique entails the multiplication of each feature by its corresponding weight, followed by the summation of the results. The primary differentiation between naive Bayes and logistic regression lies in the fact that logistic regression employs a naive Bayes approach, whereas the generative classifier utilizes a discriminative approach.

For logistic regression, the sigmoid function is known as an activation function and is described as follows:

$$f(x) = \frac{1}{1+ e^{-x}} \tag{4}$$

as,

e = natural logarithms' base

x = numerical value one wishes to transform

Data preparation for the logistic model required:

Output Binary Variable, eliminate noise, Remove Correlated Inputs from the Gaussian Distribution Fails to Converge (Rymarczyk et al., 2019).

2.5. Dataset

The CDC's dataset encompasses the majority of racial groups in the US, including white individuals, American Indians, Alaska Natives, and African Americans. It can be accessed via a Kaggle online repository. The initial dataset, comprising approximately 300 variables, was reduced to 319,796 patient records, retaining only about 18 variables. This dataset is versatile and can be utilized for a wide range of machine learning methodologies (PYTLAK, 2020). Preprocessing of the dataset is necessary to address potential errors, missing data, redundancy, noise, and other factors that may hinder the accurate utilization of the machine learning algorithm. Additional preparatory processes for the dataset may vary depending on its organization and can encompass data cleansing, transformation, imputation of missing values, normalization, feature selection, and other approaches (García et al., 2016). The datasets share similar features for heart disease. The number of attributes for each sample is 18. These characteristics are BMI, smoking, drinking alcohol, stroke, physical health, mental health, difficulty walking, sex, age category, race, diabetes, physical activity, general health, sleep duration, asthma, kidney disease, skin cancer, and heart disease. i.e. Table 2 contains details about the characteristics and sample of the dataset explained in, i.e., Table 3.

Table 2 : Fields in the dataset and their descriptions.

Attribute name	Description
BMI	Body Mass Index (had) skin cancer? Yes or No
Smoking	Smoking cigarettes or not
Drinking Alcohol	Drinking alcohol or not
Stroke	(had) a stroke or not
Physical Health	Health physically for how many days over the last 30 days, considering illnesses and injury.
Mental Health	How many days out of the last 30 have you experienced poor mental health? (0-30 days)
Difficult Walking	Have serious difficulty walking or climbing stairs
Sex	Male or female
Age	Age category
Race	Description the race of person (white, Hispanic, black ...)
Diabetic	Had diabetes? Yes or No
Physical Activity	Participated in physical activity or exercise outside of their normal employment in the last 30 days? (Yes or No)
General Health	General your health is (good, very good, fair)
Sleep Time	How much sleep do you get each night in a 24-hour period?
Asthma	Had asthma? (Yes or No)
Kidney Disease	had kidney disease (Yes or No)
Skin Cancer	(had) skin cancer? Yes or No
heart disease	Persons who have a myocardial infarction (MI) or heart disease (HD)

Table 3: Sample of dataset used for the research.

Heart Disease	BMI	Smoking	Alcohol Drinking	Stroke
No	16.6	Yes	No	No
No	20.34	No	No	Yes
No	26.58	Yes	No	No
No	24.21	No	No	No
No	23.71	No	No	No

2.6. Evaluating Model Performance

Researchers use metrics to evaluate prediction models and present the results of their performance. To show the efficacy and reliability of the test, the sensitivity, specificity, and accuracy of statistical metrics are used to evaluate the efficacy of the suggested technique. All of the research studies reviewed in our article employ accuracy as their primary performance evaluation parameter (Gupta et al., 2013).

2.6.1.Confusion Matrix

Hossin (2015) A predictive analysis tool can be described as a confusion matrix within the field of machine learning. The evaluation of a classification-based machine learning model's performance is conducted using the confusion matrix. The confusion matrix is a table that summarizes the number of correct and incorrect predictions made by a classifier or classification model for binary classification tasks. By visualizing the confusion matrix, i.e., figure. 5 illustrates the confusion matrix.

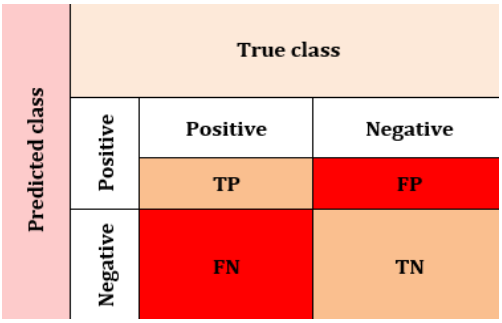

Figure 5: Confusion matrix. **(Hossin, 2015)**

TP (True Positive): It indicates that the model predicted a positive result, and the true value was indeed positive.

TN (True Negative): It represents the model displaying a negative value when the true value was negative.

FP (False Positive): This occurs when the model predicts a positive result, but it is incorrect or false.

FN (false negative): This refers to the model predicting a negative result, but it is incorrect or false. (Hossin, 2015)

Confusion matrix contains a lot of ways to calculate accuracy as following:

A. Accuracy

In general, the accuracy metric calculates the percentage of accurate predictions for all data that were considered.

$$\text{acc} = \frac{\text{TP+ TN}}{\text{TP+ FP+ TN+ FN}} \quad (5)$$

Accuracy is a metric that quantifies the proportion of correct predictions generated by your model on the complete test dataset. Accuracy is a fundamental metric that serves as a reliable measure to evaluate the performance of the model. Unbalanced datasets render accuracy an inadequate metric.

The accuracy score may not provide an accurate representation of a model's performance, and it is not the sole statistic used to evaluate a model's performance. In such scenarios, it is crucial to take into account other evaluation metrics, including precision, recall, F-score, and ROC curve.

B. Precision

Precision is a measure that indicates the proportion of accurately predicted cases that did not result . in favorable outcomes. This would confirm the reliability of our model. When the occurrence of a false positive is more troublesome than that of a false negative, accuracy serves as a useful signal (Ma J et al., 2019).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+TN}} \quad (6)$$

C. Recall

The percentage of actual positive cases that our model properly predicted is known as recall. The formula that explains recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (7)$$

The recall is higher, indicating that a significant proportion of the positive instances (true positives and false negatives) would be correctly identified as positive (true positives). This will lead to an increase in the number of FP measurements being conducted and a decrease in overall accuracy. The recall rate is low, indicating that a significant proportion of false negatives occurred when instances that should have been classified as positive were instead labeled as negative. This suggests that in the event of identifying a positive example, one can have a higher level of certainty that it is indeed a genuine positive (Vakili et al., 2020).

D. F1-score

The f-score, also known as the f-measure, is a metric that evaluates the performance of an algorithm by taking into account both precision and recall. The mathematical representation of recall and precision is based on the principles of harmonious techniques (Kabir et al., 2023).

$$\mathrm{F-score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{8}$$

E. Specificity

Specificity refers to the proportion of genuine negatives that the model correctly identifies. This suggests that there will be a specific proportion of accurate negative predictions that will be classified as positive and can be labeled as false positives (Banaei et al., 2019).

$$\mathrm{Specificity} = \frac{\mathrm{TP}}{\mathrm{P}} \tag{9}$$

F. Sensitivity

Sensitivity, also known as the true positive rate, is determined by the proportion of correctly diagnosed heart disease cases among all positive predictions made by the models (Hand, 2007).

$$\mathrm{Sensitivity} = \frac{\mathrm{TN}}{\mathrm{N}} \tag{10}$$

3. RESULT AND DISCUSSION

Comparing three machine learning techniques: Naive Bayes, K-NN, and Logistic Regression. The requisite data was initially uploaded to the KNIME environment. To access the data, which is in "CSV" format, navigate to the "IO" section in the KNIME environment and select "Read." From there, choose "CSV Reader." Upon perusing the input, In order to apply the Naive Bayes and K-NN algorithms to the analysis, it is necessary to divide the data into separate parts. Partitioning demonstrates the act of dividing or separating something into smaller parts. The dataset is partitioned into two segments: the training data and the test data. The study employed an 80% train data and 20% test data split. The overall dataset consisted of 319,795 samples, with 255,836 samples allocated for training and 63,959 samples for testing we also used . The training set evaluates the model's ability to account for the data in the target variable, while the test set assesses the model's performance on new, unseen observations. During the modeling phase, Naive Bayes learning and prediction nodes are incorporated into the model. In this case, a scorer is utilized to provide the final result. The information can be observed in the provided illustration, specifically in Figure 6.

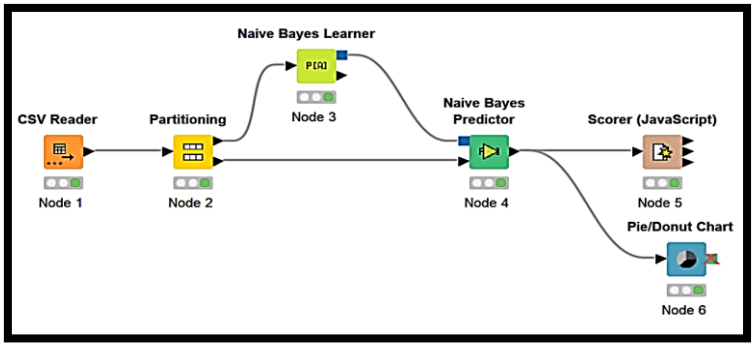

Figure 6 : Utilizing naive bayes to classify the heart disease dataset.

As a result, the overall accuracy was %86,63 with the confusion matrix, i.e., figure 7.

Scorer View

Confusion Matrix

Rows Number : 63959	No (Predicted)	Yes (Predicted)	
No (Actual)	53278	5206	91.10%
Yes (Actual)	3347	2128	38.87%
	94.09%	29.02%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
86.63%	13.37%	0.260	55406	8553

Figure 7: Confusion matrix for Naive bayes method.

Also, we can see the performance evaluation calculations from the KNIME environment, i.e., Table 4.

Table 4: Performance evaluation for naive bayes algorithm

	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure
Yes	2128	5206	53278	3347	0.389	0.29	0.389	0.911	0.332
No	53278	3347	2128	5206	0.911	0.941	0.911	0.389	0.926

When using the K-NN method, learning and prediction nodes are added to the model throughout the modeling phase; here, the scorer is the product result. We can view that, i.e., figure 8.

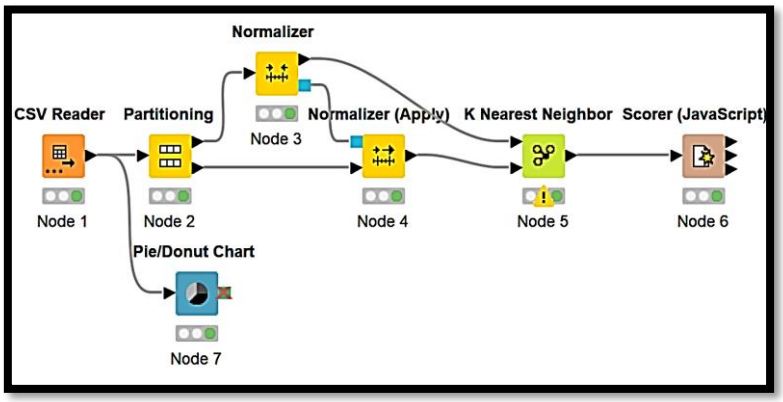

Figure 8: Heart disease dataset classification using the K Nearest Neighbor (K-NN) algorithm.

To compute the accuracy overall, it was %90,77 with the confusion matrix as a figure 9.

K Nearest Neighbor

Confusion Matrix

	No (Predicted)	Yes (Predicted)	
No (Actual)	57919	501	99.14%
Yes (Actual)	5402	137	2.47%
	91.47%	21.47%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
90.77%	9.23%	0.027	58056	5903

Figure 9: Confusion matrix for K - Nearest Neighbor

Also, we can see the performance evaluation calculations from the KNIME environment in Table 5.

Table 5: Performance evaluation for K-NN algorithm

	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure
Yes	57919	5402	137	501	0.991	0.915	0.991	0.025	0.952
No	137	501	57919	5402	0.025	0215	0.025	0.991	0.044

Finally, we used the logistic regression method for predicting heart diseases. Instead of an 80-20 split, the data is partitioned into 10-fold and 5-fold cross-validation segments using the "Partitioning" node. This ensures more robust evaluation by repeatedly splitting the dataset into training and testing subsets. Each fold represents a distinct combination of training and testing data. Learning and prediction nodes are added to the model throughout the modeling phase; here, we used a scorer for the product result. We can view that, i.e., figure 10.

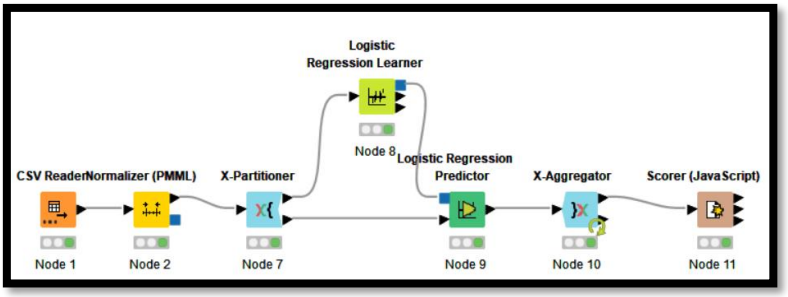

Figure 10: Classification of the Heart Disease data set using Logistic Regression

This method achieves %91,43 accuracy with the confusion matrix, i.e., figure 11.

Scorer View

Confusion Matrix

Rows Number : 319795	No (Predicted)	Yes (Predicted)	
No (Actual)	292374	48	99.98%
Yes (Actual)	27357	16	0.06%
	91.44%	25.00%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
91.43%	8.57%	0.001	292390	27405

Figure 11: Confusion matrix for Logistic Regression

Also, we can see the performance evaluation calculations from the KNIME environment, i.e., Table 6.

Table 6 : Performance evaluation for Logistic Regression algorithm

	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure
Yes	292374	27357	16	48	1	0.914	1	0.001	0.955
No	16	48	292374	27357	0.001	0.25	0.001	1	0.001

4. CONCLUSION

Heart disease is an important concern in our modern world. Therefore, there is a requirement for an automated system capable of forecasting cardiovascular illness in its early stages. Feature selection and prediction are crucial components of any automated system. Enhancing our ability to forecast cardiac disease can be achieved by carefully choosing relevant features. Within this research endeavor, we formulated three distinct methodologies for doing comparative analysis, which yielded advantageous outcomes. Machine learning techniques outperformed other methods in this analysis. The evaluation metrics include the confusion matrix, accuracy, specificity, sensitivity, and F1 score. Our objective is to enhance the precision and efficiency of forecasts by reducing the number of features and tests required. A dataset consisting of 319,795 data entries is utilized for machine learning (ML) methodologies. The algorithms are compared in Table 7.

Table 7: Result Comparison

Algorithms	Partition	Accuracy
Naïve Bayes	80 train, 20 test	86,63%
K-Nearest Neighbors	80 train, 20 test	90,77%
Logistic Regression	5-fold cross validation	91,43%

For this study, we employed three different machine-learning algorithms to forecast the occurrence of cardiac disease. Out of all the algorithms, the Logistic Regression Algorithm demonstrated the greatest accuracy rate of %91,43 in predicting heart disease. The findings demonstrate the effectiveness of machine learning algorithms in detecting heart disease and estimating the likelihood of an individual's impact. This can aid physicians in future studies by enabling them to make well-informed decisions regarding the required level of treatment intensity for patients. In the future, the prediction of heart disease using various algorithms and diverse variables holds promise for doctors and heart surgeons. The findings suggest that this approach can be beneficial for determining the likelihood of a heart attack in patients. Furthermore, the study demonstrates that a relatively simple supervised machine learning method can accurately predict heart disease, indicating its potential utility in clinical practice.

REFERENCES

Ahmed, S., Singh, M., Doherty, B., Ramlan, E., Harkin, K., Bucholc, M., & Coyle, D. (2023). An empirical analysis of state-of-art classification models in an it incident severity prediction framework. Applied Sciences, 13(6), 3843.

Alexander Fillbrunn, Christian Dietz a, Julianus Pfeuffer, René Rahn, Gregory A. Landrum, Michael R. Berthold . (2017). KNIME for reproducible cross-domain analysis of life science data. Journal of Biotechnology, pp. 1-8.

Ashok Kumar Dwivedi. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput & Applic 29, 685–693.

Banaei, N., Moshfegh, J., Mohseni-Kabir, A., Houghton, J. M., Sun, Y., & Kim, B. (2019). Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips. RSC advances, 9(4), 1859-1868.

Bernd Wiswedel, M. B. (2009). knime. (software) Retrieved from https://www.knime.com/.

Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017, July). A study of machine learning in healthcare. In 2017 IEEE 41st annual computer software and applications conference (COMPSAC) (Vol. 2, pp. 236-241). IEEE.

Dr. M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj. (2021). Heart Disease Prediction using Hybrid machine Learning Model. Coimbatore, India: 2021 6th International Conference on Inventive Computation Technologies (ICICT).

F. -J. Yang. (2018). An Implementation of Naive Bayes Classifier. International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 301-306.

Ferdous, M., Debnath, J., & Chakraborty, N. R. (2020, July). Machine learning algorithms in healthcare: A literature survey. In 2020 11th International conference on computing, communication and networking technologies (ICCCNT) (pp. 1-6). IEEE.

G. S. Sajja, M. Mustafa, K. Phasinam, K. Kaliyaperumal, R. J. M. Ventayen and T. Kassanuk, (2021). Towards Application of Machine Learning in Classification and Prediction of Heart Disease. 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 1664-1669.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), pp. 1-22.

Hand, D. J. (2007). Principles of Data Mining. Drug Safety, pp. 1-30.

Haziemeh, F.A., Darawsheh, S.R., Alshurideh, M., Al-Shaar, A.S. (2023). Using Logistic Regression Approach to Predicating Breast Cancer DATASET. The Effect of Information Technology on Business and Marketing Intelligence Systems, pp. 1-10.

Hossain, M. a. (2015). A review of evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP), pp. 1-11.

J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access, 107562-107582.

M. Ferdous, J. Debnath and N. R. Chakraborty. (2020). Machine Learning Algorithms in Healthcare: A Literature Survey. 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. (1-6).

M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel. (2009). KNIME: the Konstanz information miner. ACM SIGKDD Explorations Newsletter, 6 pages.

Ma, J., Ding, Y., Cheng, J. C., Tan, Y., Gan, V. J., & Zhang, J. (2019). Analyzing the leading causes of traffic fatalities using XGBoost and grid-based analysis: a city management perspective. IEEE Access, 7, 148059-148072.

Mahesh, B. (2020). Machine learning algorithms—a review. Int. J. Sci., 5.

Maryam I. Al-Janabi, , Mahmoud H. Qutqut and , Mohammad Hijjawi. (2018). Machine Learning Classification Techniques for Heart Disease Prediction: A Review. International Journal of Engineering & Technology, 7 (4) (2018) 5373-5379.

Md Faisal Kabir, Tianjie Chen, Simone A. Ludwig. (2023). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthcare Analytics, 9 pages.

Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui , Julian M.W. Quinn , Mohammad Ali Moni .(2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 10 pages.

Medjahed, S. A., Saadi, T. A., & Benyettou, A. (2013). Breast cancer diagnosis by using k-nearest neighbor with different distances and classification rules. International Journal of Computer Applications, 62(1).

Meysam Vakili, Mohammad Ghamsari and Masoumeh Rezaei. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. 13 pages.

Niyati Gupta, Arushi Rawal, Dr. V.L. Narasimhan, Savita Shiwani. (2013). Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data. IOSR Journal of Computer Engineering (IOSR-JCE), pp 70-73.

Patil, T. R. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. Int. J. Comput. Sci. Appl., 6.

Pavan Kumar T and Avinash Golande. (2019). Heart Disease Prediction Using Efficient Machine Learning Methods. International Journal of Current Technology, 70.

PYTLAK, K. (2020). kaggle. (Kaggle) Retrieved from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?select=heart_2020_cleaned.csv.

Rajeswari R. P, Kavitha Juliet, Dr. Aradhana. (2017). Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier. International Journal of Computer Trends and Technology (IJCTT), pp. 1-5.

Ramesh TR, Umesh Kumar Lilhore, Poongodi M, Sarita Simaiya, Amandeep Kaur and Mounir Hamdi. (2022). predictive analysis of heart diseases with machine learning approaches. Malaysian Journal of Computer Science, 132–148.

Rymarczyk, T., Kozłowski, E., Kłosowski, G., & Niderla, K. (2019). Logistic regression for machine learning in process tomography. Sensors, 19(15), 3400.

Samir S Yadav; Shivajirao M. Jadhav; Snigdha Nagrale; Niraj Patil. (2020). Application of Machine Learning for the Detection of Heart Disease. 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 165-172.

T. Vivekanandan, N. Ch Sriman Narayana Iyengar. (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, Computers in Biology and Medicine, pp 125-136.

Tauben Averbuch, Kristen Sullivan, Andrew Sauer, Mamas A Mamas, Adriaan A. Voors, Chris P. Gale, Marco Metra, Neal Ravindra and Harriette G.C. Van Spall. (2022). Applications of artificial intelligence and machine learning in heart failure. European Heart Journal - Digital Health, 311-322.

Tougui, I., Jilbab, A. & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health Technol, 1137–1144.

Uddin, S., Haque, I., Lu, H. et al. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep 12, 6256.

Umarani Nagavelli, Debabrata Samanta and Partha Chakraborty. (2022). Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering, 9 pages.