Comparison of Computational Efficiency of Various Machine Learning Algorithms in Heart Disease Diagnosis.

1.J.Sahithi(21VV1A1229)
2.K.Himabindu(21VV1A1230)
3.K.Chandra Sai Shankar(21VV1A1231)
4.K.Bhuvaneswary(21VV1A1232)

Author Affilliation
Jawharlal Nehru Technological University Gurajada, Vizianagaram
Dwarapudi, Viziznagaram, Andhra Pradesh 523003.

Author's Emails
sahithijakkilinki@gmail.com
himabindukandivalasa@gmail.com
chandukarra03@gmail.com
bhuvana190503@gmail.com

1 Abstract

This project focuses on predicting heart disease using a comparative analysis of various machine learning algorithms. By examining models such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Random Forest, the project aims to identify the most effective algorithm for detecting heart disease risk based on patient data. Key features, including age, cholesterol levels, blood pressure, chest pain type, and more, are used to build predictive models that can assist in early diagnosis. Each algorithm is evaluated on metrics such as accuracy, interpretability, and computational efficiency, with Random Forest emerging as a strong candidate for its high accuracy and ability to manage complex, non-linear relationships in healthcare data.

2 Introduction

Heart disease remains one of the leading causes of death worldwide, making early and accurate diagnosis a critical area of medical research. Advances in machine learning have opened new possibilities for predicting and diagnosing heart disease based on vast datasets of patient health information. However, while accuracy is often a focal point, the computational efficiency of these machine learning algorithms is equally vital, especially when considering large datasets or real-time applications in healthcare.

Different machine learning algorithms come with varying levels of complexity and computational demands. For instance, simpler algorithms like Logistic Regression or Decision Trees might offer faster processing times but may sacrifice predictive power in certain cases. On the other hand, more complex models such as Neural Networks and Ensemble methods (e.g., Random Forest, XGBoost) tend to provide higher accuracy at the cost of greater computational resources.

3 Need of Heart Disease Detection

Heart disease remains one of the most prevalent and deadly health conditions worldwide, contributing significantly to global mortality rates. Detecting heart disease early is crucial for effective treatment and can dramatically improve patient outcomes. However, many heart conditions remain undiagnosed until they reach a severe stage due to vague symptoms, lack of regular health monitoring, or limited access to healthcare. Early detection methods enable healthcare providers to identify at-risk individuals sooner, allowing for timely intervention that can prevent complications or fatalities.

Advanced heart disease detection is essential because heart conditions can progress rapidly if untreated, leading to life-threatening events such as heart attacks or strokes. Early-stage heart disease often presents with symptoms that are subtle and easy to overlook, making accurate diagnosis challenging. In many cases, patients ignore early warning signs, attributing them to less serious conditions. Effective detection systems can highlight potential risks before these symptoms escalate, giving patients a better chance of managing their condition through lifestyle changes, medications, or other medical interventions.

Recent technological advancements, particularly in the field of machine learning and artificial intelligence, have improved our ability to detect heart disease early. Machine learning models can analyze complex datasets and identify patterns that traditional methods might miss, helping to diagnose various forms of heart disease with higher precision. These models are especially useful in analyzing risk factors such as age, cholesterol levels, blood pressure, and lifestyle habits. By leveraging such tools, healthcare providers can predict which patients are at higher risk and monitor them more closely, potentially reducing the overall burden of heart disease.

Finally, widespread adoption of heart disease detection technologies has the potential to reduce healthcare costs significantly. Detecting heart disease early and managing it proactively is far less expensive than treating advanced-stage conditions, which often require intensive medical care, surgery, and long hospital stays. By identifying high-risk patients early, health systems can allocate resources more efficiently, reduce emergency interventions, and lower the strain on healthcare facilities. This proactive approach not only benefits patients but also helps in creating a sustainable healthcare model focused on prevention rather than crisis management.

Heart disease detection empowers individuals to take charge of their health by identifying risks before severe symptoms arise. It enables healthcare providers to offer personalized treatment plans, improving the quality of life and longevity for many. Additionally, early detection contributes to medical research, helping scientists better understand and combat heart disease at a population level.

4 Dataset Description

The dataset contains several essential features that are key indicators in assessing heart disease risk. The age and sex columns represent the patient's demographic data, with age providing context on the likelihood of heart disease as it increases with age, and sex indicating the biological differences in heart disease prevalence and severity. Research shows that older individuals, particularly males, tend to have a higher risk of heart disease, making these variables foundational for analysis and prediction.

The dataset also includes medical indicators commonly used to evaluate cardiovascular health. Cp (chest pain type) is a crucial symptom indicator, as different types of chest pain are associated with varying heart conditions. Trestbps (resting blood pressure) and chol (cholesterol levels) are vital measurements, as high blood pressure and cholesterol are known risk factors for heart disease. Fbs (fasting blood sugar) adds further insight, helping to identify any underlying diabetes, which can significantly impact cardiovascular health.

Electrocardiogram results are also represented with restecg (resting electrocardiographic results) and thalach (maximum heart rate achieved). These attributes provide data on heart rhythm and heart rate behavior under stress, which are valuable in identifying abnormalities that may indicate heart disease. Exang (exercise-induced angina) is a binary variable indicating whether chest pain occurs during exercise, often signaling heart blockages. Additionally, oldpeak records ST depression levels during exercise, helping to identify ischemic conditions in the heart muscle, which could suggest a higher risk of heart disease.

The last set of features includes slope (the slope of the ST segment), ca (number of major vessels colored by fluoroscopy), and thal (a type of blood disorder). These features are used to evaluate the severity of heart conditions further. Finally, target is the outcome variable indicating the presence or absence of heart disease, making it the central focus for predictive analysis. Each of these features provides critical insights, collectively creating a comprehensive dataset for developing models that can help predict heart disease risk.

This dataset comprises 14 columns and 1026 rows, each row representing a unique patient's data. The 14 columns, as described, include both demographic and clinical variables relevant to assessing heart disease. With features such as age, sex, chest pain type, blood pressure, cholesterol levels, and various test results, the dataset offers a well-rounded view of the typical clinical markers used to evaluate heart health. .

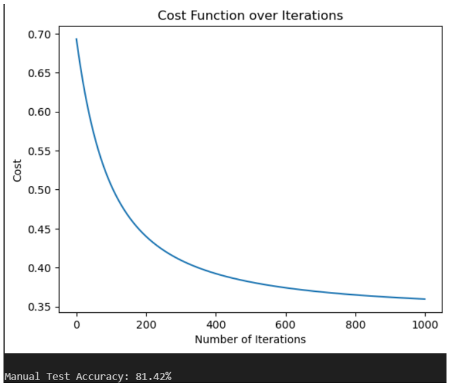
The 1026 rows provide a substantial amount of data, allowing for a robust statistical analysis and reliable model training. With this dataset size, there is enough data to ensure accurate representation of various patterns and relationships within the features, which helps in building models that generalize well to new data.

5 Algorithms Used

Programming for heart disease detection involves building machine learning models that can analyze patient data and predict the likelihood of heart disease. By leveraging a dataset of key health indicators—such as age, cholesterol, blood pressure, chest pain type, and others—programmers can preprocess the data, apply feature engineering, and use algorithms like Logistic Regression, Decision Trees, and Neural Networks to develop predictive models. These models are trained and validated on historical data to learn patterns associated with heart disease risk.

5.1 Logistic Regression

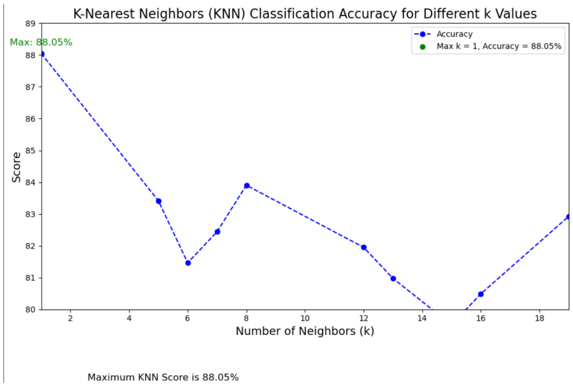
Logistic regression is a supervised machine learning algorithm used for binary classification tasks, such as predicting the presence or absence of heart disease. It estimates the probability of a target variable by applying a logistic function to a linear combination of input features. Unlike linear regression, logistic regression maps the output to a range between 0 and 1, making it ideal for predicting probabilities.



```
Manual Test Accuracy: 81.42%
```

5.2 K Nearest Neighbor Classifier

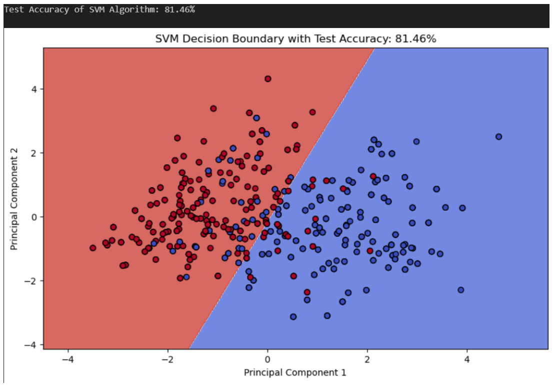
The K-Nearest Neighbor (KNN) classifier is a simple, non-parametric algorithm used for classification tasks. It predicts the class of a data point based on the majority class among its K nearest neighbors in the feature space. KNN is highly intuitive, making predictions based on proximity, where ”closeness” is measured by distance metrics such as Euclidean distance.



```
Maximum KNN Score is 88.05%
```

5.3 Support Vector Machines

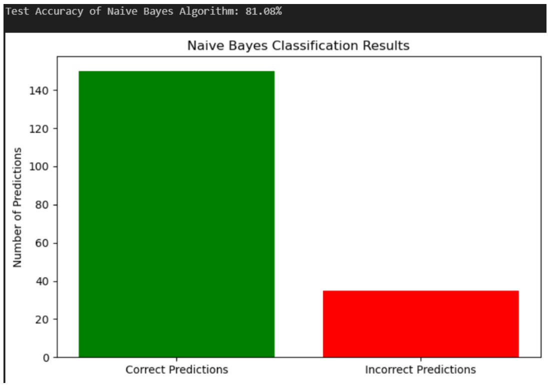
Support Vector Machine (SVM) is a powerful supervised machine learning algorithm commonly used for classification tasks. It works by finding the optimal hyperplane that best separates data points into different classes, maximizing the margin between them. SVM is particularly effective in high-dimensional spaces and with data that isn't linearly separable, where it uses kernel functions to transform data for better classification. Known for its accuracy and robustness, SVM is widely used in applications like image recognition, text classification, and medical diagnosis.



```
Test Accuracy of SVM Algorithm: 81.46%
```

5.4 Gaussian Naive Bayes

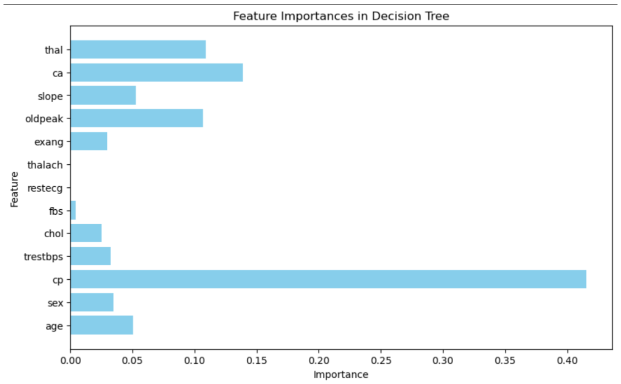
Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence among features and that data follows a Gaussian (normal) distribution. It calculates the probability of each class for a given set of features, selecting the class with the highest probability. This algorithm is particularly efficient with large datasets and works well for continuous data, making it suitable for applications like medical diagnosis.



```
Test Accuracy of Naive Bayes Algorithm: 81.08%
```

5.5 Decision Tree Classifier

A Decision Tree Classifier is a machine learning algorithm that splits data into branches based on feature values to classify outcomes. It uses a tree-like structure where each node represents a decision based on a feature, leading to further splits until reaching a leaf node that represents the final class label. Decision Trees are popular for their interpretability, as the branching paths make the classification logic transparent and easy to follow.



5.6 Random Forest Classifer

The Random Forest Classifier is an ensemble learning algorithm that builds multiple decision trees and merges their outputs to make accurate predictions. Each tree is trained on a different subset of data, adding randomness that improves the model's robustness and reduces overfitting. It is particularly effective for classification tasks, handling high-dimensional data well, and can rank feature importance for further insights. Random Forest is popular for its high accuracy, stability, and ability to handle both structured and unstructured data.

6 Comparison

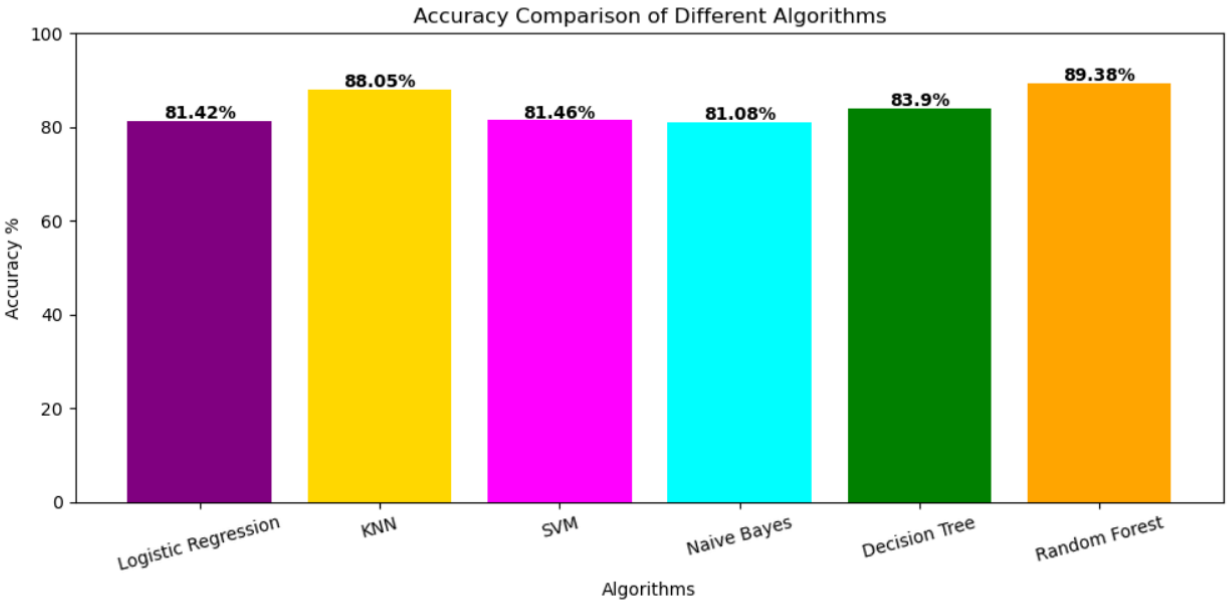
Logistic Regression is a statistical method often used as a baseline model due to its simplicity and interpretability. For heart disease detection, Logistic Regression is useful in determining the probability of a patient being at risk based on binary and numerical input features. It performs well with linear relationships, making it easy to interpret the influence of each feature on the target. However, its performance may be limited if the data is complex or has non-linear relationships, as is often the case in medical datasets.

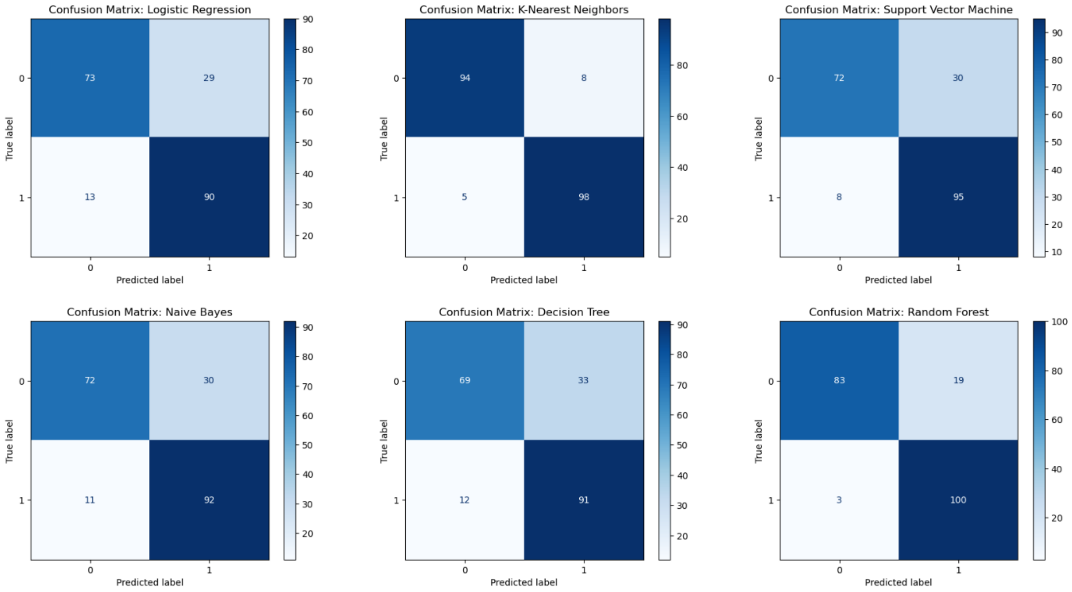
K-Nearest Neighbors (KNN) is a non-parametric algorithm that classifies data points based on the classes of their nearest neighbors. It is straightforward and highly interpretable, and it adapts well to complex datasets. For heart disease prediction, KNN can be particularly effective when patterns in patient data are similar among risk groups. However, KNN tends to be computationally expensive, especially with larger datasets, as it calculates distances for each new instance. Additionally, its performance may degrade if features are not properly scaled, and it is more prone to overfitting, especially if the dataset is noisy.

Support Vector Machine (SVM) is another powerful classifier that is widely used in heart disease detection for its ability to handle high-dimensional data and define clear decision boundaries. SVM tries to maximize the margin between classes, making it robust in cases where classes are well-separated. It works well in complex, non-linear datasets, especially when paired with kernel functions. However, SVMs are sensitive to parameter selection and require careful tuning, which can be time-consuming. Additionally, it can be computationally expensive with large datasets, and interpretability may be lower than simpler models.

Naive Bayes is a probabilistic classifier that makes predictions based on the Bayes theorem, assuming feature independence. It is efficient and fast, making it suitable for large datasets, and it performs surprisingly well despite its simplicity. For heart disease detection, Naive Bayes can be effective if the data largely meets the independence assumption. However, in practice, heart disease risk factors are often correlated, and Naive Bayes may not capture complex interactions between features as effectively as other models, which may result in reduced accuracy.

Decision Trees and Random Forest are popular choices in heart disease prediction due to their flexibility and robustness. A Decision Tree divides the data into smaller subsets based on feature values, making it easy to interpret but prone to overfitting. The Random Forest, an ensemble of Decision Trees, addresses this limitation by averaging predictions across many trees built from random data samples, improving accuracy and reducing variance. Random Forest is powerful for heart disease detection, handling complex interactions among features and ranking feature importance, which aids in understanding key risk factors.


Accuracy Comparison of Different Algorithms
Accuracy %
0
20
40
60
80
100
81.42%
88.05%
81.46%
81.08%
83.9%
89.38%
Logistic Regression
KNN
SVM
Naive Bayes
Decision Tree
Random Forest
Algorithms


Confusion Matrix: Logistic Regression
73
29
13
90
Confusion Matrix: K-Nearest Neighbors
94
8
5
98
Confusion Matrix: Support Vector Machine
72
30
8
95
Confusion Matrix: Naive Bayes
72
30
11
92
Confusion Matrix: Decision Tree
69
33
12
91
Confusion Matrix: Random Forest
83
19
3
100
True label
Predicted label
0
1

7 References

Uddin, S., Khan, A., Hossain, M. E., Moni, M. A. (2019). "Comparing different supervised machine learning algorithms for disease prediction." BMC Medical Informatics and Decision Making, 19(1), 1-16. This study compares various machine learning models, including Logistic Regression, SVM, and Random Forest, for medical data applications.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I. (2017). "Machine learning and data mining methods in diabetes research." Computational and Structural Biotechnology Journal, 15, 104-116. Although focused on diabetes, this paper provides a comprehensive overview of machine learning algorithms, many of which are applicable to heart disease detection.

Chaurasia, V., Pal, S. (2014). "Data mining techniques: to predict and resolve breast cancer survivability." International Journal of Computer Science and Mobile Computing, 3(1), 10-22. This paper explores Decision Trees, Naive Bayes, and KNN for medical predictions, including insights that can be transferred to heart disease studies.

Amin, M. S., Chiam, Y. K., Varathan, K. D. (2019). "Identification of significant features and data mining techniques in predicting heart disease." Tehnički vjesnik, 26(1), 149-154. This article compares models like KNN, SVM, and Decision Trees, focusing on their predictive capabilities in heart disease contexts.

Jabbar, M. A., Deekshatulu, B. L., Chandra, P. (2013). "Heart disease prediction using lazy associative classification." International Journal of Computer Science and Engineering, 2(2), 1-4. This paper discusses KNN and Naive Bayes for heart disease prediction, providing performance insights useful for comparison.

National Center for Biotechnology Information (NCBI): https://www.ncbi.nlm.nih.gov/

Kaggle: https://www.kaggle.com/

Towards Data Science (on Medium): https://towardsdatascience.com/

Analytics Vidhya: https://www.analyticsvidhya.com/

Scikit-Learn Documentation: https://scikit-learn.org/stable/

8 Future Sccope

The future of heart disease detection using machine learning holds promising developments, especially with the integration of wearable technology and real-time health monitoring. Wearable devices such as smartwatches and fitness trackers now capture continuous data on metrics like heart rate, ECG, and physical activity, all of which can be analyzed using machine learning algorithms to monitor cardiovascular health. This integration allows for real-time detection of irregularities and immediate alerts for medical intervention, potentially preventing heart attacks and other severe outcomes. With such advancements, healthcare systems can shift towards more proactive and preventive care, reducing the reliance on hospital visits and improving patient outcomes.

Another promising area is the advancement of personalized medicine, where machine learning models can be tailored to account for an individual's unique profile, including genetics, lifestyle habits, and environmental factors. Future models could consider these variables to provide personalized risk assessments and customized prevention strategies for each patient, which would be especially valuable for high-risk individuals. This approach could improve prediction accuracy and offer treatment recommendations based on each patient's specific needs, ultimately enhancing the effectiveness of heart disease prevention and management on a global scale.

9 Conclusion

In comparing algorithms for heart disease detection, each model—Logistic Regression, KNN, SVM, Naive Bayes, Decision Tree, and Random Forest—has unique strengths and limitations. Logistic Regression is useful for its simplicity and interpretability, especially when there is a linear relationship among features. KNN is effective when there are clear similarities among patient profiles but can struggle with large datasets. SVM excels in handling high-dimensional, non-linear data but requires careful tuning. Naive Bayes offers computational efficiency, yet its assumption of feature independence can limit its accuracy in complex medical datasets. Decision Trees are easily interpretable and adapt to non-linear data, though they are prone to overfitting; Random Forests counter this by combining multiple trees to improve accuracy and stability.

Ultimately, the choice of algorithm depends on factors like data complexity, model interpretability, and computational efficiency. Random Forest often emerges as a robust option for heart disease detection due to its high accuracy and ability to manage non-linear relationships and feature interactions, making it ideal for complex healthcare data. However, simpler models like Logistic Regression or Naive Bayes might be preferred in settings where interpretability and fast computation are key. Thus, the ideal model choice may vary based on specific project requirements, data characteristics, and the balance needed between interpretability and predictive performance.