STM JOURNALS

Research & Reviews: A Journal of
Bioinformatics

https://journals.stmjournals.com/rrjobi

ISSN: 2393-8722
Volume 11, Issue 3, 2024
September—December
DOI (Journal): 10.37591/RRJoBI

Review RRJoBI

Performance Analysis of Machine Learning Algorithms for Disease Prediction

Prasad Kathalkar[1], Eshwar Satale[2], Jeet Gate[3], Chandrashekhar G. Patil[4,*]

Abstract

In this 21st century, where Digitization makes humans measure, record, analyze and to manipulate the huge amount of data as per the requirement, prediction of the decease based on Machine Learning models will be representing one of the good applications of the efficient data handling. An Automatic Decease Prediction system based on the symptoms would be the great boon for the medical practitioners. The Supervised Machine Learning models, such as Logistic regression, Random Forest, K-Nearest neighbor, Decision Tree and Navie Bays models are tested here for the deployment of the automation of the detection of the decease. The different set of the data of the symptoms (both online available data and the data collected from hospitals) of the diseases like Heart disease, Breast cancer, Diabetes are undertaken for the comparison of these models. On comparison of the performance of each of the models, the Random Forest model is found to be performed optimally under some conditions.

Keywords: Machine learning, disease prediction, heart disease, breast cancer, diabetes

INTRODUCTION

Background

The field of healthcare has witnessed significant advancements in recent years, with technology playing a pivotal role inrevolutionizing the way diseases are diagnosed and managed. Machine learning, a branch of artificial intelligence, has proven to be a highly effective tool within the healthcare industry. It offers fresh possibilities for predicting, diagnosing, and managing a variety of diseases with greater efficiency. Machine learning algorithms can analyze large datasets to detect patterns and make highly accurate predictions. This allows healthcare professionals to make better-informed decisions, promotes early diagnosis, and ultimately enhances patient outcomes. Machine learning techniques have shown great potential in disease prediction and management. By analyzing large volumes of patient data. Machine learning algorithms can uncover insights and generate predictions with precision. This empowers healthcare providers to make data-driven decisions, improve early diagnosis, and enhance patient care.

***Author for Correspondence**
Chandrashekhar G. Patil
E-mail: chandrashekhar.patil_skncoe@sinhgad.edu

[1-3]Student, Department of Electronics & Telecommunication Engineering, Smt. Kashibai Navale College of Engineering, Vadgaon(BK), Pune, Maharashtra, India.
[4]Associate Professor, Department of Electronics & Telecommunication Engineering, Smt. Kashibai Navale College of Engineering, Vadgaon (BK), Pune, Maharashtra, India.

Received Date: July 16, 2024
Accepted Date: October 02, 2024
Published Date: November 11, 2024

Citation: Prasad Kathalkar, Eshwar Satale, Jeet Gate, Chandrashekhar G. Patil. Performance Analysis of Machine Learning Algorithms for Disease Prediction. Research & Reviews: A Journal of Bioinformatics. 2024; 11(3): 9–18p.

Relevance

The "Disease Prediction Using Machine Learning" project is profoundly relevant in the contemporary healthcare landscape. With the global burden of diseases on the rise and the increasing strain on healthcare systems, the need for early prediction has never been more critical. Leveraging the power of machine learning, this project not only offers the potential to save lives by enabling early intervention but also presents a pathway to optimize healthcare resources and reduce costs. In an era of data-driven healthcare, where personalized

medicine and precise decision-making are paramount, the project aligns perfectly with these evolving healthcare paradigms. By improving the accuracy of disease prediction, it enhances the quality of patient care, empowers individuals to take charge of their health, and contributes to ongoing research in the field of healthcare technology. Additionally, as global health challenges continue to emerge, the project's proactive approach can play a pivotal role in early detection and response. Thus, the "Disease Prediction Using Machine Learning" project is not just relevant but also a transformative endeavor with the potential to reshape healthcare delivery and improve public health outcomes.

This paper is a groundbreaking attempt to harness the potential of machine learning in healthcare. It encompasses data collection, preprocessing, model development, and deployment, aiming to create an accurate and accessible disease prediction system. It strives to empower healthcare professionals with a tool that can make early disease detection more efficient, personalized, and cost-effective. By integrating technology and healthcare, this project addresses the ever-increasing burden of diseases and promises to revolutionize the way we approach healthcare delivery, with the potential to enhance patient outcomes and improve overall public health.

Chapter 1 introduces a general overview of the project under consideration and the relevance of a disease prediction system using machine learning. In Chapter 2, research papers were studied and the body of knowledge on machine learning algorithms used to predict various diseases was examined. Chapter 3 outlines the project's approach, including how the hospital dataset was trained and tested using various machine learning algorithms to achieve accurate results. In Chapter 4, disease prediction using machine learning is implemented with algorithms like Logistic Regression, K-Nearest Neighbor, Naïve Bayes, and Decision Tree for accurate results. Chapter 5 presents the project's results and discussion, demonstrating that a high accuracy rate in disease prediction was achieved by the machine learning models, indicating their potential to significantly impact healthcare by facilitating early diagnosis and improving patient care. Chapter 6 showcases the potential of advanced technology to revolutionize disease prediction through the 'Disease Prediction Using Machine Learning' project. Chapter 7 lists the references consulted for the project report.

The "Disease Prediction Using Machine Learning" project leverages advanced machine learning techniques to accurately predict diseases early. Through the collection and analysis of extensive healthcare data, the project showcases its ability to enhance patient outcomes, streamline resource management, and lower healthcare expenses. The project's results suggest that this approach could revolutionize disease diagnosis and management, offering significant benefits to both healthcare professionals and patients.

LITERATURE SURVEY (RELATED WORK)

In [1], a disease prediction system uses predictive modelling to identify a user's illness based on the symptoms they input. The system processes the provided symptoms and predicts the probability of the illness. A decision tree classifier is employed to achieve disease prediction. The likelihood of the illness is calculated by a decision tree classifier. An increasing number of organizations in the biological and healthcare fields are utilizing big data to support early disease detection and enhance patient care. The development of a system that will allow people to forecast chronic illnesses without having to see a doctor or medical professional for a diagnostic is necessary by watching patient signs and using a variety of machine learning modelling methods, different illnesses can be identified. Text and organized data processing do not follow any standard method. The suggested paradigm would examine both organized and random material. Prediction precision can be increased through machine learning to diagnose various diseases by analysing patient symptoms through different machine learning techniques. Medical professionals face challenges in accurately assessing symptoms and identifying diseases early due to the vast amount of data [2]. There are two approaches to determining whether a patient has a specific disease. Manually it takes a lot of time. Applying machine learning algorithms makes the work easy, such that, there is a specific application particularly for each disease. There is no common system or

application for predicting multiple diseases. Hence, the proposed system contains single application which can predict various diseases, with the help of user inputs. The algorithms used in the proposed system like SVM, Logistic Regression, Decision Tree, KNN algorithm. The system's results are presented in a binary format, such as "yes" or "no".

The disease prediction involves estimating the likelihood of a patient's illness by analyzing the combination of their symptoms [3]. Monitoring a patient's condition and health information during the initial examination allows for more effective treatment. This analysis in healthcare can streamline and accelerate patient care. Previous research has focused on machine learning models, such as Support Vector Machine (SVM), K-nearest neighbours (KNN), and RUS boost to detect diseases based on symptom inputs. The proposed model improves the accuracy and performance of earlier models by addressing their limitations. It uses a medical dataset from Kaggle, transforming the data by assigning weights based on rarity. The dataset is trained using a combination of machine learning algorithms, including Random Forest, Long Short-Term Memory (LSTM), and SVM. Additionally, the patient's medical history is analyzed using LSTM, and SVM is employed to determine the likely disease. The proposed model achieves better accuracy and reliability compared to state-of-the-art methods. In [4], machine learning techniques have significantly advanced healthcare by enabling precise and timely disease prediction. Predicting multiple diseases at once can enhance early diagnosis and treatment, improving patient outcomes and reducing costs. This research explores the application of machine learning in multi-disease prediction, examining its benefits, challenges, and future directions. It reviews various machine learning models and data sources commonly used for disease prediction. The importance of feature selection, model evaluation, and integrating multiple data types for improved disease prediction is also discussed. The findings highlight the potential impact of machine learning on public health by enabling multi-disease prediction. The model trains itself using sample data to predict whether an individual is affected by certain diseases. The paper [5] aims to help individuals predict their diseases through symptom input and vital signs. Many people avoid addressing health issues due to high medical fees, leading to severe problems later. If uninsured, medical bills can be burdensome. This website seeks to ease the burden by estimating the disease and its severity based on symptoms, age, and gender. The diagnosis system not only predicts whether a user has a specific disease but also provides diet plans and exercises to reduce the effects of the disease. It offers a simple yet effective approach for predicting diseases when accurate vital sign data is available. In [6], disease prediction is accomplished by processing symptoms provided by users and outputting the probability of a disease. The Naïve Bayes classifier, a supervised machine learning algorithm, is employed for this purpose. With the growing amount of healthcare data, accurate medical analysis benefits early disease detection and patient care. The system predicts diseases, such as Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis using linear regression and decision tree models.

In [7], the importance of healthcare has become more evident, especially during the pandemic. When feeling unwell, people often check their temperature to decide whether to seek medical help. Similarly, a medical disease prediction app can provide a preliminary diagnosis, helping individuals decide whether to consult a doctor or try home remedies for temporary relief. Integrating machine learning with an application interface enables users to interact with the model for more accurate predictions. In [8], the "Disease Prediction" method focuses on predictive modeling, using the user's symptoms as input to estimate the likelihood of a disease. The random forest classifier is used for this purpose. [9] states that, while the human immune system protects the body, it sometimes fails to prevent diseases caused by environmental factors and lifestyle choices, leading to high mortality rates. Diagnosing these diseases can be challenging, and timely detection is crucial for effective treatment. With the increase in medical data, researchers are using machine learning algorithms to aid healthcare professionals in diagnosing diseases. This paper surveys various supervised learning models, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Naïve Bayes, and Random Forest (RF), analyzing their performance. In [10], a disease prediction system uses predictive modeling to estimate the likelihood of a disease based on symptoms input by the user. The system employs a decision

tree classifier to analyze the symptoms and calculate the probability of the disease. The growing volume of medical data in healthcare and biomedical fields enhances early disease detection and patient care through accurate analysis.

PROBLEM STATEMENT

Determining the presence of a specific disease in the human body through traditional methods is often time-consuming. Although various applications exist for individual diseases, users must switch between different tabs if they wish to check for multiple conditions.

PROPOSED SOLUTION

This project focuses on developing an application that integrates multiple disease predictions within a user-friendly interface. A viable solution to this challenge is to implement machine learning algorithms to create a unified model capable of predicting the presence or absence of several diseases at once, eliminating the need for users to navigate between different applications or tabs. Hence, users can easily check for the probability of having as many diseases present in that application.

SYSTEM IMPLEMENTATION (METHODOLOGY)

Gather relevant data from various sources, clean the data to remove noise, handle missing values, and engineer features to extract valuable information for prediction divide the dataset into training, validation, and testing sets. Further preprocessing may involve normalization or scaling of data. Choose an appropriate machine learning model, train it using the training data, and perform hyper parameter tuning to optimize its performance. Evaluate the trained model using the validation data, assess its performance, and test it on the separate testing dataset. Measure the model's effectiveness using various metrics. Deploy the trained model into a production environment and integrate it with the healthcare system or application where it will be utilized. Create a user-friendly interface to facilitate interaction. Continuously monitor the model's performance in real-world conditions, gather user feedback, and update the model regularly to ensure it maintains accuracy. Feedback on the model's predictions should be used to refine and improve its performance over time, ensuring it remains effective and relevant. Additionally, implement robust security measures to protect patient data and ensure compliance with healthcare regulations (Figure 1).

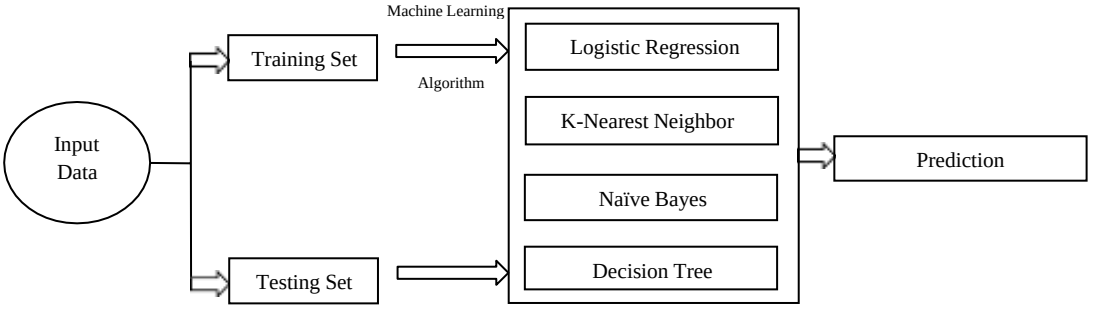


Figure 1. Block Diagram of disease prediction using machine learning.

The block diagram within the "Disease Prediction Using Machine Learning" project report serves as a schematic representation of the project's intricate workflow. It visually illustrates the journey of data, starting from data collection through various sources, progressing to data preprocessing, and ultimately feeding into machine learning models for disease prediction. This diagram encapsulates the project's technical architecture.

IMPLEMENTATION

Some of the algorithms are Support Vector Machine, KNN, Random Forest, Logistic Regression are used for classification problems.

Logistic Regression

Logistic regression, a crucial statistical technique, serves as the core predictive model in our project. It is a versatile tool used for binary classification, enabling us to determine the probability of a particular event occurring based on a set of predictor variables. With its mathematical formulation and use of the sigmoid function, we can understand how changes in the predictors influence the likelihood of the event of interest. Our project leverages logistic regression to make data-driven predictions, contributing to more informed decision-making in the context of disease prediction. Further, the results and findings obtained through this technique have been explored, shedding light on its relevance in achieving our project's objectives (Figure 2).

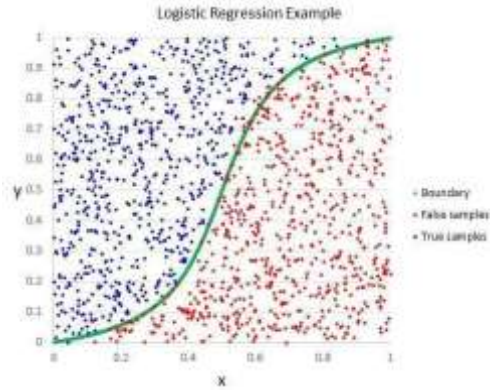


Figure 2. Logistic Regression.

$$\text{Mathematical Equation: } (y = 1) = \frac{1}{1+e^{-z}} \quad \text{(i)}$$

NEAREST NEIGHBOR

In our project, K-Nearest Neighbors (K-NN) is a prominent machine learning algorithm employed for disease prediction. K-NN is a non-parametric classification technique that assigns data points to the class most common among their "k" nearest neighbors. This straightforward yet effective method has been instrumental in achieving accurate predictions. It considers the proximity of data points to make decisions, making it particularly suitable for our task of classifying diseases based on patient data. Our project report will delve into the specific implementation of K-NN, its performance, and its role in enhancing the accuracy and reliability of our predictivemodel (Figure 3).

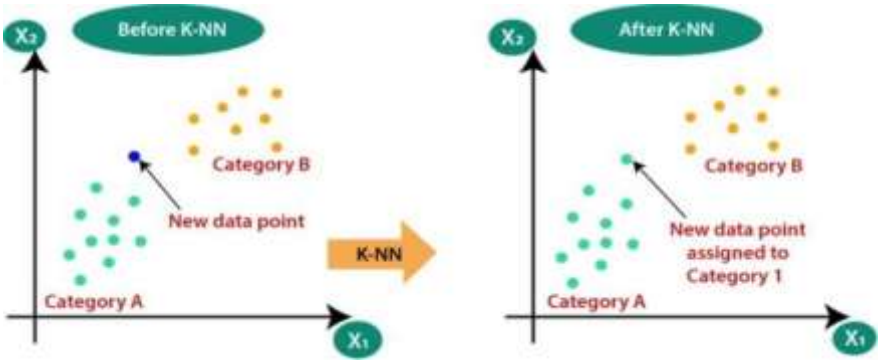


Figure 3. K-Nearest Neighbors.

$$\text{Mathematical equation: Euclidean distance } = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad \text{(ii)}$$

RANDOM FOREST

Random Forest, a powerful ensemble learning algorithm, plays a pivotal role in our project for disease prediction. This technique combines the predictions of multiple decision trees to yield robust and accurate results. The model's ability to handle complex data, mitigate over fitting, and provide feature importance insights makes it a valuable component of our project. In the report, the implementation of Random Forest, its impact on predictive performance, and how it contributes to enhancing the reliability and interpretability of our disease prediction model (Figure 4) is detailed.

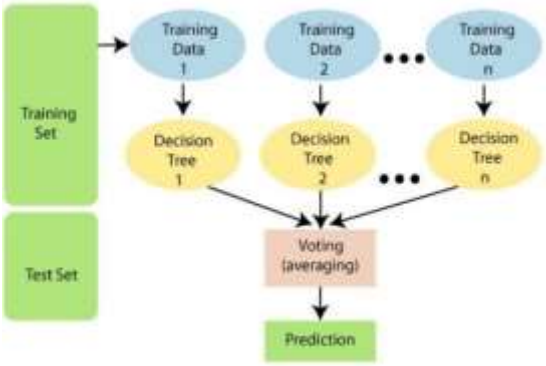


Figure 4. Random Forest.

Mathematical Equation: $fi = \frac{=\sum j:node\ j\ splits\ on\ feature\ i\ nij}{\sum k \in all\ nodes\ nik}$ (iii)

RESULTS & DISCUSSIONS

In this section, we delve into the heart of the "Disease Prediction Using Machine Learning" project, presenting the results and engaging in a comprehensive discussion. The results provide insights into the project's technical accomplishments, specifically the performance of machine learning models in disease prediction. Furthermore, the discussion explores the broader implications of these results, focusing on their impact on healthcare, patient outcomes, and resource allocation. This segment combines the technical evaluation of the models with their real-world relevance, setting the stage for the conclusions that follow, and highlighting the project's significance in the field of healthcare technology.

Heart Disease

Figure 5 shows the performance comparison between Logistic Regression and K-Nearest Neighbors (KNN) models for heart disease prediction. The confusion matrices provide detailed insights into the classification outcomes:

Logistic Regression Confusion Matrix

- *True Positives:* 90.
- *True Negatives:* 71.
- *False Positives:* 31.
- *False Negatives:* 13.

K-Nearest Neighbors Confusion Matrix

- *True Positives:* 94
- *True Negatives:* 91
- *False Positives:* 11
- *False Negatives:* 9

The bar charts compare the accuracy and training accuracy of both models. KNN demonstrates better overall performance with higher accuracy and fewer false predictions. The training accuracy plot shows that both models are well-fitted, with KNN displaying a slight edge in performance. These results suggest that KNN may be more effective for heart disease prediction in this scenario.

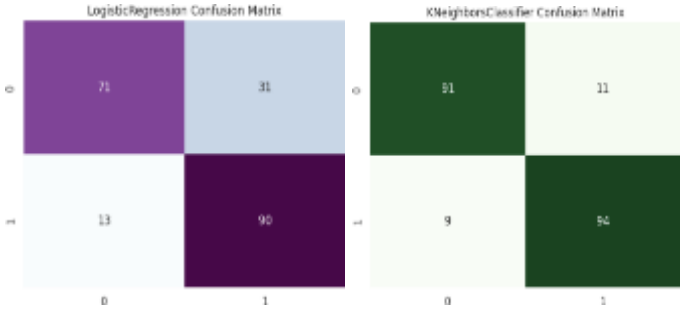
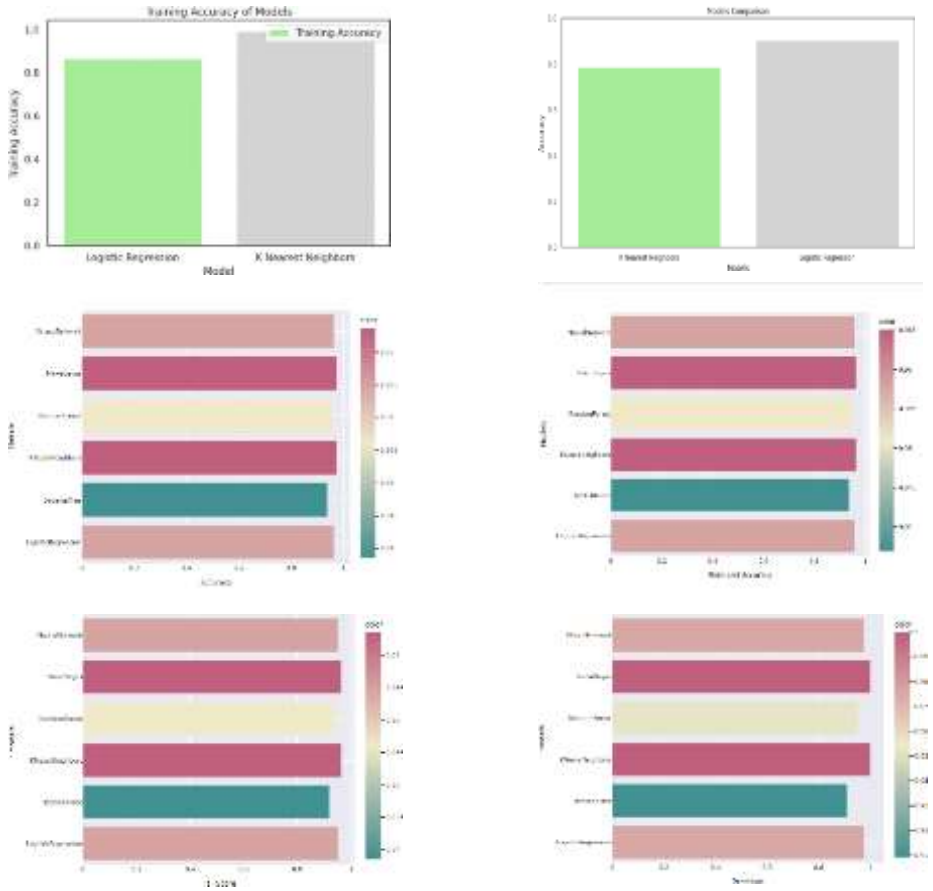


Figure 5. Results of Heart Disease Classification Models.

Breast Cancer

Figure 6 provides a comparison of multiple models used for breast cancer prediction, including Neural Network, Naive Bayes, Random Forest, K-Nearest Neighbors, Decision Tree, and Logistic Regression. The metrics shown include,

- *Accuracy:* Measures the overall correctness of the model predictions.
- *Balanced Accuracy:* Considers the balance of class distributions in accuracy.
- *F1 Score:* Represents the harmonic mean of precision and recall.
- *Precision:* Indicates the proportion of true positive predictions among all positive predictions.
- *Recall:* Reflects the ability to identify all true positive instances.



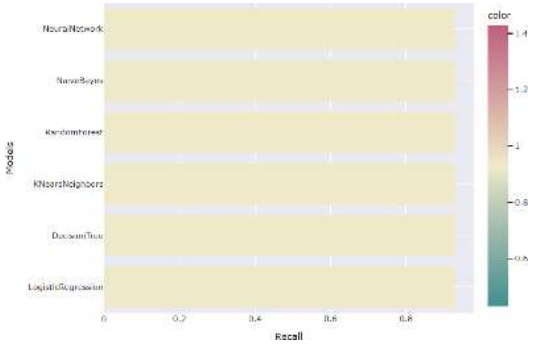


Figure 6. Results of Breast Cancer.

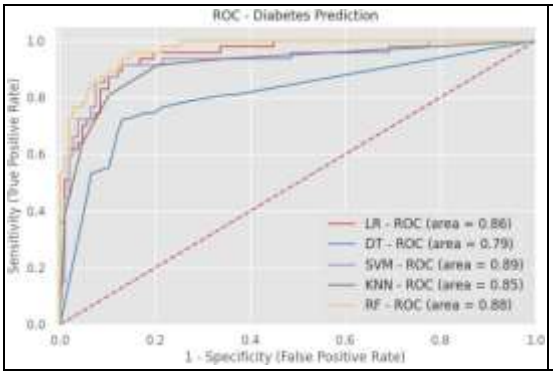
Neural Network and Naive Bayes models show the highest performance across most metrics, followed closely by Random Forest and K-Nearest Neighbors. These metrics highlight the strengths and predictive quality of each model, aiding in the selection of the best algorithm for breast cancer prediction.

Diabetes

Figure 7 presents the performance evaluation of various machine learning models – Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF) – for diabetes prediction. The bar chart compares each model's accuracy and ROC scores, highlighting their effectiveness.

- *Accuracy (%):* SVM achieves the highest accuracy, closely followed by KNN and RF.
- *ROC (%):* The ROC curve illustrates the area under the curve (AUC) for each model:
  - *SVM:* 0.89
  - *Logistic Regression:* 0.86
  - *Random Forest:* 0.88
  - *KNN:* 0.85
  - *Decision Tree:* 0.79

These results demonstrate the strong predictive capabilities of SVM and Random Forest, making them suitable choices for diabetes prediction tasks.



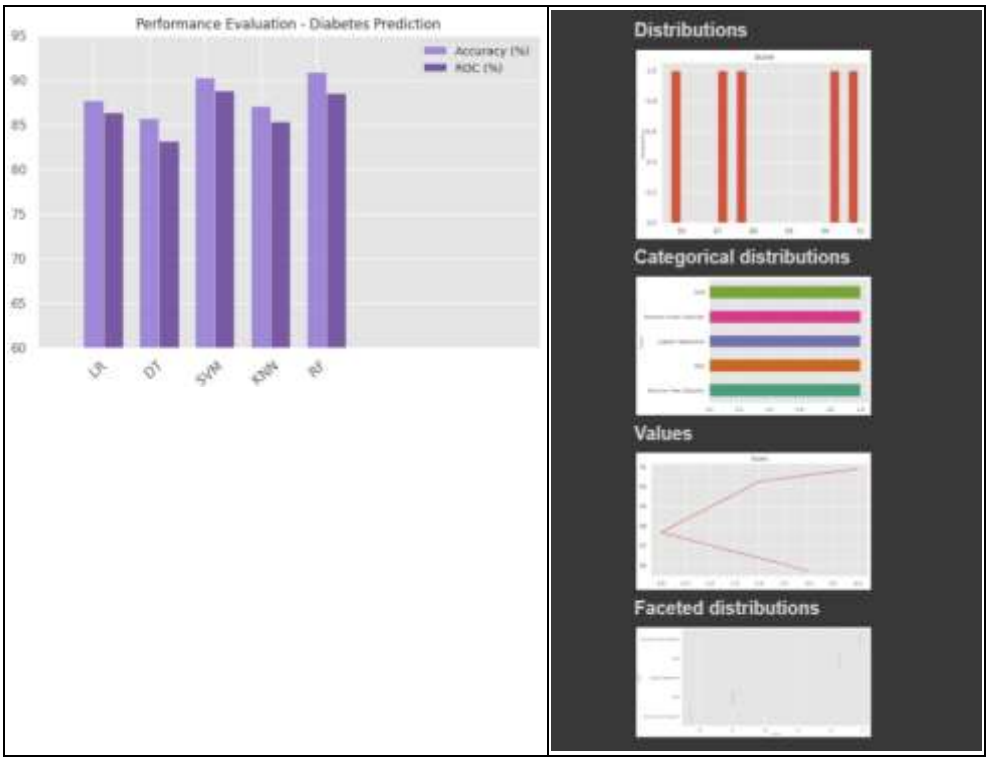


Figure 7. Result of diabetes.

CONCLUSIONS

The "Disease Prediction Using Machine Learning" project has partially demonstrated the remarkable potential of machine learning in transforming disease prediction and management. Through systematic data collection, preprocessing, and model development, it is shown that early disease detection is not only achievable but can significantly enhance patient care and reduce healthcare costs. As the project moves towards deployment, it promises to empower healthcare professionals with a powerful tool for decision support, aligning with the evolving landscape of data-driven healthcare. In conclusion, this project represents a significant step forward in the intersection of technology and healthcare, with the capacity to revolutionize disease diagnosis, treatment, and ultimately improve public health outcomes. Each of these algorithms was trained and tested on a dataset of medical records, with the following accurate results:

- *Logistic Regression:* Achieved accuracy of around 78%.
- *K-Nearest Neighbors (KNN):* Achieved an accuracy of around 73%.
- *Random Forest:* Demonstrated the highest accuracy of around 98%.

REFERENCES

1. Khan MN, Srivastava A. Disease prediction using machine learning. Int J Eng Manag Res (IJEMR). 2023 Jun;13(3):23–30.
2. Raju K, Priya H, Supraja M. Multiple disease prediction using machine learning. J Emerg Technol Innov Res (JETIR). 2023 Apr;10(4):88–95.
3. Gaurav K, Kumar A, Singh P, Kumari A, Kasar M, Suryawanshi T. A detailed review on disease prediction models that use machine learning: Human disease prediction using machine learning techniques and real-life parameters. Int J Eng (IJE). 2023 Jun;36(6):78–85.
4. Parshant, Rathee A. Multiple disease prediction using machine learning. IRE J. 2023 Dec;6(12):45–50.

5. Bhatt A, Singasane S, Chaube N. Disease prediction using machine learning. Int Res J Modern Eng Technol Sci (IRJMETS). 2022 Jan;4(1):11–18.
6. Gomathy CK, Naidu AR. The prediction of disease using machine learning. Int J Sci Res Eng Manag (IJSREM). 2021 Oct;5(10):62–70.
7. Takke K, Bhaijee R, Singh A, Patil A. Medical disease prediction using machine learning algorithms. Int J Res Appl Sci Eng Technol (IJRASET). 2021 May;10(5):102–110.
8. Reddy PP, Babu DM, Kumar H, Sharma S. Disease prediction using machine learning. Int J Creat Res Thoughts (IJCRT). 2021 May;9(5):1234–1242.
9. Farooqui ME, Ahmad J. A detailed review on disease prediction models that use machine learning. Int J Innov Res Comput Sci Technol (IJIRCST). 2020 Jul;8(4):40–48.
10. Chauhan RH, Naik DN, Halpati RA, Patel SJ, Prajapati AD. Disease prediction using machine learning. Int Res J Eng Technol (IRJET). 2020 May;7(5):56–63.

