IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION USING RANDOM FOREST

[1]Narra Lavanya, [2]Maddu Nithisha Swetha, [3]Karumuri Keerthi Sai, [4]Damerla Harshitha

[1,2,3,4] UG Students

[1]Information Technology,

[1]Vasireddy Venkatadri Institute of Technology, Nambur, India

Abstract: Heart disease remains a pervasive global health concern, maintaining its status as a primary cause of mortality. Early detection is pivotal for effective intervention strategies. This research leverages the Random Forest algorithm to predict the risk of heart disease, utilizing a comprehensive dataset encompassing patient attributes such as age, gender, cholesterol levels, blood pressure, and other pertinent clinical features. Renowned for its robustness and accuracy, the Random Forest model undergoes training on historical data to discern intricate patterns and relationships within the dataset. Through a meticulous evaluation process, we showcase the model's proficiency in accurately categorizing individuals into heart disease risk groups. This underscores its significance as a valuable tool for healthcare professionals in the assessment of patient cardiovascular health. This study represents a meaningful contribution to the ongoing endeavors aimed at harnessing machine learning to enhance heart disease prevention and elevate the standard of patient care.

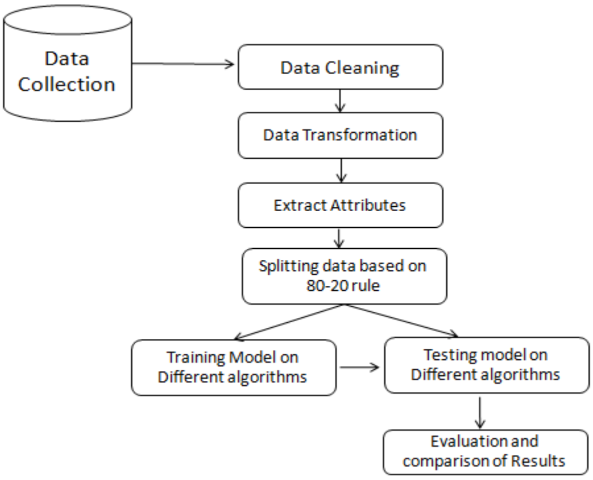
***Index Terms* - Machine Learning, Random Forest Algorithm, Risk Prediction and Accuracy.**

1. INTRODUCTION

Heart Diseases persist as a formidable global health challenge, claiming millions of lives annually. Timely and accurate identification of individuals at risk is paramount for implementing preventive measures and personalized interventions. In recent years, machine learning (ML) techniques have emerged as potent tools for harnessing insights from large and intricate datasets. Among these techniques, Random Forest, an ensemble learning algorithm, has shown promise in various domains due to its adaptability, interpretability, and robust performance. Random Forest, with its ability to handle high-dimensional datasets and capture non-linear relationships, emerges as a compelling candidate for enhancing the accuracy and interpretability of heart disease prediction models. In the dynamic landscape of heart health research, the ongoing evolution of predictive models demands adaptability to encompass emerging risk factors. Traditional parameters, though invaluable, may inadvertently overlook pivotal contributors to heart disease. In response, this study extends the framework of our previous heart disease prediction model, augmenting the dataset to include salient attributes such as smoking and alcohol intake, alongside the contemporarily significant parameter-details related to COVID-19. Understanding that risk prediction is inherently multifaceted, we acknowledge the evolving understanding of heart disease etiology. The inclusion of smoking and alcohol intake, recognized as influential lifestyle factors, serves to enrich the granularity of our predictive model. Furthermore, acknowledging the global health paradigm shift brought about by the COVID-19 pandemic, integrating pertinent details enables a more nuanced analysis of their potential impact on heart health. Our hypothesis posits that the inclusion of lifestyle factors and COVID-19 details will contribute significantly to the model's discernment of nuanced relationships, potentially influencing the accuracy of heart disease predictions. We delve into the intricacies of our methodology, present our findings in detail, and provide nuanced insights into the implications of our results, thereby enhancing the robustness of predictive models in cardiovascular health research.

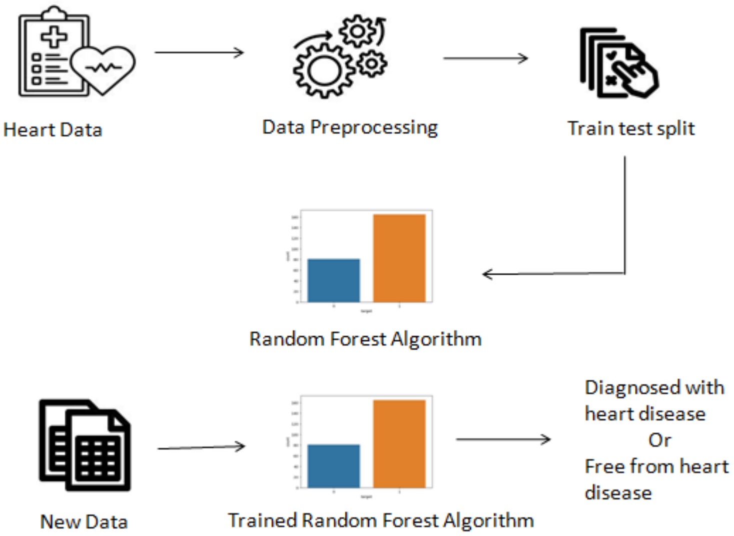
2. EXISTING SYSTEM

In this system, we take data from the patients and we will use different types of ml techniques to obtain results like data preprocessing, feature scaling, model building, and one hot coding. In this we will use different types of python modules like pandas, NumPy to visualize the data. We use five different models and take the best one with high accuracy to obtain our final result.


Data Collection
Data Cleaning
Data Transformation
Extract Attributes
Splitting data based on 80-20 rule
Training Model on Different algorithms
Testing model on Different algorithms
Evaluation and comparison of Results

3. PROPOSED SYSTEM

In the context of the heart disease prediction model, it is imperative to acknowledge the evolving landscape of risk factors. Consequently, our updated dataset incorporates crucial parameters such as smoking habits, alcohol intake, and relevant COVID-19 details. Recognizing the pivotal role these attributes play in Heart health, their inclusion aims to enhance the predictive accuracy of our model. As we integrate these additional parameters, we anticipate a refinement in the precision and comprehensiveness of our heart disease prediction system. This nuanced approach reflects our commitment to advancing the efficacy of predictive analytics in the domain of Heart health. These adjustments, rooted in contemporary medical insights, underscore the ongoing evolution of our proposed system for heart disease prediction, positioning it at the forefront of informed and progressive healthcare methodologies



4. RESEARCH METHODOLOGY

4.1 Data Collection

Accumulation of a comprehensive dataset inclusive of diverse patient records. Incorporation pertinent features, including traditional risk factors (e.g., age, blood pressure, cholesterol levels) and novel contributors (e.g., smoking habits, alcohol intake, and COVID-19 details.

4.2 Data Preprocessing

Cleaning and handling missing values to ensure data integrity. Standardization or normalization of numerical features to bring them to a common scale. Encoding categorical variables for compatibility with the Random Forest algorithm.

4.3 Feature Selection

Identification of relevant features through techniques such as feature importance ranking. Emphasis on including newly added parameters (smoking, alcohol intake, and COVID-19 details) as potential discriminators in the model.

4.4 Training the Random Forest Model

Division of the dataset into training and testing sets to facilitate model evaluation. Training the Random Forest ensemble on the training data, where multiple decision trees are constructed with different subsets of the data and features.

4.5 Model Evaluation

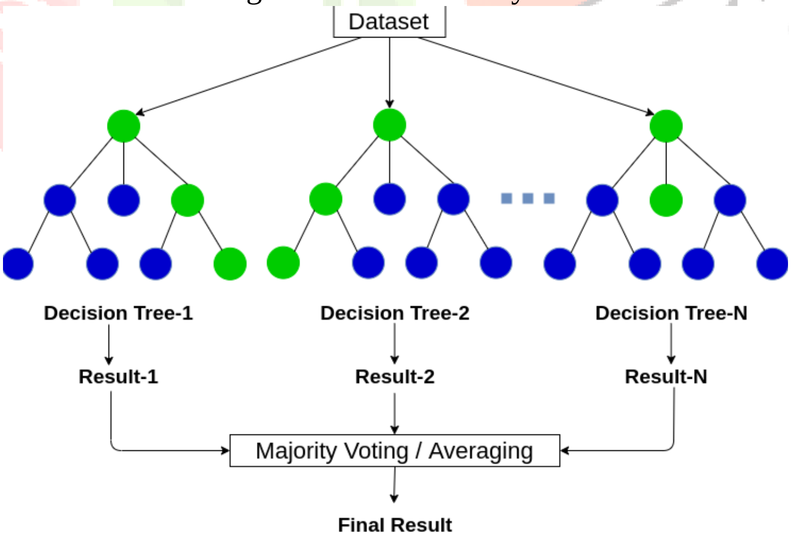
Assessment of the model's performance using metrics such as accuracy. Validation of the model on the testing dataset to ensure its generalization to unseen data.

4.6 Prediction and Interpretation

Application of the trained model to make predictions on new, unseen data. Interpretation of the model's decisions, leveraging feature importance scores to understand the relative impact of different factors on heart disease prediction.

5. RANDOM FOREST ALGORITHM

Random Forest is a flexible and extensively used machine learning technique that may be used for classification, regression, feature selection, and other applications. It is especially beneficial when you need a strong and robust model that can handle large datasets with many characteristics.



5.1 ADVANTAGES

- High Accuracy
- Reduced Over fitting
- Feature Importance
- Robustness
- Ease of use

6. LIBRARIES

6.1 NumPy (Numerical Python)

It was designed to accommodate huge, multi-dimensional arrays and matrices, as well as a set of high-level mathematical functions for operating on these arrays. NumPy is the Python basis for scientific computing and data analysis, allowing for efficient and quick numerical calculations.

6.2 Pandas

It was created to give Python with simple data structures and data analysis capabilities. Pandas have two basic data structures: Data Frames for dealing with structured data and Series for dealing with one-dimensional data, making it a useful library for data manipulation and analysis.

6.3 Matplotlib

It was created in Python to produce static, animated, and interactive visualizations. Matplotlib is a popular and adaptable data visualization library that provides great control over plot components.

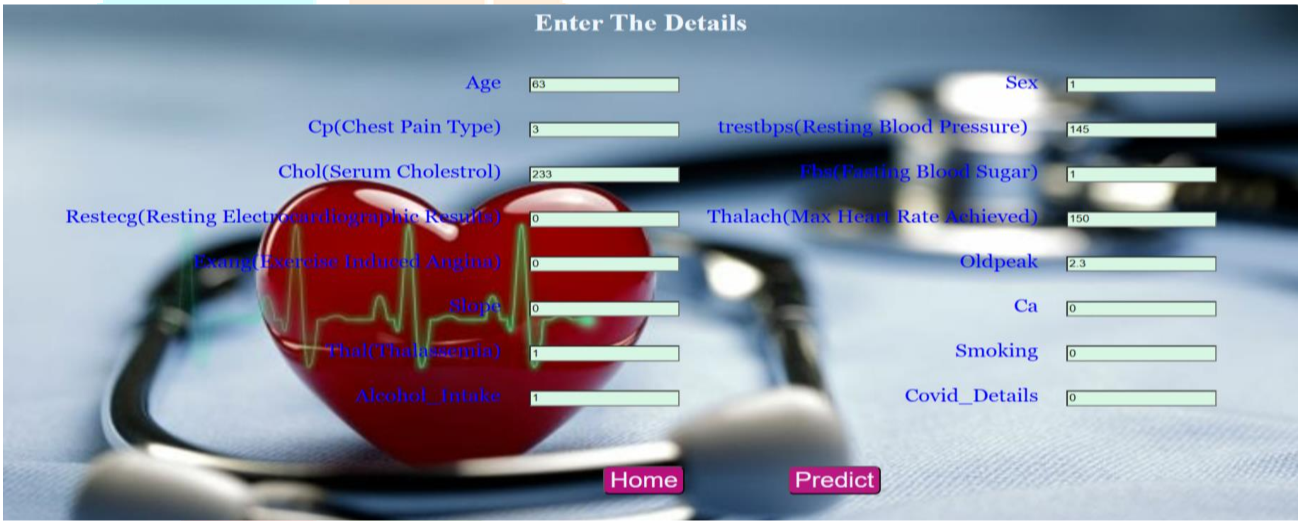
6.4 Seaborn

The name Seaborn is not derived from anything specific, but rather symbolizes the company's desire to be visually appealing and focused on statistical data presentation. It is based on Matplotlib and offers a high-level interface for producing useful and appealing statistical visualizations. Seaborn makes it easy to create complicated visualizations and is especially well-suited for working with statistical data and datasets.

6.5 Sklearn

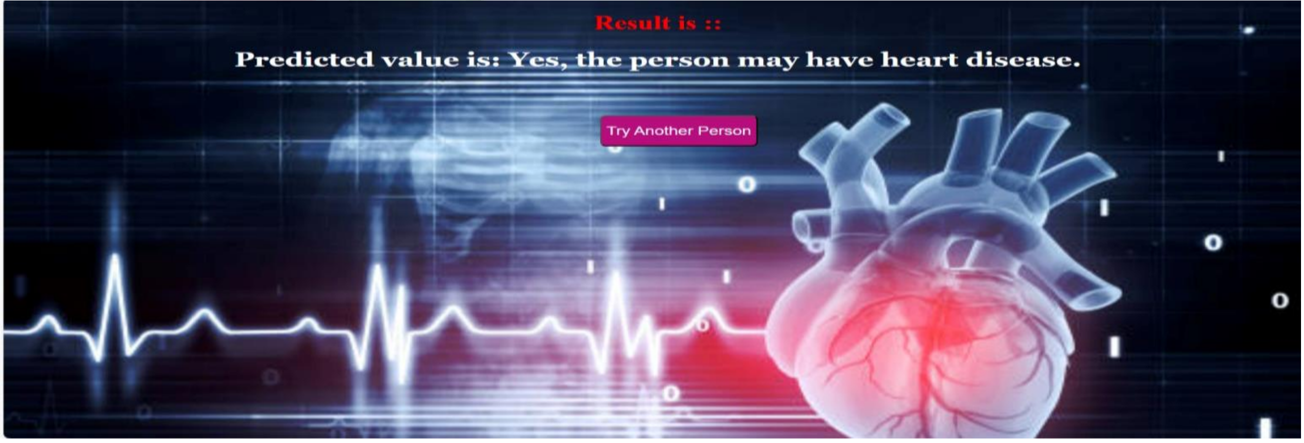
Sklearn, sometimes known as "scikit-learn," is a popular Python machine learning package. It is free and open-source, and it is built on top of well-known libraries like NumPy, SciPy, and Matplotlib.

7. RESULT



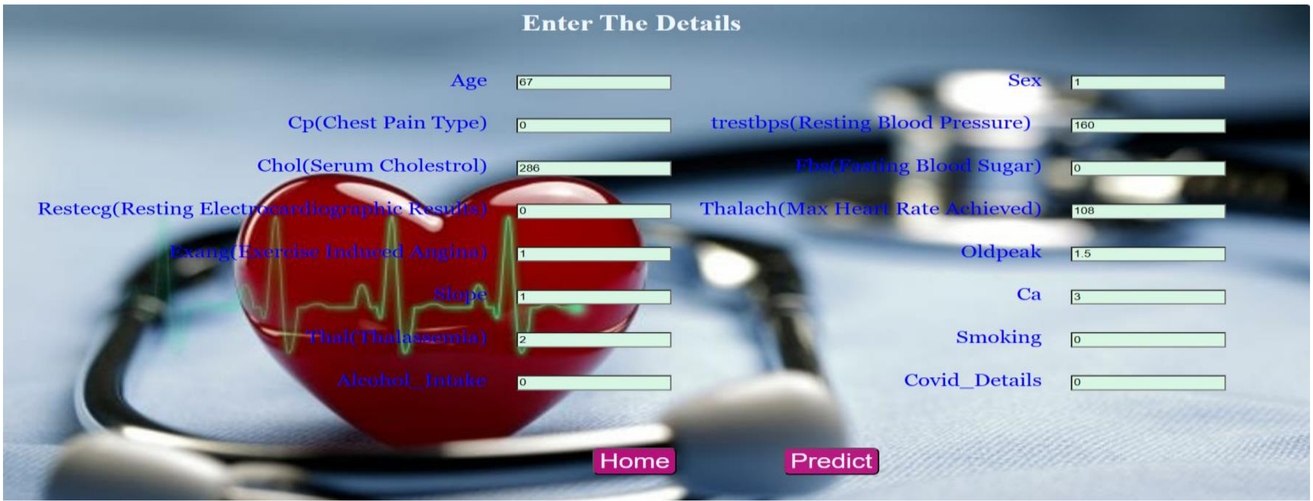
Enter The Details

Age	63	Sex	1
Cp(Chest Pain Type)	3	trestbps(Resting Blood Pressure)	145
Chol(Serum Cholestrol)	233	Fbs(Fasting Blood Sugar)	1
Restecg(Resting Electrocardiographic Results)	0	Thalach(Max Heart Rate Achieved)	150
Exang(Exercise Induced Angina)	0	Oldpeak	2.3
Slope	0	Ca	0
Thal(Thalassemia)	1	Smoking	0
Alcohol_Intake	1	Covid_Details	0



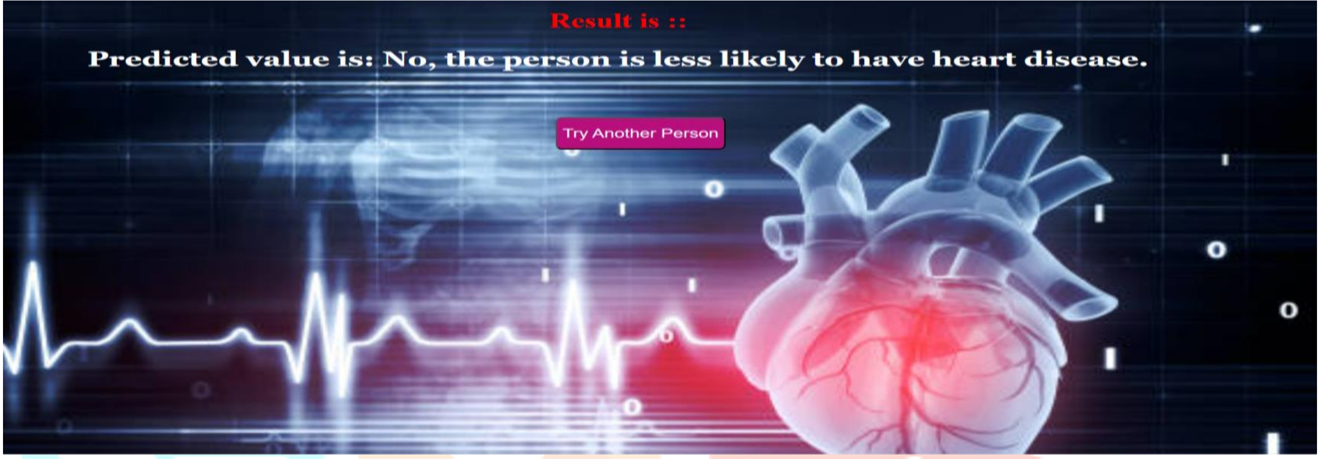
Result is ::

Predicted value is: Yes, the person may have heart disease.



Enter The Details

Age	67	Sex	1
Cp(Chest Pain Type)	0	trestbps(Resting Blood Pressure)	160
Chol(Serum Cholestrol)	286	Fbs(Fasting Blood Sugar)	0
Restecg(Resting Electrocardiographic Results)	0	Thalach(Max Heart Rate Achieved)	108
Exang(Exercise Induced Angina)	1	Oldpeak	1.5
Slope	1	Ca	3
Thal(Thalassemia)	2	Smoking	0
Alcohol_Intake	0	Covid_Details	0



Result is ::

Predicted value is: No, the person is less likely to have heart disease.

8. CONCLUSION

In our pursuit of refining heart disease prediction models, this study ventured beyond conventional parameters by augmenting the dataset with influential attributes such as smoking, alcohol intake, and COVID-19 details. These additions were prompted by the evolving understanding of Heart risk factors, acknowledging the contemporary relevance of lifestyle choices and the impact of recent health events. Our initial model, trained on a dataset that lacked these crucial attributes, showcased a commendable accuracy. However, the subsequent enhancement, achieved through the integration of lifestyle and COVID-19 data, propelled the Random Forest model to a refined accuracy. This substantial improvement underscores the pivotal role of these attributes in augmenting the predictive power of the model. The nuanced analysis of the refined model's performance revealed its heightened ability to discern intricate relationships within the data. Lifestyle factors, such as smoking and alcohol intake, emerged as influential determinants, shedding light on their significant roles in cardiovascular risk. Furthermore, the inclusion of COVID-19 details provided valuable insights into the interplay between recent health events and heart disease, emphasizing the need for adaptive models in the era of evolving health landscapes. Our findings not only contribute to the growing body of knowledge in Heart health but also advocate for a paradigm shift towards more holistic risk assessment. The model's adaptability to contemporary risk factors positions it as a valuable tool in clinical decision-making, empowering healthcare professionals with insights that extend beyond traditional predictors. In conclusion, this study not only demonstrates the tangible impact of attribute integration on model accuracy but also underscores the importance of adapting predictive models to reflect the dynamic nature of health considerations. The comprehensive approach presented herein serves as a foundation for future research endeavors, advocating for precision and relevance in heart disease prediction models.

REFERENCES

1. Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, SagarPande, Parneet Singh, Heart Disease Prediction Using a Combination of ML, Volume7,Publish Year: 2021.
2. Mohd Faisal Ansari, BhavyaAlankar, HarleenKaur,Heart Disease Prediction Using a Combination of ML ,Volume -6,Publish Year: 2020.
3. T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Map reduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN:2278- 3075, Volume-8 Issue-3, January 2019.
4. A. S. Abdullah and R.R.Rajalaxmi, ``A data mining model for predicting the coronary heartdisease using random forest classi_er," in Proc. Int. Conf. Recent Trends Comput. Methods,Commun. Controls, Apr. 2012, pp. 22_25.
5. A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ``Using PSO algorithmfor producing best rules in diagnosis of heart disease," in Proc. Int. Conf. Comput. Appl. (ICCA),Sep. 2017, pp. 306_ 311.
6. N. Al-milli, ``Back Propagation neural network for prediction of heart disease," J. Theor.Appl.Inf. Technol., vol. 56, no. 1, pp. 131_135, 2013.
7. C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika,``Analysis of neural networks based heart disease prediction system," in Proc. 11th Int. Conf.Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233_239.
8. P. K. Anooj, ``Clinical decision support system: Risk level prediction of heart disease usingweighted fuzzy rules," J. King Saud Univ.- Computer. Inf. Sci., vol. 24, no. 1, pp. 27_40, Jan. 2012.doi: 10.1016/j.jksuci.2011.09.002.
9. L. Baccour, ``Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets," Expert Syst. Appl., vol. 99, pp. 115_125, Jun. 2018. doi:10.1016/j.eswa.2018.01.025.
10. C.-A. Cheng and H.-W. Chiu, ``Anarti_cial neural network model for the evaluation of carotid artery stenting prognosis using a nationalwide database," in Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017, pp. 2566_2569.