iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62177

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Naive Bayes - Random Forest Ensemble Model Analysis and HeartDisease Prediction

Nivyn Bybin[1], Parvathy Gopan[2], Ramkrishna K[3], Ryan Sebastain Jimmy[4], Anu Eldho[5], Rotney Roy Meckamalil[6]
Department of Computer Science and Engineering Mar Athanasius College of Engineering, Kothamangalam, Kerala

Abstract: ***The critical task of accurately diagnosing heart disease can potentially save lives. Leveraging the UCI machine learning heart disease dataset and employing ensemble machine learning techniques, including Gausian Naive Bayes and Random Forest algorithms, the research investigates 13 primary characteristics to predict heart disease risks. The conventional machine learning approaches are utilized to analyze the dataset, unveiling correlations between features and heart infection risks. The aim is to develop a user-friendly interface, the Heart Disease Clinical Decision Support System (HDCDSS), where patients input their clinical details to receive a predictive analysis of their coronary disease. The system, built in Python with Flask and Bootstrap, provides clients with personalized reports on their heart health, enhancing diagnostic accuracy and potentiallystreamlining healthcare processes.***

Index Terms: heart disease, risks, infections, user friendly,machine learning, dataset.

I. INTRODUCTION

Heart disease stays a pressing worldwide fitness issue, stressful available and green diagnostic tools. This projectaddresses this want by means of developing a scientific recordsevaluation device for coronary heart disorder prediction, utilising ensemble strategies. With cardiovascular sicknesses (CVDs) ranking as the leading motive of world mortality in line with the World Health Organization (WHO), early and accurate analysis is vital for effective treatment and improved affected person results. The project's goal is to leverage the strengths of Gaussian Naive Bayes and Random Forest algorithms via an ensemble version for enhanced coronary heart ailment prediction.

Traditional diagnostic methods for heart disorder can be time-ingesting, highly-priced, and invasive. Despite numerous machine learning fashions available for heart ailment pre- diction, there often exists a exchange-off among accuracy and simplicity of implementation. This challenge proposes a person-pleasant system that utilizes device learning to mitigate these boundaries. By imparting a handy threat evaluation tool, the machine aims to improve accessibility, probably lessen healthcare costs, and facilitate early detection for higherpatient effects.

The machine encompasses capabilities which include con- sumer registration and authentication, user input of applicable medical attributes, a educated device gaining knowledge of version for threat evaluation, and a user-friendly interface.Flask (Python) serves because the web application back- end, interacting with a PostgreSQL database, whilst Scikit- research libraries educate and combine device studying mod- els. HTML/CSS/Bootstrap make contributions to constructing an intuitive interface, making sure ease of use for both patientsand healthcare specialists.

The ensemble model, combining Gaussian Naive Bayes andRandom Forest. Although person models carried out well withaverage accuracy's around eighty-four percent, the ensemble technique continually outperformed them, emphasizing the advantage of combining algorithms for improved prediction accuracy.

In conclusion, this mission amalgamates internet improvement with gadget studying strategies to create an software facilitating coronary heart disorder risk assessment. By leveraging Flask for net improvement and Scikit-studyfor device getting to know, the utility gives an green and person-friendly platform for predicting coronary heart disorder threat based on man or woman medical attributes. Moving forward, deploying the confirmed version in medical settings, increasing prediction abilities to other cardiovascular diseases,and continuous version updates constitute promising avenues for future upgrades.

II. RELATED WORKS

A. Cardiovascular Disease Prediction

Heart disease is a really serious illness that can lead todeath for a lot of people. But figuring out if someone hasheart disease can be tricky for doctors.

Nowadays, because heart problems are so common, predicting who might have heart disease has become really important.

Researchers wanted to find the best way to predict if someone might have heart disease by looking at different factors that could be related to it. So, they used computer programs called Machine Learning techniques to help them. These programs learn from data to make predictions. They tried out a bunch of different techniques like Decision Tree, Support Vector Machine, and others to see which one was the best at predicting heart disease. They tested their techniques using information from three big sets of data, and they found that one technique, called Decision Tree, was the most accurate. It was right about ninety-nine percent of the time in predicting if someone might have heart disease.

B. Early Heart Diseases Diagnosis

Heart disease is a major cause of death worldwide, but if we can catch it early, we can treat it better and save lives. To help doctors make better diagnoses, researchers came up with a computer system called a Clinical Decision Support System. This system uses data from patients to help doctors figure out if someone might have heart disease. Now, the researchers wanted to make this system even better, so they created a special model called a Heart Disease Prediction Model. This model uses a few fancy techniques to improve how it works. First, they clean up the data by getting rid of any weird or unusual bits. Then, they make sure the data they have is balanced and represents different types of patients well. Finally, they use XGBoost, which is like a smart computer program, to predict if someone might have heart disease based on their data. They tested this model using information from two big sets of data that other researchers have shared. They compared their model to other ones people have used before, like Naive Bayes, Logistic Regression, and others. The results showed that their model was better at predicting heart disease than the others. It was accurate about ninety-six percent of the time for one dataset and ninety-eight percent for another. They also made a prototype of a computer program based on their model to help doctors diagnose heart disease in patients faster. This means doctors can catch heart disease earlier and start treatment sooner, which could save lives.

C. Smart Healthcare Monitoring System For Heart Diseases

Predicting heart disease accurately is really important for treating patients before they have a heart attack. There have been a lot of computer systems recently that use machine learning to try to predict heart disease, but they have some limitations. These systems struggle when they have a lot of different kinds of data to work with, like data from sensors and medical records. Also, they often aren't very good at picking out the most important information from all that data. So, in this study, researchers came up with a smarter system for predicting heart disease. First, they combined data from sensors and medical records to get a better picture of each patient's health. Then, they used a technique to pick out the most important information and ignore the stuff that doesn't matter as much. They also figured out a way to give different importance to different pieces of information depending on the situation. Finally, they used a special kind of computer program called an ensemble deep learning model to make predictions about heart disease. When they tested their system with real heart disease data, they found that it was really accurate, getting it right about ninety-five percent of the time. This means their system is better at predicting heart disease than other systems people have made before.

D. Decision Support System For Heart Disease Prediction

Cardiovascular disease is a major cause of death worldwide these days. To tackle this problem, doctors and researchers are using a technique called data mining, which helps them sift through lots of medical data to find useful patterns. This study looked at 25 different research papers that used machine learning to predict heart disease. It examined the different types of machine learning models used and the features they looked at to make predictions about who might get heart disease. The study found that some models performed better than others in predicting heart disease, and it also looked at the tools that were used in these research papers. Overall, the study identified some gaps in the research and suggested areas for future work to improve how we predict heart disease using machine learning.

III. PROPOSED MODEL

The program combines rich and diverse datasets with at- tributes important for heart health assessment including age, gender, blood pressure and fat mass This dataset contains 14 attributes with a total of 14 attributes one target is included, and provides a comprehensive basis for training and predic- tion of evaluating models Model evaluation Alternatively, the method adopts a robust method of stratified k-fold cross- validation by partitioning the dataset into k clusters while preserving the class distribution, this approach increases the reliability of model performance assessment across different data subsets. An important part of the work is the use of cohort learning techniques, particularly the benefits of soft voting classifier.

This sophisticated approach combines predictions of different base models, including Gaussian naive base and ran- dom forest classifiers. By combining the individual strengths of these models, the team approach leverages the collective forecasting power, producing more robust and accurate resultsThe slower boat gathers potential resources the results gener- ated by each base model edge, culminating in a final predictionof the group consensus. In parallel, the project uses Flask, a versatile Python framework, to create dynamic and interactiveweb applications. With the power of Flask, users have easy access to the predictive model, allowing them to input their medical data and receive personalized risk assessments for cardiovascular disease This user-friendly interface not only enables scalability but encourages greater use of predictive tools, proactively for individuals

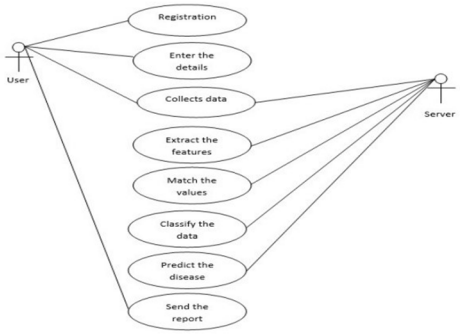


Fig. 1. Use Case Diagram of Our Proposed Model

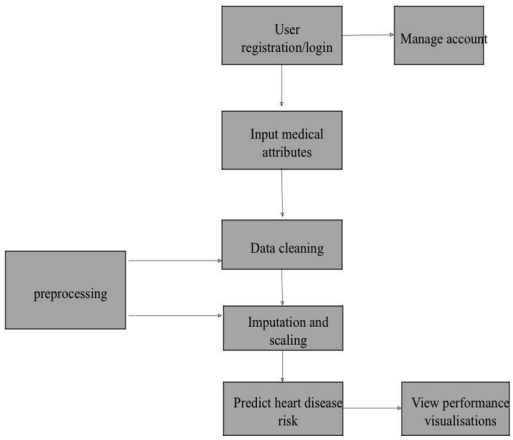


Fig. 2. Data Flow Diagram of Our Proposed Model

IV. IMPLEMENTATION

Naive Bayes - Randomized forest model analysis and design for cardiovascular disease prediction uses a multidimensional model that incorporates multiple interacting variables and observational steps that it is robust and accurate in predicting cardiovascular risk and body composition This set of data, including lifestyle factors, serves as a foundational input for model development and subsequent phases of analysis. The initial phase involves careful data preprocessing and feature engineering to clean up the raw data set and convert it into a structured format suitable for analysis. This includes handling missing values, normalizing or scaling numerical features, encoding categorical variables, and potentially performing dimensionality reduction techniques to increase sample efficiency. Then, the preprocessed data goes through stratified k-factor cross-validation, a complex checklist designed to evaluate model performance on different subsets of the data set while maintaining squared balance This iterative procedure checks find that the predictive model generalizes well to unseen data and reduces the risk of overfitting orbiased evaluation results.

A. Data Collection

The foundation of a machine learning project lies in thedata used to train the models. For this project, we leveraged a publicly available dataset: the Cleveland heart disease dataset from the UCI Machine Learning repository. This dataset contains information for 303 individuals with heart disease, characterized by 14 variables. These variables likely include factors such as age, blood pressure, and medical history, all relevant to predicting heart disease

B. Data Preprocessing

Raw data often requires preprocessing before feeding it intomachine learning models. To clean, transform, integrate, and reduce the dataset, several techniques were applied. Firstly, missing values were handled using the K-nearest neighbors (KNN) imputation technique provided by the KNN Imputer from scikit-learn. This technique allows for the estimationof missing values based on the values of neighboring data points. Additionally, feature scaling and encoding were performed on the dataset. Numerical features were scaled to ensure that all features are on a similar scale, which is crucial for many machine learning algorithms. Categorical variables were encoded into numerical representations to ensure effective processing by machine learning algorithms. These preprocessing steps help to ensure that the data is suitable for training machine learning models.

C. Data Splitting

The preprocessed dataset is then split into training and testing sets. The training set is used to train the machinelearning models, while the testing set is used to evaluate their performance using techniques such as k-fold cross-validation.

This division allows us to assess the performance and generalization ability of our models accurately. One common technique used for evaluating model performance on the testing set is k-fold cross-validation. In k-fold cross-validation,the training set is further divided into k subsets or folds. The model is trained k times, each time using k-1 folds for training and the remaining fold for validation. This process is repeated k times, with each fold serving as the validation set exactly once.

D. Machine Learning Models

The heart of this project lies in the machine learning models responsible for making predictions, particularly an ensemble learning technique known as soft voting classifier. Ensemble learning works by combining predictions from multiple base models to achieve higher accuracy and robustness. In this instance, the soft voting classifier merges predictions from two base models: Gaussian Naive Bayes and Random Forest Classifier. The Gaussian Naive Bayesmodel assumes independence between features, making it efficient for datasets with many features, while the Random Forest Classifier builds multiple decision trees on subsets of data and aggregates their predictions, leading to improvedgeneralization and reduced overfitting. The soft voting classifier combines these individual classifiers' predictions through averaging, leveraging their diverse perspectives to achieve superior performance compared to each classifier on its own.

4.3 Data Flow diagram

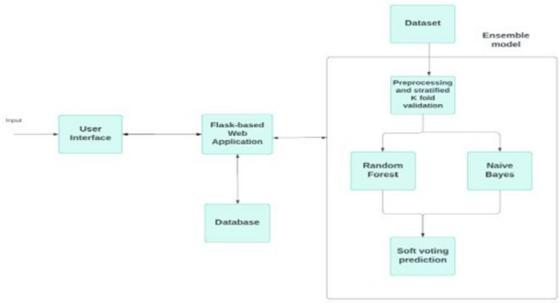


Figure 4.3: Data Flow Diagram

Fig. 3. Application Flow of our Proposed Model

E. Model Training and Evaluation

With the dataset preprocessed and the model selection made, the application proceeds to train each chosen model on the preprocessed heart disease dataset. During this phase, the models learn from the training data, adjusting their internal parameters to minimize prediction errors and improve their ability to generalize to unseen data. The training process is iterative, with each model gradually refining its predictive capabilities through repeated exposure to the training Dataset. Once the models are trained, the application evaluates their performance using robust evaluation techniques such as k-foldcross-validation. In k-fold cross validation, the dataset is partitioned into k equal-sized folds, with each fold servingas a validation set while the remaining folds are used for training. This process is repeated k times, ensuring that each data point is included in the validation set exactly once. By averaging the performance metrics obtained across multiple iterations of cross-validation, the application obtains a more reliable estimate of each model's predictive performance.

F. Visualization

The application goes beyond just training and evaluating machine learning models by offering insightful visualizations to enhance understanding and interpretation of their performance. These visualizations are crucial for stakeholders, providing deeper insights into the models' predictive abilities and aiding informed decision-making.

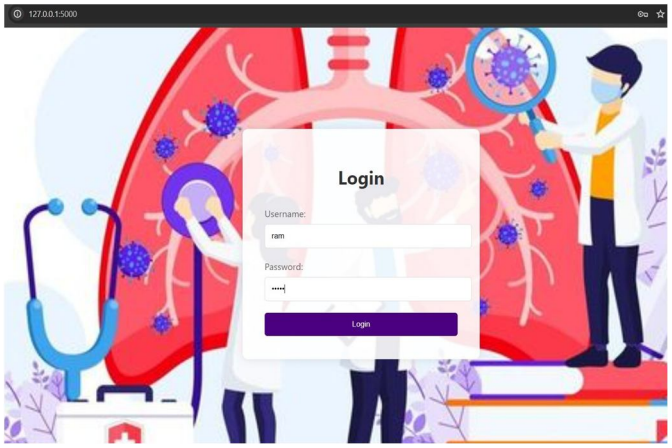


Fig. 4. Login Page of Proposed Model

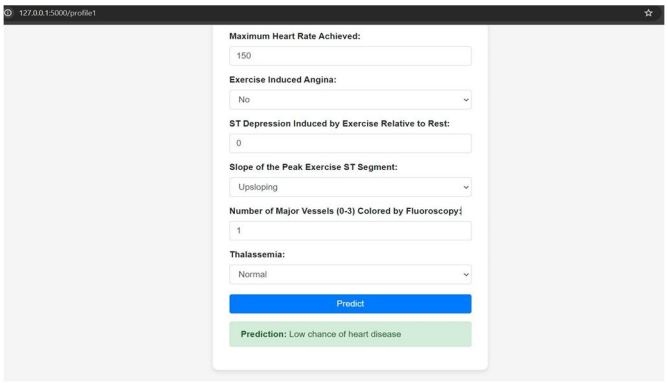


Fig. 5. Form of Proposed Model

Key visualizations include confusion matrices, which summarize the model's predictions by comparing actual and predicted class labels,and ROC curves, which illustrate the trade-off between true positive and false positive rates across different classification thresholds. These tools empower stakeholders to bettercomprehend and utilize the models' capabilities.

G. Model Integration And Deployment

Utilizing Flask as the underlying framework, the application leverages dynamic web page generation to provide users with an interactive and responsive user experience. Flask, being a lightweight and flexible web framework for Python, facilitates the seamless integration of machine learning models into web applications, enabling real-time predictions based on user input data.

H. Continuous Improvement

To ensure the effectiveness and relevance of the deployed heart disease prediction system, it's imperative to establish mechanisms for ongoing monitoring, feedback collection, and model adaptation. Continuous performance monitoring allows for the tracking of key metrics such as prediction accuracy, response times, and user satisfaction levels. The system should be designed to seamlessly integrate new data streams, enabling the continuous updating of machine learning models to adapt to emerging trends or variations in language. This iterative approach ensures that the models remain up-to-date and capable of capturing evolving patterns in heart disease diagnosis. The overall architecture and design of the system may vary based on specific project requirements, available resources, and the complexity of the analysis. However, the iterative process of experimentation, evaluation, and feedback-driven refinement is fundamental to building an effective and robust heart disease prediction system.

V. RESULTS

The ensemble model, which combines Gaussian Naive Bayes and Random Forest, achieved the highest accuracy,indicating the best performance. While the individual models, Gaussian Naive Bayes and Random Forest, performed well with an average accuracy of around eighty-four percent, the ensemble approach consistently outperformed them, highlight- ing the advantages of combining algorithms for improved statistical emphasis. The average accuracy for Naive Bayes is eighty-three point eight seven percent and for Random Forest is eighty-three point nine percent. The final precision of the ensemble model, rounded to two decimal places, was zero point eight-five.

Fold	GNB Accuracy	RF Accuracy	Ensemble Accuracy	F1 Score
1.0	81.25	68.75	81.25	0.8
2.0	75.0	75.0	81.25	0.82
3.0	81.25	87.5	81.25	0.73
4.0	73.33	80.0	73.33	0.67
5.0	86.67	86.67	86.67	0.86
6.0	100.0	93.33	100.0	1.0
7.0	100.0	93.33	100.0	1.0
8.0	86.67	100.0	93.33	0.93
9.0	100.0	86.67	100.0	1.0
10.0	80.0	80.0	80.0	0.73
11.0	80.0	93.33	80.0	0.77
12.0	86.67	80.0	93.33	0.93
13.0	80.0	86.67	80.0	0.77
14.0	86.67	73.33	80.0	0.77
15.0	66.67	66.67	66.67	0.62
16.0	86.67	80.0	86.67	0.83
17.0	93.33	100.0	93.33	0.92
18.0	60.0	66.67	60.0	0.62
19.0	86.67	86.67	86.67	0.83
20.0	86.67	86.67	86.67	0.83

Fig. 6. Evaluation Matrix

VI. FUTURE SCOPE

Success of the program in implementing Naive Bayes- Random Forest group. The cardiovascular prediction model in the Cleveland data set opens the doors to interesting possibilities of future development. First, clinical planning offers important opportunities. Rigorous testing and validationis necessary to ensure the model is developed with effortand reliability in real-world conditions. Seamless integration with existing It will facilitate user interaction and access to healthcare systems data medical experts. Moreover, itis closely monitored and investigated on the clinical side.Provision will be required to identify any issues and ensurethe preservation of the sample. It improves over time. The power of the model can be extended beyond cardiovascular predictions to cover a wide range of cardiovascular diseases.

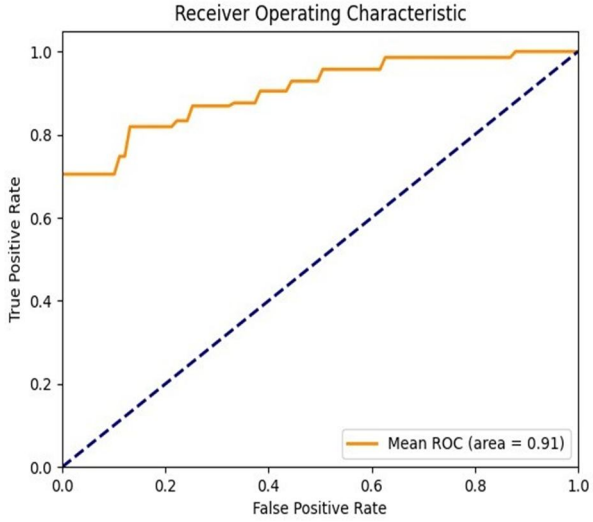


Fig. 7. ROC Curve

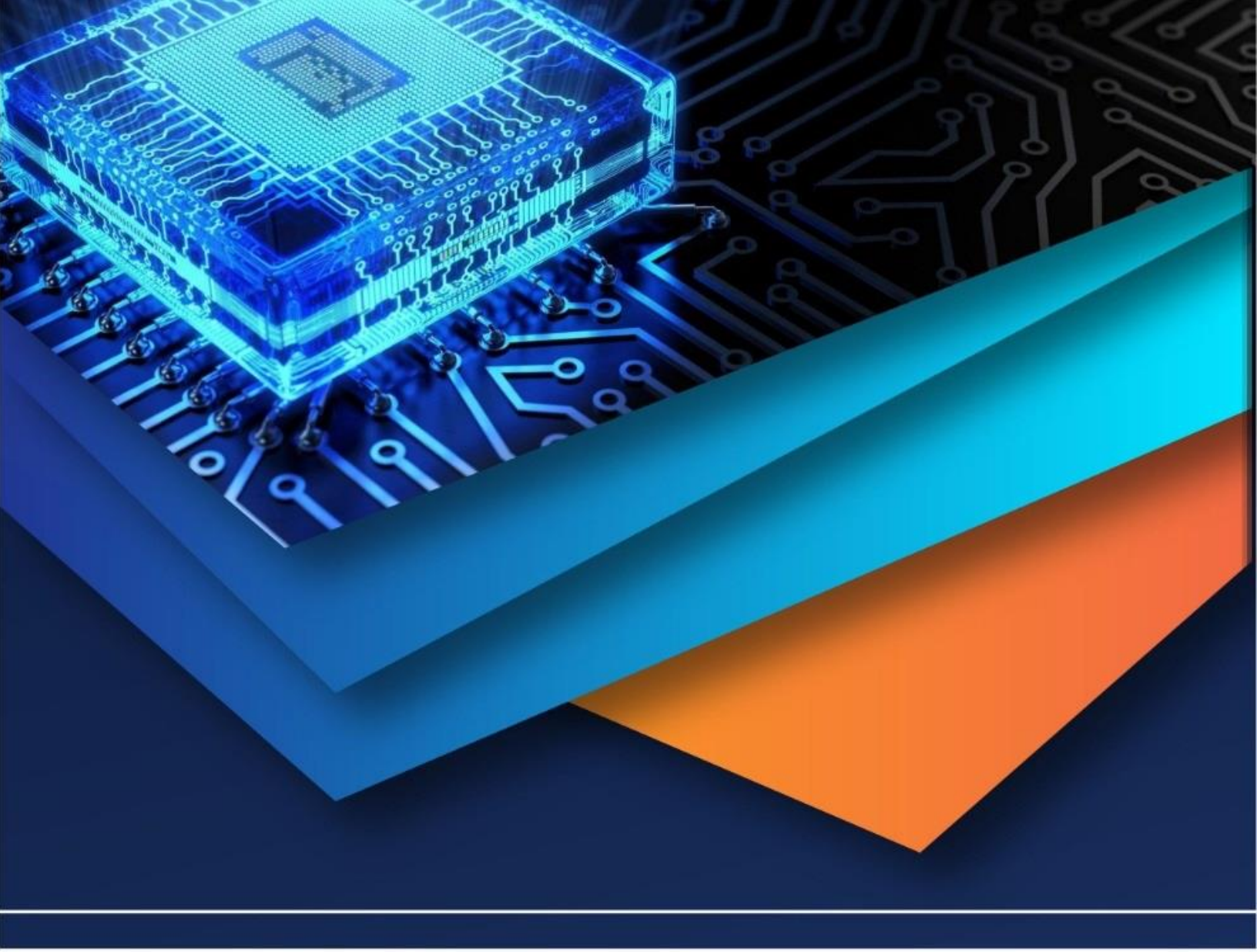
In addition to the Relevant data sets specific to other diseases such as stroke or hypertension, and optimization.The algorithm of the model to the specific symptoms of each disease, the task. It can provide comprehensive support tohealthcare professionals. This is an expansion. The specific risks will need to be carefully considered. Any cardiovascular disease. Finally, maintaining model accuracy and relevance over time. Continued improvement strategies are needed. Updating the model regularly given the need for new health information and research findings. And then he speaksof the challenge with data imbalances that are common real-world datasets is crucial.Examination of heterogeneous, disease-specific models. The future could bring even greater accuracy.

VII. CONCLUSION

This project successfully combined web developmentand tools to develop a user-friendly application for cardiac monitoring and assessing the risk of disease. Leveraging Flask for the web interface and scikit-learn for the machine learning algorithms, the application provides an effective platform for individuals to gain insight into their potential cardiovascular risk based on personalized treatment features. The project examined the effectiveness of group learning by incorporating predictions from Gaussian Naive Bayes and Random Forest distributions, achieving impressive results with an averageaccuracy of about eighty-four point eight five percent across all data types. While individual distributions such as Gaussian Naive Bayes and Random Forest performed well with average accuracy's of about eighty-three point eight seven percent and eighty-three point nine zero percent respectively, the ensemble model consistently outperformed them in most cases of clustered data. This highlights the potential of combining different algorithms to achieve high prediction accuracy. Furthermore, the ensemble model demonstrated consistent performance, exhibiting slightly higher F1 scores and accuracy's compared to individual distributions and clusters of trees. This robustness underscores the effectiveness of the model in analyzing cardiac disease risk, providing users with valuable insights, and opening up possibilities for integration into clinical settings through continuousimprovement and validation.

REFERENCES

[1] T. Christensen, A. Frandsen, S. Glazier, J. Humpherys and D. Kartchner,"Machine Learning Methods for Disease Prediction with Claims Data," 2018 IEEE International Conference on Healthcare Informatics (ICHI),
[2] New York, NY, USA, 2018, pp. 467-4674
[3] G. R. Thummala, R. Baskar and N. Thiyaneswaran, "Prediction of Heart Disease Using Naive Bayes in Comparison with KNN Based on Accuracy," 2022 International Conference on Cyber Resilience (ICCR),
[4] Dubai, United Arab Emirates, 2022, pp. 1-4
[5] N. Sabri et al., "HeartInspect: Heart Disease Prediction of an Individ-ual Using Na¨ıve Bayes Algorithm," 2023 IEEE 11th Conference on Systems, Process and Control (ICSPC), Malacca, Malaysia, 2023, pp. 350-354
[6] Khandaker Mohammad Mohi Uddin, Rokaiya Ripa, Nilufar Yeasmin, Nitish Biswas, Samrat Kumar Dey, Machine learning-based approachto the diagnosis of cardiovascular vascular disease using a combined dataset, Intelligence-Based Medicine, Volume 7,2023
[7] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Ef- fective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY