IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

Heart Disease Prediction System using Supervised Machine Learning Algorithms

Dr. K Satyanarayana Raju[1], N Rakesh[2], M Vinay[3], N Nikhil[4], M Chandra Kanth[5]
Assistant Professor, Department of Information Technology[1]
Students, Department of Information Technology[2,3,4,5]
S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh, India
rakeshnakka9999@gmail.com

Abstract: *This project aims to help prevent heart disease by predicting it early and recommending ways to reduce the risk. Heart problems can affect people of all ages nowadays, so prevention is really important. We're using information about your health and lifestyle to figure out if you're at risk of having heart issues soon. The good thing about our system is that we're using machine learning, which means we've trained computers to analyze a lot of data and make pretty accurate predictions about heart disease. However, the downside is that existing systems usually just tell you if you're at risk or not. They don't give you any advice on how to lower that risk. Plus, these systems aren't easily accessible for everyone to use.We're testing different models like Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, K-nearest Neighbor, and XGBoost to see which one works best for predicting heart disease.*
This way, we hope to make it easier for people to take control of their heart health and reduce the chances of having problems in the future. Additionally, we're planning to deploy our website so that everyone can access it easily. This means you won't need any special software or knowledge to use our system—it'll be available to anyone with an internet connection.

Keywords: heart disease

I. INTRODUCTION

Heart disease refers to conditions affecting the heart and blood vessels, a leading global cause of death. Risk factors include high blood pressure, high cholesterol, smoking, diabetes, obesity, and inactivity. Symptoms vary but may include chest pain, shortness of breath, and fatigue. Early detection and lifestyle changes can help manage and prevent heart disease.

Sudden cardiac arrest (SCA) and heart attacks are distinct but related cardiovascular events. SCA occurs when the heart suddenly stops beating, leading to a loss of blood flow and consciousness. It's often caused by an electrical malfunction in the heart, and immediate intervention with CPR and defibrillation is crucial for survival.

On the other hand, a heart attack, or myocardial infarction, results from a blockage in the blood vessels supplying the heart. This blockage can damage or destroy part of the heart muscle, causing symptoms like chest pain, shortness of breath, and nausea. Prompt medical attention is essential to minimize heart damage during a heart attack.

While both are serious, SCA is more abrupt and can be fatal within minutes, emphasizing the need for quick response, whereas a heart attack may involve more gradual symptoms, allowing for a window of intervention. Awareness, preventive measures, and emergency preparedness are vital in addressing these cardiovascular events.

II. LITERATURE SURVEY (RELATED WORK)

[1] The study **"Effective Heart Disease Prediction Using Machine Learning Techniques",** by Chintan Bhatt, Parth Patel, Tarang Ghetia, and Pier Luigi Mazzeo, published in 2023, focuses on utilizing machine learning techniques for heart disease prediction. The algorithms employed include Random Forest, Decision Tree Classifier, Multilayer Perceptron, and XGBoost. However, the study has limitations such as considering only a limited set of variables, not evaluating model performance on a held-out test set for generalizability, and not assessing the interpretability of

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

clusters formed by the k-modes algorithm. Despite these limitations, the research contributes to the field of heart disease prediction using machine learning techniques.

[2] The survey **"Heart Disease Prediction using Machine Learning Techniques: A Survey",** by V V Ramalingam and AyantanDandapath, published in 2018, examines the utilization of machine learning algorithms such as SVM, Random Forest, and Ensemble Models for heart disease prediction. Key limitations include overfitting issues with Decision Trees and challenges in handling high-dimensional datasets. Despite these limitations, SVM, Random Forest, and Ensemble Models demonstrated superior performance in predicting heart disease risk.

[3] The study **"A Web-based Heart Disease Prediction System using Machine Learning Algorithms"** by Md. Mahbubur Rahman, Morshedur Rahman Rana, Nur A Alam Munna, and Md. Saikat Islam Khan, published in 2022, develops a web-based system using eight ML algorithms and achieves high accuracy, especially with Decision Tree and Random Forest. The gaps identified include lower accuracy of some existing methods and the scope to improve accuracy further using more advanced techniques.

[4] In their study **"A Novel Web-Based Multi-Class Heart Disease Prediction Using Machine Learning Algorithms,"**Toqeer Ahmed and Saeed Mian Qaiser, published in 2022, introduce a novel approach to multi-class heart disease prediction utilizing a stacking ensemble classifier. The system achieved high accuracy, precision, and recall, with deployment via API and web interface using the Flask framework. The study's novelty lies in its multi-class classification approach and deployment methods. However, there are opportunities to enhance accuracy further and improve the web interface.

[5] The project **"Heart Disease Prediction using Machine Learning Algorithms,"** authored by Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain, and Preeti Nagrath, and published in 2021, developed a heart disease prediction system employing 3 ML classifiers on a small dataset. Among the classifiers, KNN demonstrated the highest accuracy at 88.52%. However, key limitations include the small dataset size, limited algorithms tested, absence of cross-validation, and binary prediction outcome. Expanding the dataset, exploring more algorithms, and predicting more granular outcomes could potentially improve the system's performance.

[6] The paper titled **"An Improved Auto Categorical PSO with ML for Heart Disease Prediction,"**authored by Animesh Kumar Dubey, Amit Kumar Sinhal, and Richa Sharma, and published in 2022, proposes a novel approach to improve the prediction of heart disease at an early stage. The study combines machine learning algorithms with the Improved Auto Categorical Particle Swarm Optimization (IACPSO) technique. IACPSO aids in selecting an optimal feature set, while ML algorithms categorize the data. This combined approach aims to enhance the accuracy and efficiency of heart disease prediction, particularly in identifying early stages. The study addresses the limitations of current diagnostic approaches and offers a potential solution for more effective detection of heart diseases.

[7] The paper titled **"An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches,"** authored by Subhash Mondal, Soumadip Ghosh, and Amitava Nag, and published in 2022, addresses key gaps in existing heart disease prediction models. The study introduces techniques to tackle overfitting and underfitting scenarios of machine learning (ML) models by employing hyperparameter tuning and a dual-stage stacking ensemble approach. Additionally, the study applies k-fold cross-validation to enhance generalization and mitigate overfitting. The proposed model achieves superior performance in terms of accuracy, recall, and ROC-AUC compared to previous studies utilizing the same datasets. Overall, the study contributes to advancing risk prediction models for heart diseases by addressing critical gaps in existing methodologies.

[8] The project **"Predict2Protect - Machine Learning Application in the Prediction of Heart Disease,"** authored by Ankita Mandal and published in 2023, utilizes four machine learning algorithms to predict heart disease risk, with the Decision Tree model exhibiting the best performance. However, the study identifies several limitations, including the reliance on Decision Tree, which may lead to overfitting, a limited prediction timeline, and the exclusion of certain factors that could potentially improve accuracy if incorporated. The study highlights the importance of addressing these limitations to enhance the accuracy and reliability of heart disease prediction models.

[9] The paper titled **"Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,"** authored by Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava, and published in 2019, proposes a hybrid machine learning method called HRFLM for heart disease prediction. The study demonstrates that HRFLM achieves

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

higher accuracy than existing methods when applied to the Cleveland UCI dataset. However, the study identifies several limitations, including the reliance on a single dataset, the potential for exploring more sophisticated machine learning approaches, and the need for evaluation on additional heart disease datasets. Overall, while HRFLM shows promise in heart disease prediction, further research is needed to validate its performance across diverse datasets. [10] The paper titled **"Heart Disease Prediction using Different Machine Learning Algorithms,"** authored by Rajani Pk, Kalyani Patil, Bhagyashree Marathe, Prerna Mhaisane, and Atharva Tundalwar, and published in 2023, presents a comparison of various machine learning algorithms for heart disease prediction. However, the study has several limitations, including the use of a small and generic dataset, limited algorithm evaluation, absence of hyperparameter tuning, lack of model evaluation on unseen data, and inadequate analysis of model performance. Strengthening the study with a larger and more diverse dataset, incorporation of advanced algorithms, optimization of model parameters, rigorous evaluation methodology, and deeper analysis of model performance could enhance the reliability and usefulness of the findings.

III. EXISTING SYSTEM

Existing systems for heart disease prediction leverage machine learning techniques that have been refined and validated using extensive datasets, enabling them to provide reasonably accurate forecasts regarding the probability of heart disease based on input data. These systems capitalize on algorithms such as Random Forest, XG-Boost, K-Nearest Neighbors (KNN), Logistic Regression, and Support Vector Machines (SVM), which have demonstrated efficacy in analyzing various factors associated with heart health. By analyzing a range of variables, including medical history, lifestyle habits, and physiological metrics, these systems offer valuable insights into the likelihood of heart disease occurrence.

However, a notable drawback of these existing systems is their limited scope, as they primarily focus on predicting whether an individual is at risk of heart disease or not, without offering specific recommendations for preventive measures to mitigate the risk. This lack of actionable advice can hinder efforts to proactively manage heart health and prevent cardiovascular issues. Furthermore, accessibility remains a challenge with these systems, as they are often not readily available for widespread use. This limitation restricts the potential impact of these predictive tools, preventing broader access to valuable insights and recommendations that could aid in early detection and prevention of heart disease. Thus, while existing systems provide valuable predictive capabilities, addressing these limitations is crucial to enhancing their utility and effectiveness in promoting heart health on a broader scale.

IV. PROPOSED SYSTEM

Our proposed system is dedicated to predicting whether an individual's heart is at risk or normal, while also providing personalized preventive measures to reduce the risk of heart disease. To ensure universal accessibility, we have deployed our system on pythonanywhere.com, a hosting platform that provides free lifetime access.

Python Anywhere allows us, as developers, to choose our preferred Python version and provides a comprehensive hosting service that includes a database, MySQL, and a Python terminal for installing any necessary libraries for our code. Once the website is deployed, users can access it by simply clicking on the shared website link.

Upon accessing the website, users will be directed to a homepage where they can navigate through various options. When they choose the "predict" option, they will be redirected to a dedicated page where they can input their information for heart disease prediction.

By leveraging PythonAnywhere's hosting solution, we ensure that individuals from all backgrounds have access to our heart disease prediction and prevention platform, empowering them to take proactive steps towards better heart health.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

V. SYSTEM ARCHITECTURE

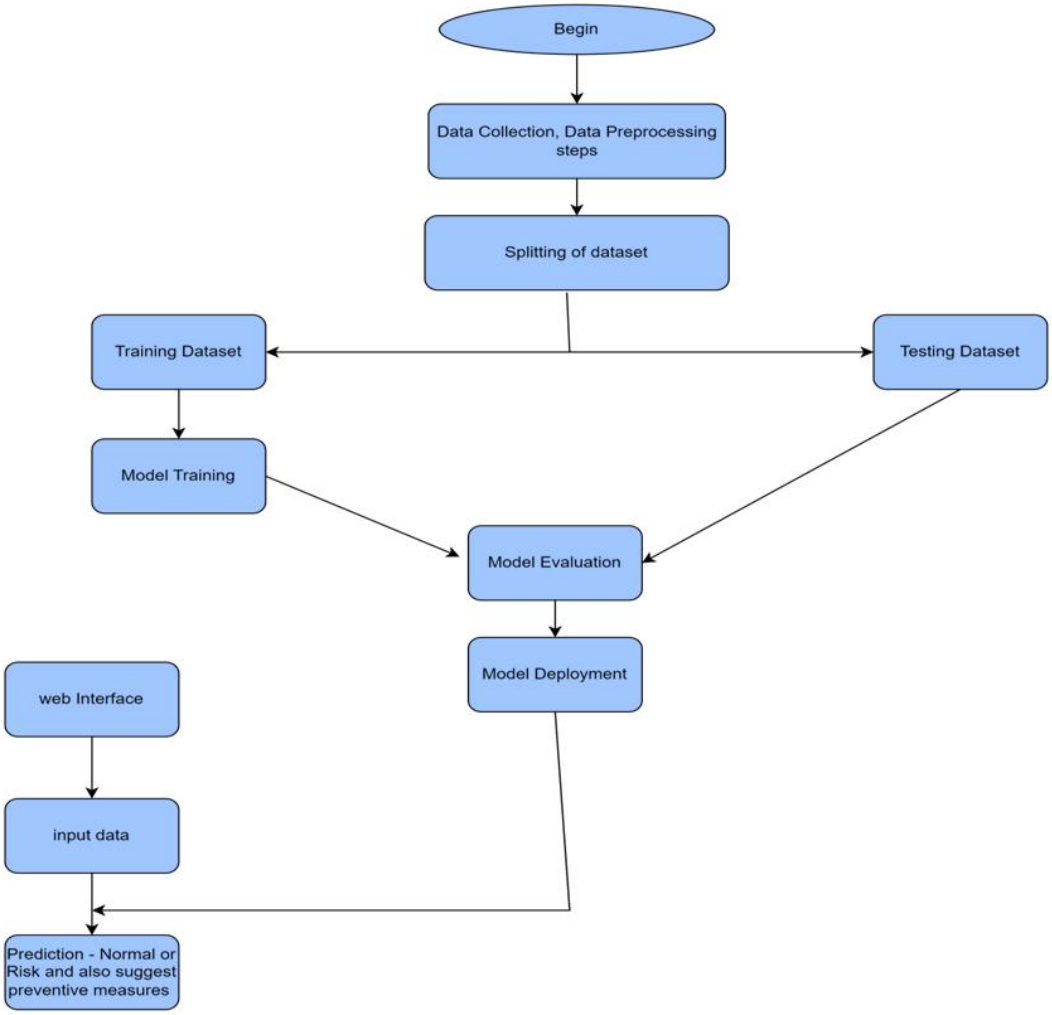


Fig 1 Architecture diagram

VI. METHODOLOGY

The various modules used in this paper are-

Sklearn- The open-source Python toolkit Scikit-Learn, sometimes referred to as "Sklearn" is used to create machine learning algorithms. For classification, clustering, and regression, there are numerous machine learning techniques. It enables the import of machine learning models and provides accuracy and confusion matrices.

Pandas - It is simple to use. It produces quick results and is also simple to comprehend. It is a free library that anyone can use. It has been used to analyse and explore the data

Pickle - The pickle module in Python is crucial for saving trained machine learning models as serialized objects, enabling easy storage and reuse of models without retraining.

NumPy - Essential for numerical computing, NumPy provides efficient array operations and mathematical functions, serving as a foundation for scientific computing tasks.

Matplotlib- A versatile plotting library, Matplotlib enables creation of various visualizations with customizable features, facilitating data exploration and presentation.

Seaborn- Built on Matplotlib, Seaborn simplifies statistical data visualization, offering attractive plots and advanced features for insightful data analysis.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

Steps:

Data Collection: Taken the dataset from Kaggle which consists of 13 features. They are : id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active, and cardio.

Dataset consists of 70,000 samples.

```
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           70000 non-null  int64
 1   age          70000 non-null  int64
 2   gender       70000 non-null  int64
 3   height       70000 non-null  int64
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64
 6   ap_lo        70000 non-null  int64
 7   cholesterol  70000 non-null  int64
 8   gluc         70000 non-null  int64
 9   smoke        70000 non-null  int64
 10  alco         70000 non-null  int64
 11  active       70000 non-null  int64
 12  cardio       70000 non-null  int64
dtypes: float64(1), int64(12)
```

Fig 2 Dataset Info

Data Cleaning and Preprocessing: Handle missing values, address outliers, and preprocess the data. This involves scaling numerical features, encoding categorical variables, and ensuring data quality.

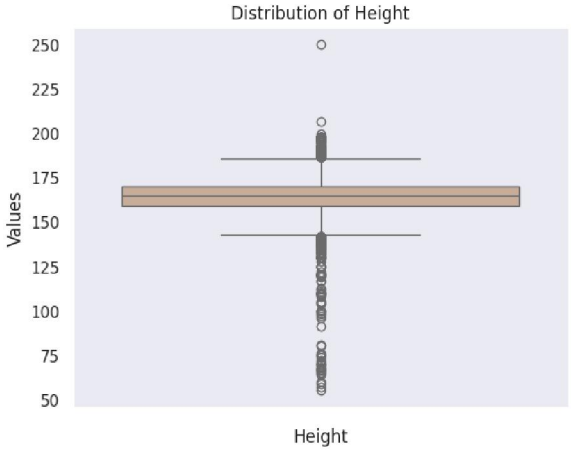


Fig 3 Before Removal of Outliers For Height

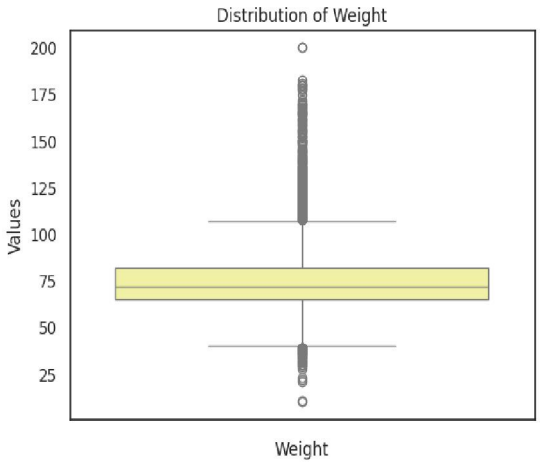


Fig 4 Before Removal of Outliers For Weight

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

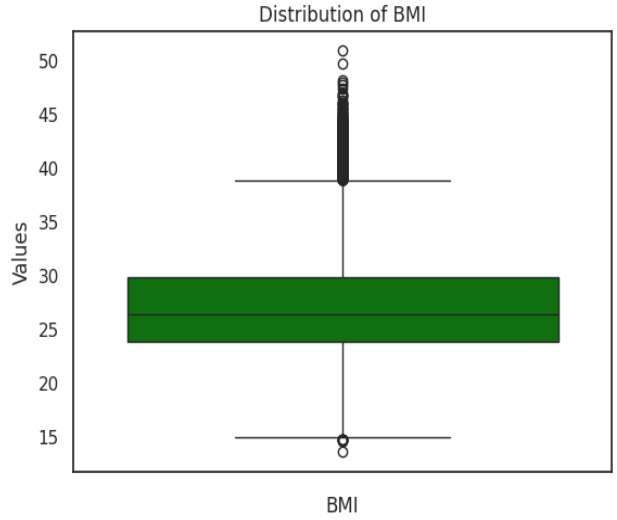


Fig 5 Before Removal of Outliers For BMI

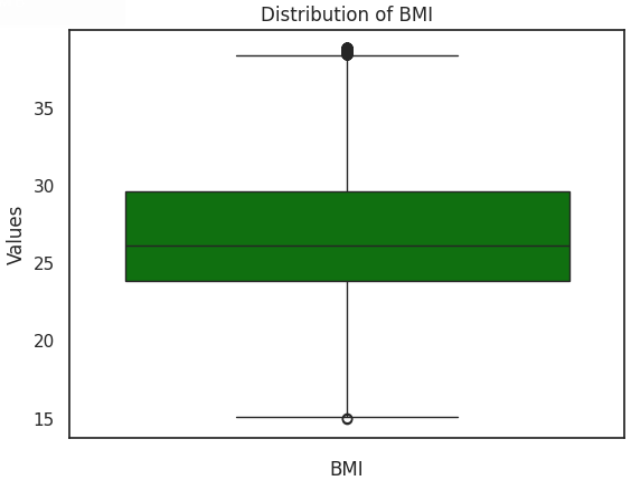


Fig 6 After Removal of Outliers For BMI

Exploratory Data Analysis (EDA): Analyze the dataset to gain insights into the distribution of features, identify patterns, and understand the relationships between variables.

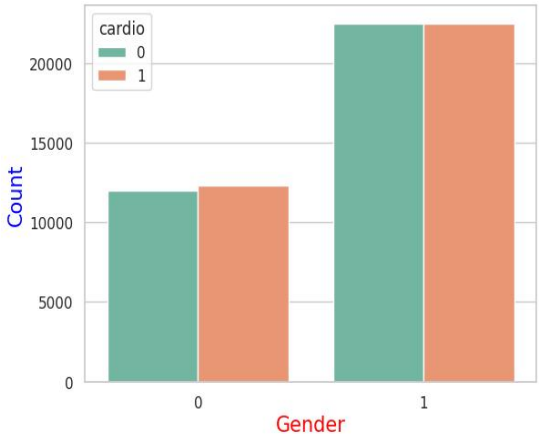


Fig 7 Distribution of Gender Vs Cardio

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

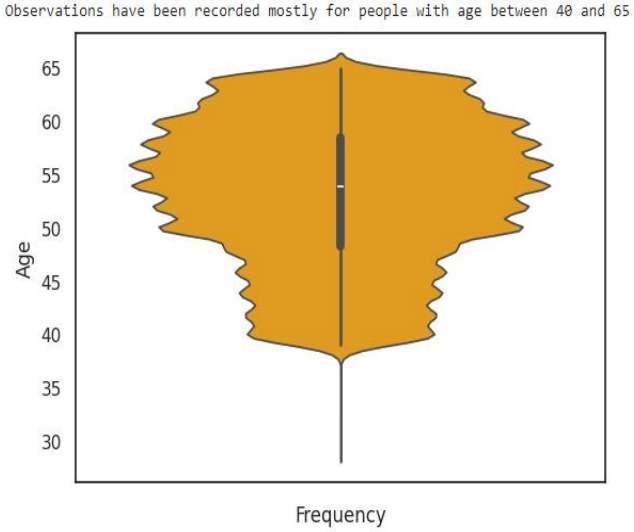


Fig 8 Violin Plot for Age

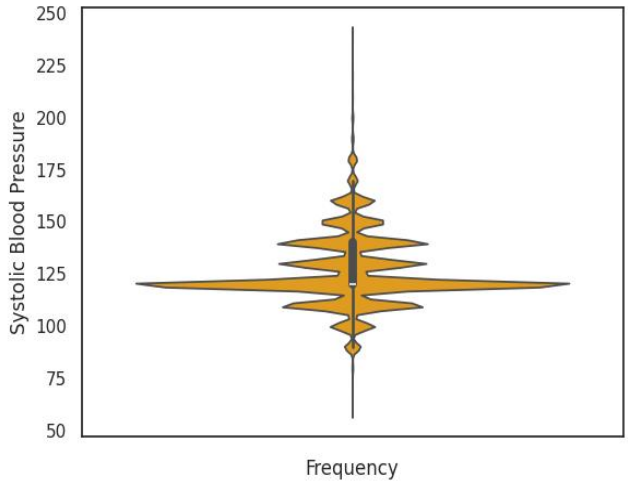


Fig 9 Violin Plot For Systolic Blood Pressure

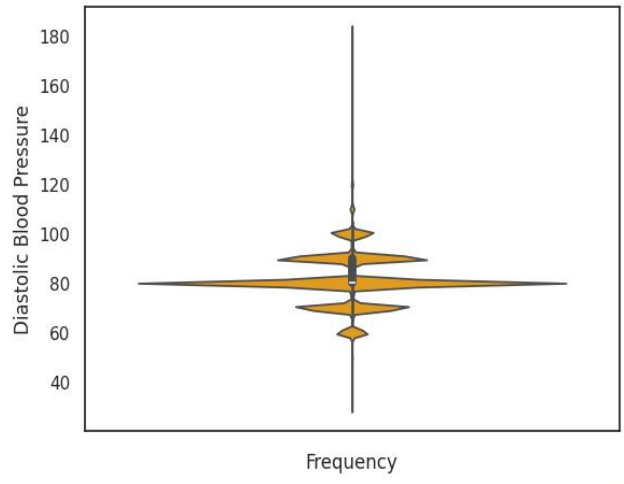


Fig 10 Violin Plot For Diastolic Blood Pressure

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

Feature Selection: Choose relevant features based on EDA and statistical methods to improve the model's performance.

Data Splitting: Divide the dataset into training and testing sets to train the machine learning models and evaluate their performance.

Splitting the dataset into 70:30 ratio.

70% of data is taken for training set.

30% of data is taken for testing set.

Model Training: Utilize machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, K-Nearest Neighbor, and XGBoost to train the models on the training dataset.

Model Evaluation: Assess the performance of each model using the testing dataset, employing metrics accuracy, F1 score.

Hyperparameter Tuning: Fine-tune the parameters of the selected models to optimize their performance.

Deploy the model- Created an website for the model's prediction and to take the data. we will consider the health and lifestyle related data to figure out if you're at risk of having heart issues soon or not.

VII. RESULT AND DISCUSSIONS

The results of our suggested system are displayed in this section; it performs faster, more accurately, and more reliably than the current system. A variety of machine learning techniques were used to get the outcomes.

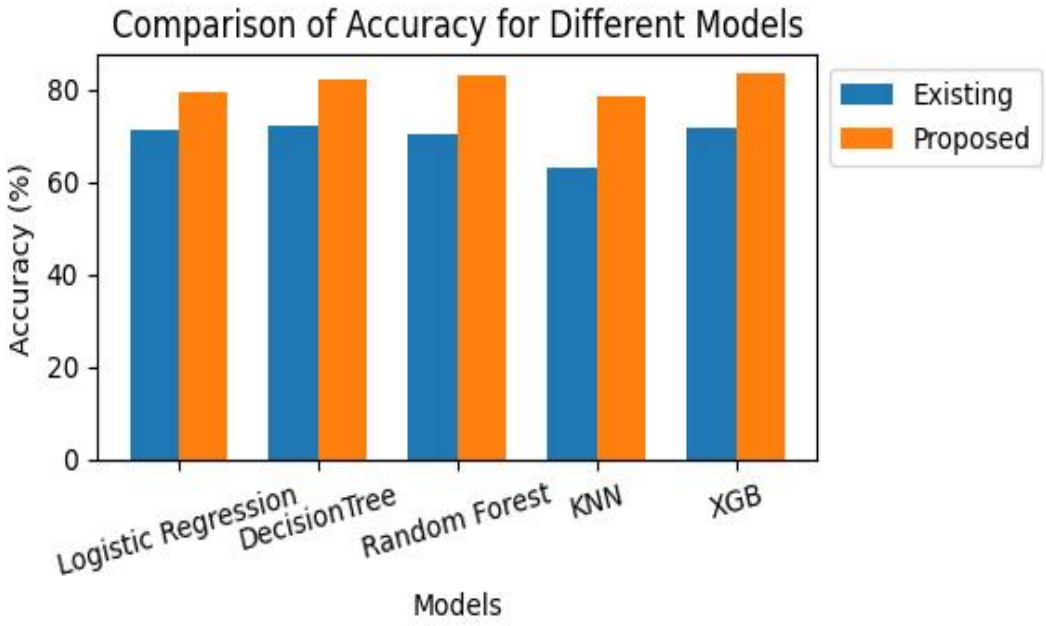


Fig 11 Comparison of Models Over Existing and Proposed

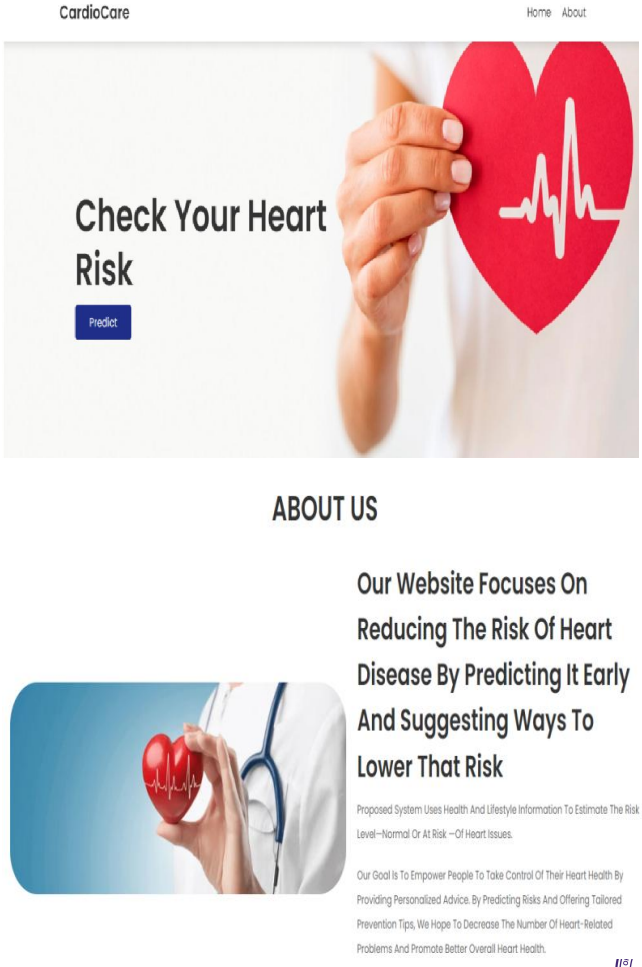
After careful consideration of the provided accuracy and F1-score metrics for each model, we have selected XGBoost, Support Vector Machines (SVM), and Random Forest as the top three models for further evaluation. XGBoost demonstrates the highest accuracy (82.95%) and F1-score (0.81), indicating robust performance across both metrics. SVM follows with an accuracy of 80.17% and an F1-score of 0.78, showcasing its strong predictive capabilities. Despite having slightly lower accuracy, Random Forest is chosen for its balanced performance, with an accuracy of 79.62% and an F1-score of 0.79, making it a reliable choice for classification tasks. These selections are based on the models' overall performance across multiple evaluation criteria, ensuring a comprehensive approach to model selection for heart disease prediction. which are listed in the table below.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

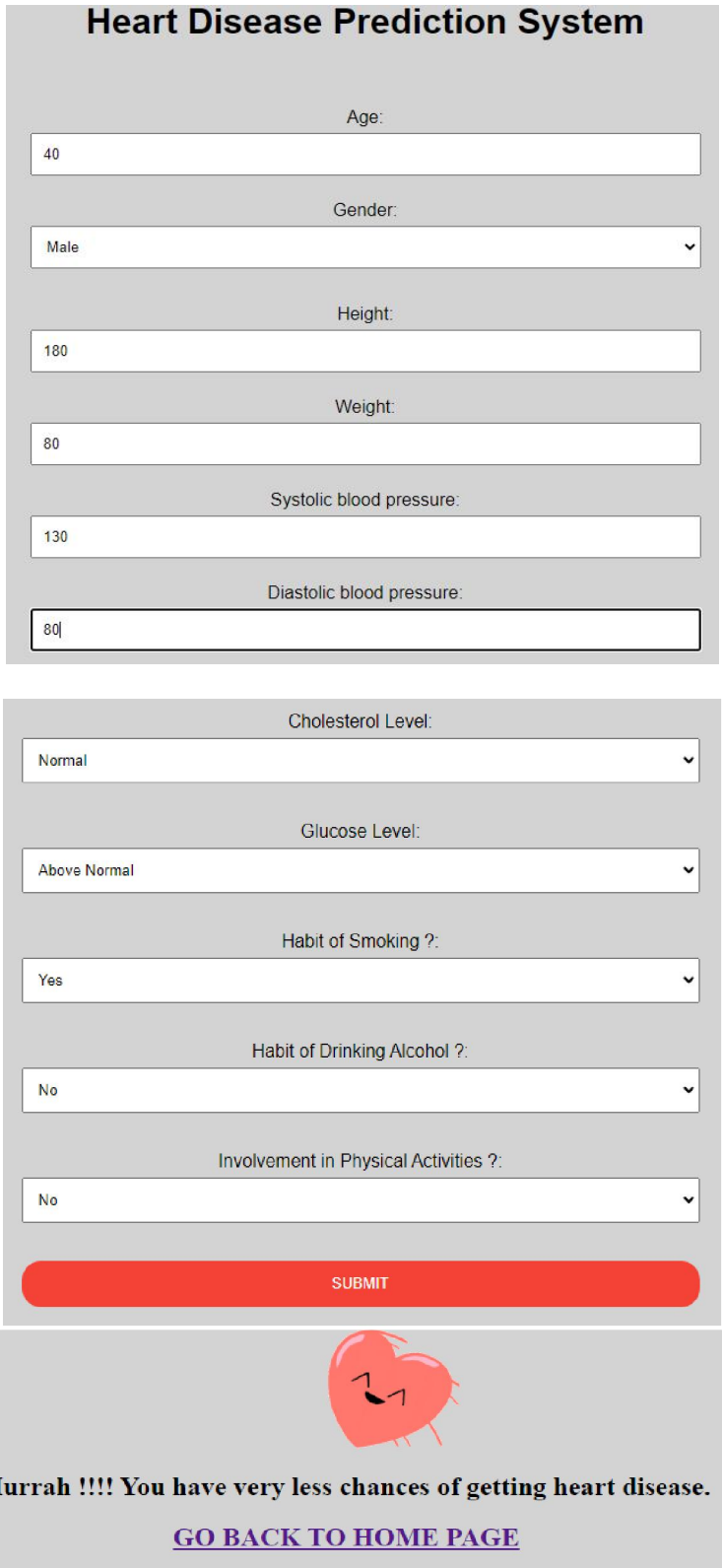
	Accuracy in %	F1-score
Logistic Regression	78.74	0.77
Decision Tree	73.31	0.73
Random Forest	79.62	0.79
Support Vector Machines	80.17	0.78
K-nearest Neighbors	78.42	0.78
XGBoost	82.95	0.81

Fig 12 Accuracy and F1-Score Table

In the proposed system, the utilization of advanced machine learning techniques, particularly XGBoost and Random Forest models, significantly enhances predictive accuracy compared to the existing system. Key to this improvement is the implementation of hyperparameter tuning through methods like GridSearchCV. By dynamically optimizing parameters such as 'max_depth', 'learning_rate', 'n_estimators', 'min_child_weight', 'subsample', and 'colsample_bytree', tailored toeach model, the system identifies the best parameter values for the given data. This fine-tuning process ensures more accurate predictions by optimizing model performance, making the proposed system a robust solution for heart disease prediction.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 5, April 2024

VIII. CONCLUSION

After performing hyperparameter tuning using GridSearchCV, the XGBoost (XGB) classifier emerged as the best model for heart disease prediction. With an accuracy of 83.26% and an F1-score of 0.82, the XGB classifier demonstrated superior performance compared to the Random Forest classifier.

The XGB classifier offers a balance of accuracy and precision, making it a reliable choice for predictive modeling in heart disease diagnosis. Following the tuning process, the optimized XGB classifier model was saved as a pickle file for future use.

Unlike existing systems that only provide outcome predictions without preventive measures, our project goes a step further. It not only categorizes the risk level (normal or at risk) but also offers preventive measures, empowering individuals to take control of their heart health. The proposed system uses a dataset from Kaggle with attributes like age, gender, height, weight, and various health metrics.

REFERENCES

[1] Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. J. Hepatol. 2018, 69, 896–904.

[2] Ramalingam, V V&Dandapath, Ayantan& Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering & Technology.7. 684. 10.14419/ijet.v7i2.8.10557.

[3] Rahman, Md & Rana, Morshedur& Munna, Nur & Khan, Md &Mohi Uddin, Khandaker Mohammad. (2022). A web-based heart disease prediction system using machine learning algorithms.12. 64-80.

[4] Ahmed, Toqeer& Qaisar, Saeed. (2023). A Novel Web-Based Multi-Class Heart Disease Prediction Using Machine Learning Algorithms.

[5] Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072

[6] Sharma, Richa &Sinhal, Dr. Amit & Dubey, Animesh. (2022). An Improved Auto Categorical PSO with ML for Heart Disease Prediction. Engineering, Technology and Applied Science Research.
12. 8567–8573. 10.48084/etasr.4854.

[7] Mondal, Subhash & Maity, Ranjan & Omo, Yachang& Ghosh, Soumadip& Nag, Amitava. (2024). An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches. IEEE Access. PP.1-1. 10.1109/ACCESS.2024.3350996.

[8] Mandal, Ankita. (2023). PREDICT2PROTECT - MACHINE LEARNING APPLICATION IN THE PREDICTION OF HEART DISEASE. International Journal of Advanced Research.11. 927-933. 10.21474/IJAR01/17783.

[9] M, S., C. T & S. G. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.” IEEE Access 7 (2019):81542–81554.doi: 10.1109/ACCESS.2019.2923707

[10] Pk, Rajani & Patil, Kalyani & Marathe, Bhagyashree &Mhaisane, Prerna &Tundalwar, Atharva. (2023). Heart Disease Prediction using Different Machine Learning Algorithms. International Journal on Recent and Innovation Trends in Computing and Communication.11. 354-359. 10.17762/ijritcc.v11i9s.7430.