Heart Disease Prediction

Nitish kumar Singh

Date: 05.03.2024

(TASK-0)

Abstract:

The escalating global incidence of heart disease underscores the urgency for early prognosis and lifestyle interventions to mitigate cardiovascular risks. This paper presents a statistical model for heart disease prediction, utilizing basic patient health parameters to aid medical professionals in forecasting heart disease. Employing three machine learning classifier models - Logistic Regression, K-Nearest Neighabors, and Random Forest - clinical features crucial for heart disease diagnosis are analyzed using the University College Irvine (UCI) Dataset. By assessing the accuracy of these models, this study contributes to enhancing predictive capabilities in heart disease prognosis, thereby facilitating proactive healthcare interventions. Keywords: Data Mining, Machine Learning, Logistic Regression, KNN, Random Forest, Heart Disease Prediction

INTRODUCTION:

The prevalence of cardiovascular diseases (CVD) has surged at an alarming rate, posing significant challenges in terms of prediction, management, and prevention. Traditional methods of diagnosis and risk assessment often fall short due to human error and limitations in processing vast amounts of data. Consequently, early detection becomes imperative to mitigate the escalating risks associated with CVD.

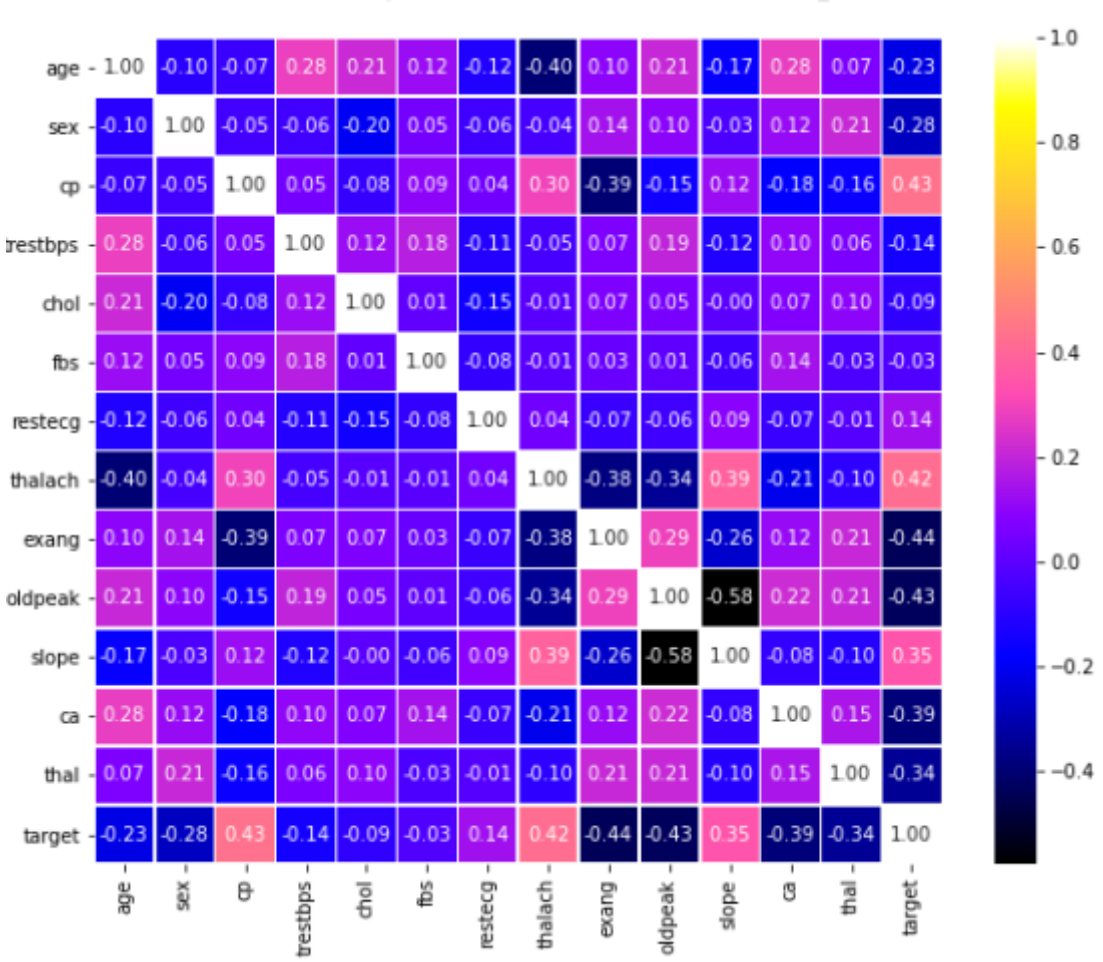
Data mining, coupled with deep learning techniques, has emerged as a powerful tool in forecasting various scenarios across diverse fields. Leveraging this approach, we aimed to develop a predictive model using the University College Irvine (UCI) Dataset for Heart Disease research. By employing three distinct machine learning classifier algorithms - K-Nearest Neighbor

Classifier, Logistic Regression Classifier, and Random Forest Classifier - we sought to enhance the accuracy and reliability of heart disease prediction.

The escalating burden of CVD, responsible for approximately 1.5 million deaths annually worldwide, underscores the urgency for accurate prediction and risk assessment [1]. Alarmingly, 82 percent of premature deaths occur in low and middle-income countries, with cardiovascular diseases contributing significantly to the mortality rate [1]. Early identification of risk factors and lifestyle behaviors associated with CVD is crucial for effective prevention and intervention strategies.

Machine learning algorithms play a pivotal role in the medical sector, utilizing vast databases and patient histories to predict and diagnose illnesses. Parameters such as age, blood pressure, cholesterol levels, and obesity serve as critical indicators for heart disease prediction. Various machine learning techniques, including KNN, decision trees, logistic regression, and random forest classifiers, are employed to analyze these variables and predict the likelihood of heart disease.

In this study, we conducted exploratory data analysis (EDA) to understand the nature of the dataset and applied standardization techniques to address missing data. Subsequently, the dataset was split into training and testing sets to train and evaluate the performance of the predictive models. Preliminary analysis revealed key triggers of heart disease, such as age, cholesterol levels, and old peaks, as evident from the heatmap analysis.



Heatmap of Different parameters of CVDs

In our study, we utilized logistic regression classifier for its reliance on predefined variables to predict outcomes based on patient medical reports. This model outputs the risk of heart attack based on input parameters and previously trained results, necessitating a high degree of precision. Additionally, the K-Nearest Neighbors (KNN) algorithm was employed during model training. KNN operates by determining the nearest neighbors based on distance metrics, allowing our model to iteratively refine predictions without repeatedly recalculating results.

A heatmap provides a graphical representation of data using color-coding, facilitating the visualization of relationships and trends. In our analysis, the heatmap (Fig.1) illustrates the interplay of various parameters in cardiovascular disease (CVD). Key features such as age, chest pain type (cp), ST depression induced by exercise relative to rest (slope), maximum heart rate achieved (thalach), and exercise-induced angina (exang) emerge as significant

predictors of heart disease risk. This visualization aids in identifying critical factors influencing CVD, contributing to the development of more accurate predictive models.

Problem Statement:

Early detection of heart disease is challenging due to its complex etiology and the presence of multiple risk factors. Traditional risk assessment methods often rely on individual risk factors such as age, gender, blood pressure, cholesterol levels, and smoking status. However, these approaches may lack accuracy and fail to capture the interactions between different risk factors. There is a need for a more comprehensive and accurate predictive model that can leverage diverse data sources to identify individuals at high risk of heart disease.

Market/Customer/Business Need Assessment of Heart Disease Prediction:

Market Assessment:

- **Prevalence Analysis:** Globally, cardiovascular diseases (CVDs) are the leading cause of death, with an estimated 17.9 million deaths annually, according to the World Health Organization (WHO). Locally, data from healthcare organizations and government agencies can provide insights into the prevalence of heart disease in specific regions.
- **Market Landscape:** Assess existing heart disease prediction tools and technologies, including traditional risk assessment methods and emerging predictive analytics solutions. Identify key players in the market, such as healthcare providers, medical device manufacturers, and technology companies.
- **Regulatory Compliance:** Understand regulatory requirements and standards, such as HIPAA in the United States or GDPR in Europe, governing the collection, storage, and analysis of patient data for heart disease prediction.

Customer Assessment:

- **Healthcare Providers**: Conduct interviews or surveys with cardiologists, primary care physicians, and other healthcare professionals to understand their pain points and challenges in accurately predicting heart disease. Identify specific needs, such as real-time risk assessment tools or predictive models tailored to specific patient populations.
- **Patients**: Gather feedback from patients through focus groups or online surveys to understand their preferences and expectations regarding heart disease prediction tools. Explore factors such as ease of use, accessibility, and trust in the technology.

Business Need Assessment:

- **Financial Implications**: Estimate the costs associated with developing and implementing heart disease prediction solutions, including data acquisition, software development, and training for healthcare professionals.
- **Return on Investment (ROI)**: Analyze the potential ROI for healthcare providers investing in predictive analytics for heart disease. Consider factors such as reduced hospital admissions, improved patient outcomes, and increased efficiency in resource allocation.
- **Revenue Generation**: Identify opportunities for revenue generation, such as licensing predictive models to healthcare organizations or offering predictive analytics as a service.
- **Scalability and Sustainability**: Evaluate the scalability and sustainability of the business model for heart disease prediction solutions. Consider factors such as market demand, technological advancements, and competition from other predictive analytics tools.

Target Specifications and Characterization:

1. **Healthcare Providers:**
    - **Cardiologists:** Specialists in diagnosing and treating heart conditions, including heart disease. They require accurate and reliable predictive models to assist in risk assessment and treatment planning.
    - **Primary Care Physicians:** Frontline healthcare providers who often encounter patients with risk factors for heart disease. They need user-friendly tools that can quickly assess patients' risk and guide appropriate interventions.
    - **Healthcare Institutions:** Hospitals, clinics, and healthcare systems are interested in implementing scalable and cost-effective solutions to improve patient outcomes and operational efficiency.
2. **Patients:**
    - **Adults at Risk:** Individuals with risk factors for heart disease, such as high blood pressure, high cholesterol, obesity, diabetes, and a family history of heart disease. They seek tools that can help them understand their risk and take preventive measures.
    - **Elderly Population:** Aging adults are at increased risk of heart disease and may benefit from personalized risk assessment tools to guide lifestyle modifications and medical interventions.

- **Technology-Aware Individuals:** Patients who are comfortable with technology and prefer digital health solutions for managing their health and wellness.

3. **Characteristics and Preferences:**
    - **Accuracy:** Healthcare providers and patients require predictive models that are highly accurate in assessing the risk of heart disease to ensure appropriate interventions are implemented.
    - **Ease of Use:** User-friendly interfaces and intuitive design are essential for widespread adoption among healthcare providers and patients.
    - **Interpretability:** Healthcare providers prefer predictive models that provide clear explanations of risk factors and recommendations to facilitate informed decision-making.
    - **Accessibility:** Solutions should be accessible across various devices and platforms, including desktop computers, tablets, and smartphones, to accommodate different user preferences and workflows.
    - **Privacy and Security:** Patients and healthcare providers prioritize data privacy and security, requiring solutions to comply with regulatory standards such as HIPAA and GDPR to protect sensitive health information.
    - **Cost-Effectiveness:** Healthcare institutions seek cost-effective solutions that provide value in terms of improved patient outcomes, reduced healthcare costs, and increased operational efficiency.

Literature Review:

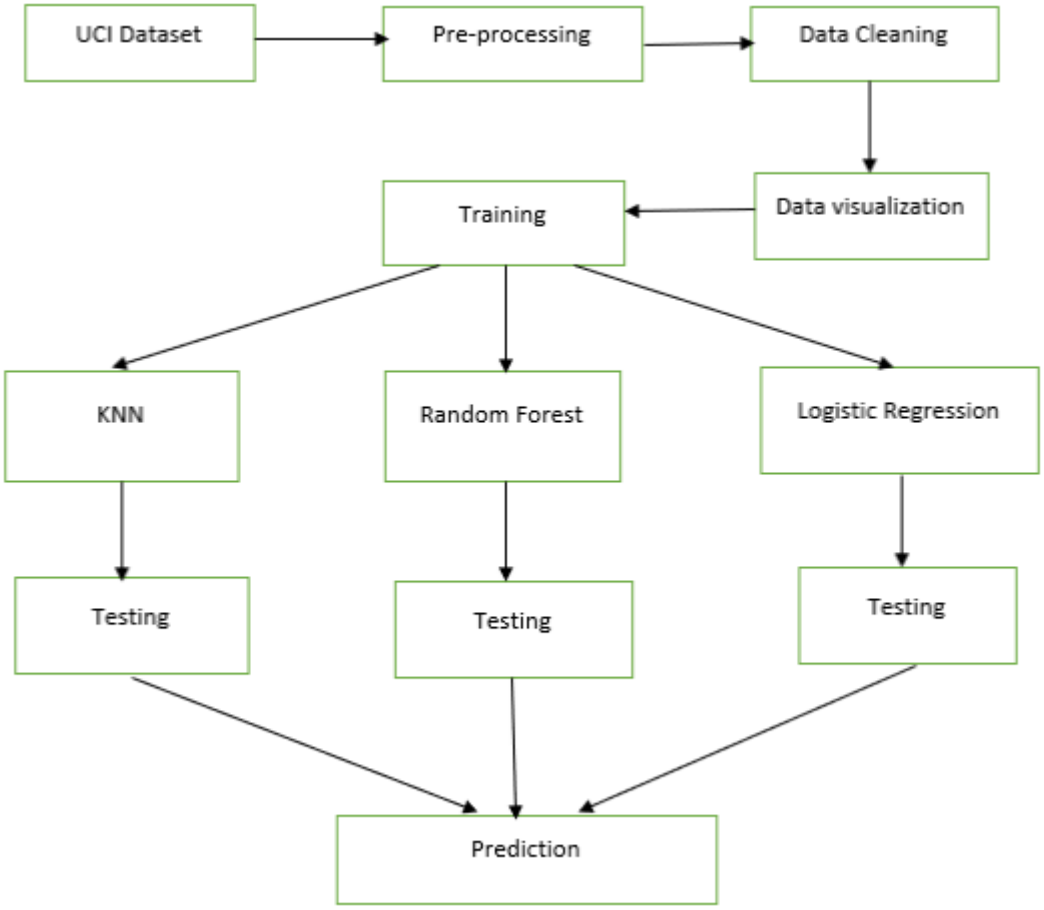
Current research focuses on enhancing the prediction of cardiovascular diseases using various methods to improve efficiency and precision across different parameters. In a 2013 study, a heart disease prediction model utilized decision tree algorithms and various machine learning techniques such as ID3 Naïve Bayes, Gain Ratio DT, ART kernel density, bagging algorithm, and SVM. While aiming for greater accuracy, challenges arose due to distortions caused by DTs of the Gain Ratio.

Similarly, in the same year, another study explored heart disease estimation using a combination of SVM, DT, and logistic regression on the CHDD dataset. Although the method leveraged DT's rule-based algorithm incorporating classification, regression, and correlation principles, it faced limitations in ensuring better outcomes and yielded order-based results.

With technological advancements, machine learning (ML) has emerged as an innovative field. In a recent 2018 study, an innovative ML algorithm for heart disease prediction was

introduced. This approach utilized the MLP algorithm, enhancing efficiency and precision. However, limitations were observed in MLP's performance under extreme levels.

These studies demonstrate ongoing efforts to improve heart disease prediction through innovative techniques, highlighting both advancements and challenges in the field of machine learning for healthcare applications.



Model Flowchart

In recent years, advancements in heart disease prediction have been driven by various machine learning algorithms. A 2019 study introduced an improved approach combining Random Forest and Linear systems, showcasing its application on raw data. However, its theoretical nature limits practical implementation. Another 2019 research explored SVM, DT, Logistic Regression, and Naïve Bayes individually on a UCI dataset, aiming for enhanced precision but faced challenges with model complexity and delayed results.

Comparative research from 2011 highlighted the efficacy of decision tree and Bayesian algorithms in heart disease prediction, especially when combined with genetic algorithms to optimize data size. Subsequent studies, including one in 2013 using data mining techniques

on the Chandigarh dataset, demonstrated the effectiveness of algorithms in identifying specific heart diseases.

In 2014, a framework combining neural networks and fuzzy logic showed promising precision, particularly suitable for physicians as a predictive tool. However, further validation on larger datasets with more parameters is warranted to ensure accuracy. Additionally, a prototype developed in 2013 aimed to train nurses and physicians in disease prediction, providing not only predictive capabilities but also explanatory insights. Though precise, its limited dataset and testing pose constraints on its reliability. Overall, these studies underscore ongoing efforts to enhance heart disease prediction through diverse methodologies, emphasizing the need for rigorous validation and practical applicability.

Clinical features	Description
Num	Diagnosis of heart disease
Exang	Exercise-induced angina
Age	Instance age in years
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect
Restecg	Resting electrocardiographic results
Cp	Chest pain type
Ca	Number of significant vessels (0-3) colored by fluoroscopy
FBS	Fasting blood sugar
Sex	Instance gender
Slope	The slope of the peak exercise ST segment
Thalach	Maximum heart rate achieved
Trestbps (mmHg)	Resting blood pressure
Oldpeak	ST depression induced by exercise relative to rest
Chol (mg/dl)	Serum cholesterol

In 2011, a web-based, handy, and accurate prediction framework for heart disease was proposed[10]. On the UCI machine learning dataset, it implemented a weighted related classifier algorithm. It is for general research purposes, as the UCI dataset has been used. Symptoms do, however, vary from place to place. So, for practical use, it should be applied on the local dataset. The J48, REPTREE, and SIMPLE CART [11] were used to construct another prediction model. The methods were applied to the South African medical practitioner's collected patient dataset. The results obtained from these techniques of classification were accurate. In another study for cardiac prediction analysis[12], data mining techniques were also applied. This research shows that the various algorithms have distinct precision, but the algorithm of the neural network gives the highest precision and then trees of decision. When combined with a genetic algorithm, the algorithm becomes more efficient.

Business Model:

1. **Subscription Model for Healthcare Providers:**
    - Offer subscription-based access to the heart disease prediction platform for healthcare providers, including hospitals, clinics, and medical practices.
    - Charge a monthly or annual fee based on the number of users or volume of predictions generated.
    - Provide tiered subscription plans with varying levels of features and support.
2. **Pay-per-Use Model for Individual Users:**
    - Allow individual users, such as patients or individuals at risk of heart disease, to access the prediction platform on a pay-per-use basis.
    - Charge a fee for each prediction generated, offering flexibility for users who may not require frequent predictions.
    - Provide discounted rates for bulk purchases or subscription plans for regular users.
3. **Freemium Model with Premium Features:**
    - Offer a basic version of the heart disease prediction platform for free to attract users and build a user base.
    - Introduce premium features, such as advanced analytics, personalized risk assessments, and real-time monitoring, available through paid subscriptions.
    - Utilize the freemium model to upsell premium features and convert free users into paying customers.
4. **Licensing Model for Healthcare Institutions:**
    - License the heart disease prediction platform to healthcare institutions, allowing them to integrate the predictive model into their existing systems.
    - Charge a one-time licensing fee or annual royalties based on the institution's size and usage volume.
    - Provide training, support, and customization services as additional revenue streams.
5. **Data Insights and Analytics Services:**
    - Offer data insights and analytics services to healthcare providers and research institutions based on the aggregated data generated by the prediction platform.

- Charge a fee for access to advanced analytics dashboards, custom reports, and predictive modeling services.
- Monetize data insights by offering anonymized datasets to pharmaceutical companies, research organizations, and government agencies for research purposes.

APPROCH:

The proposed methodology for heart disease prediction begins with obtaining the publicly available UCI dataset. The data is then subjected to cleaning and normalization as part of the preprocessing step to ensure consistency and prepare it for analysis. Subsequently, the cleaned and normalized data is visualized to identify trends and relationships among attributes.

After preprocessing, the dataset is divided into training and testing sets. Machine learning algorithms are applied to the training dataset to train the predictive model. The trained model is then evaluated using the testing dataset to assess its performance and predictive accuracy.

A comparison of various classification algorithms implemented in the model is conducted to determine the most effective approach. The entire workflow of the methodology is depicted in Fig, providing a visual representation of the research process.

- ## UCI Dataset

  The UCI dataset's open-source dataset registry is used in analysis. It has numerous disease-related databases. For academic purposes, these freely accessible databases are used. In this model, the UCI dataset of heart disease is used. It is the dataset existence category data category. With 303 instances and 75 properties, it is a multivariate dataset. The dataset includes both knowledge that is helpful and attributes that are not useful. So, in pre-processing the useful data is selected and data cleaning is done to remove the null values

- ## Pre-Processing

  As at this pre-processing level, this is the important step; meaningful data is derived from the dataset of heart disease. This phase is compulsory because the raw data is not reliable and unfinished, so pre-processing is performed for more steps to render ready raw data. In this approach, the UCI Heart Disease Dataset, the data contains 75 attributes, and during pre-processing, 14 [6] attributes are extracted to understand the nature of patients' health better. The extricated 14 attributes include BP, sex, heart rate, chest, and others. The attribute’s values are normalized and converted into numerical form.

- ## Data Clean

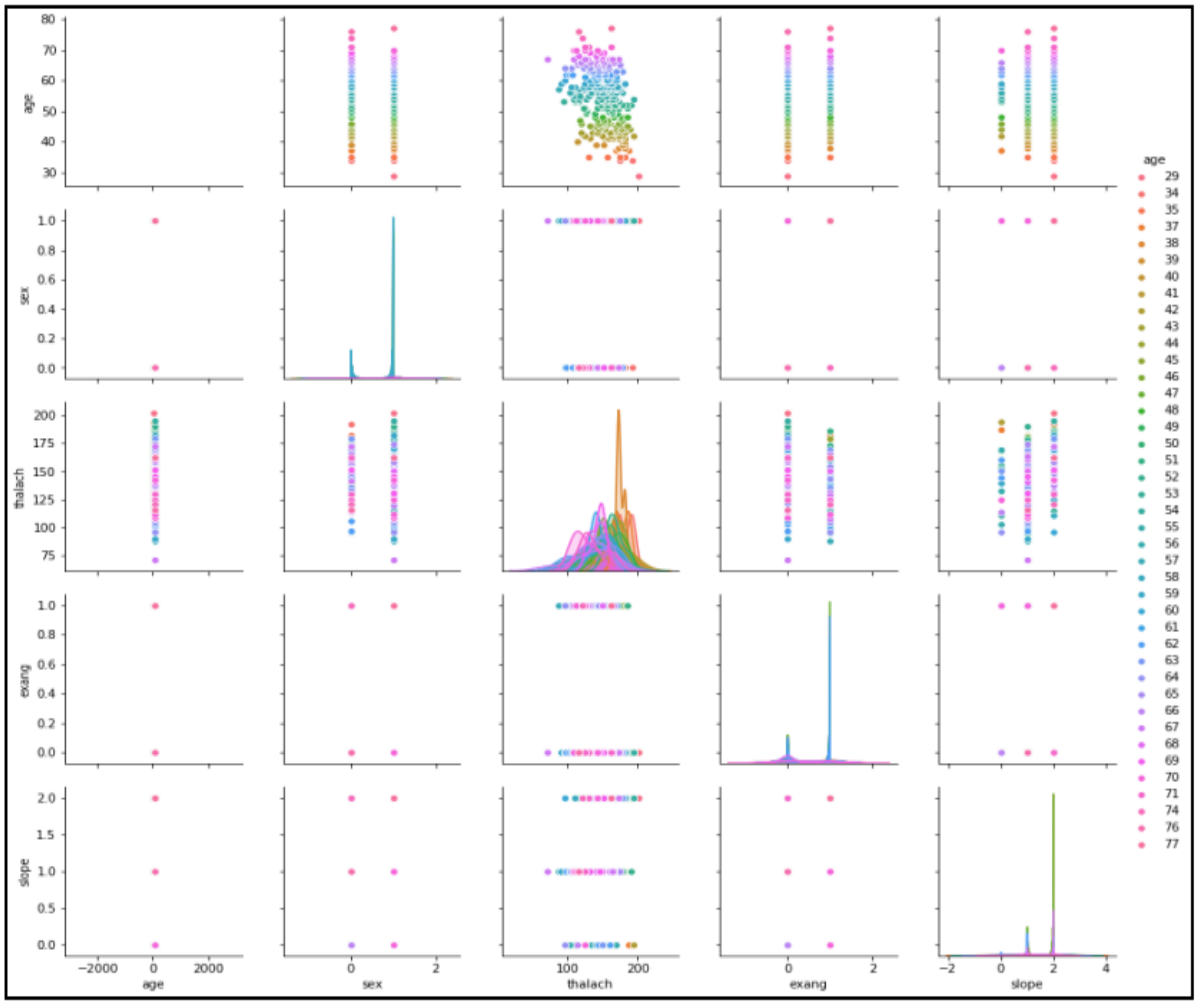
  The quality of data plays an essential role, and the most carefully depicted thing to be. For this research, data cleaning has improved the quality of our dataset. Data cleaning is necessary as it

removes unnecessary or irrelevant attributes of data from the dataset. This step of the model will make the dataset more precise and exact. In this part of approach, the Null (NaN) values are removed from the dataset to make it more useful as these values decrease the productivity of the algorithm [7]. At the data cleaning stage, the dataset is also normalized to not have any ambiguity after cleaning

Visualization:

The dataset is in tabular form and it is hard to observe and understand the data in this or any other form. So the data is visualized Graphically, as seen in Fig.3 below. It helps in knowing the trend of the data. Data visualization in this approach is a graphical representation of the data. In this analysis, using bar charts and scatter plots, the cleaned data acquired by pre-processing is visualized. It illustrates the actions of data attributes. It makes it easy to grasp the attribute's complicated relationship by graphical representation.

As mentioned above, this visualization plays a crucial role in data exploration. The various parameters of the dataset plotted dependent on the age of patients as seen in Fig.3 below. Using this Pairplot method from the Seaborn library of Python, it can be inferred the cardiac disorder, aka. Cardiovascular disorders rely primarily on the patient's age.



Exploratory Data Analysis based on Age

TABLE I. TEST PROPORTION TO DETERMINE THE BEST RATION FOR HIGHEST MODEL ACCURACY (ROUNDED TO 4 DIGITS)

Algorithms	Testing/Training Accuracy Ratios								
	0.1	***0.2***	***0.3***	***0.4***	***0.5***	***0.6***	***0.7***	***0.8***	***0.9***
KNN	0.7419	0.6885	0.6263	0.6393	0.6118	0.5989	0.6244	0.646	0.6043
Logistic Regression	0.8609	**0.8788**	0.8351	0.836	0.8289	0.8169	0.8169	0.7942	0.7912
Random forest	0.7419	0.8032	0.8681	0.8196	0.7828	0.7802	0.7652	0.7942	0.7032

Training & Testing:

In machine learning, algorithms rely on data for both input and output. The input data, known as the training dataset, is used to train the model and verify its functionality. The heart disease dataset is divided into training and testing databases. Algorithms are applied to the training dataset to generate models, which are then evaluated using the testing dataset. Random signals and inputs are utilized during testing to ensure the model operates correctly.

Algorithms play a crucial role in constructing machine learning and data science models. These algorithms can include supervised, semi-supervised, unsupervised, and reinforcement learning algorithms. In this research paper, three supervised algorithms are utilized: Random Forest, Logistic Regression, and K-Nearest Neighbor.

Random Forest:

Random Forest is a tree-based classifier technique in machine learning. It creates multiple decision trees based on different attributes, and the algorithm's performance is the mean of the predicted results of these trees. Random Forest employs bootstrap aggregating (bagging) for tree learning.

Logistic Regression:

Logistic Regression predicts the likelihood of a binary outcome by modeling the relationship between input variables and the probability of the outcome. It is used to calculate the probability of success and involves estimating regression coefficients using Maximum Likelihood Ratio (MLR).

K-Nearest Neighbor (KNN):

K-Nearest Neighbor extracts data points from the dataset and estimates the closest output. It provides high predictive precision and is suitable for pattern recognition tasks. KNN utilizes Euclidean distance to measure similarity between data points.

TABLE II. COMPARATIVE ANALYSIS WITH PREVIOUS STUDIES

Paper	Data set	Year	Technique	Accuracy	Error rate
An Analysis of Heart Disease Prediction using Different Data Mining Techniques [12]	UCI	2012	*Naïve Bayes*	52.33%	47.67%
			Decision Tree	52%	48%
			KNN	45.67%	54.33%
Heart Disease Prediction using Machine Learning Algorithms	UCI	-	*Logistic Regression*	87.88%	12.12%
			KNN	74.19%	25.81%
			Random Forest	86.81%	13.19%
Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers [8]	UCI	2011	*WAC*	57.75%	42.25%
			CBA	58.28%	21.72%
			CMAR	53.64%	46.36%
			CPAR	52.32%	47.68%
A Heart Disease Prediction Model using Decision Tree [1]	UCI	2013	*J48 Unpruned tree*	72.8%	27.2%
			J48 Pruned tree	73.79%	26.1%
			J48 Reduced Error Pruning	75.73%	24.27%
Improving Heart Disease Prediction Using Feature Selection Approaches [3]	UCI	2019	*Logistic Regression*	82.56%	17.44%
			Random Forest	84.17%	15.83%
			Naïve Bayes	84.24%	15.76%
			LR- SVM	84.85%	15.15%
			Decision Tree	82.22%	17.78%

Final Product Prototype Process:

1. Data Preprocessing: The heart disease dataset is cleaned and normalized to ensure consistency and accuracy.
2. Model Training: Random Forest, Logistic Regression, and K-Nearest Neighbor algorithms are applied to the training dataset to generate predictive models.
3. Model Evaluation: The trained models are evaluated using the testing dataset to assess their performance and predictive accuracy.
4. Comparison of Algorithms: The performance of Random Forest, Logistic Regression, and K-Nearest Neighbor algorithms is compared to determine the most effective approach for heart disease prediction.
5. Integration: The selected algorithm is integrated into the final product prototype, providing a user-friendly interface for inputting patient data and obtaining real-time predictions.

RESULTS:

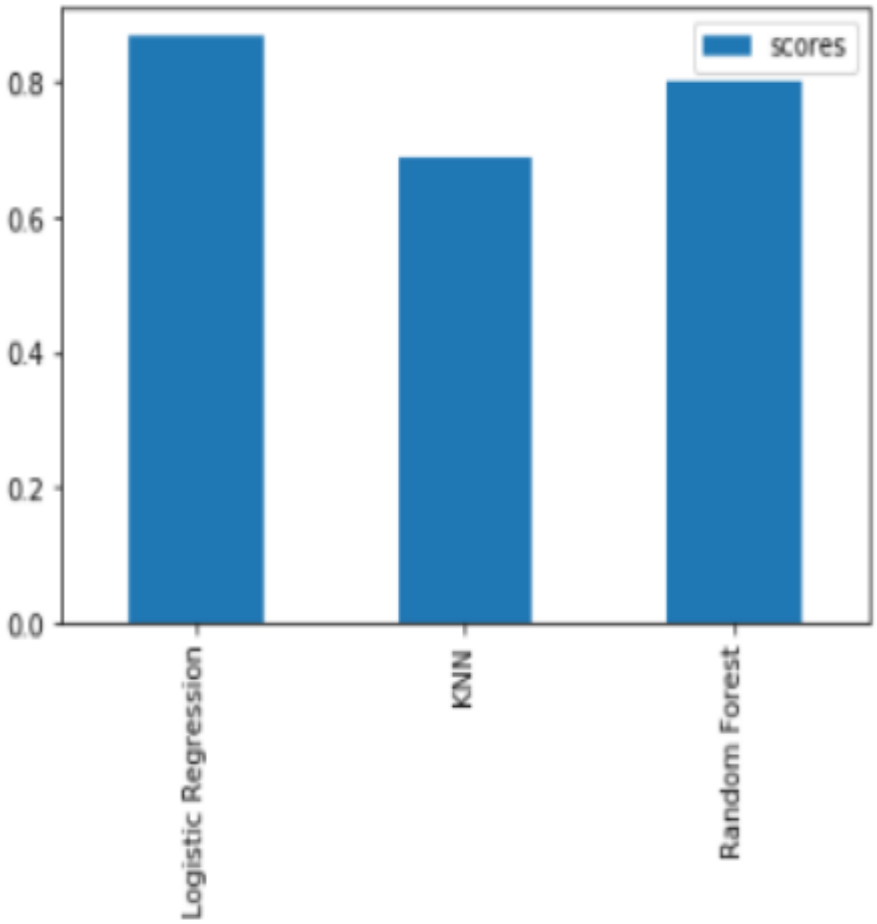
Datasets are applied to Logistic Regression, KNN, and Random Forest. [12] Concerning the algorithm and the proportion of test data, the results differ. When the proportion of test data is 0.2 percent, the maximum accuracy achieved is 87 percent by logistic regression. It is also shown in Table II above.

As mentioned below in Table II, different models constitute different results, accuracy levels, and error rates. Throughout time, the UCI Heart Disease dataset has been used due to its easy availability and ease of usage, which can be can sees since many approaches are being used. Table III gives a thorough comparative analysis of the models used in this paper with the previous research models.

CONCLUSION:

In this paper, a comprehensive machine learning-based model is proposed to aid medical practitioners in diagnosing heart diseases at an early stage, enabling patients to take precautionary measures within a rectification window. Utilizing three separate classifiers in the model, it is evident that the proportion of test and training data significantly impacts the classification model's performance.

The experimental findings, as depicted in Figure, visually compare the accuracies of different algorithms. It is observed that Logistic Regression performs best with a test data size of 0.2. However, Random Forest also yields promising results at a test data size of 0.3. Therefore, it is recommended to utilize logistic regression and random forest classification algorithms for achieving optimal heart attack predictions.



This research contributes to refining previous works in this domain, which may be considered outdated. Previous research lacked precision, with accuracies ranging from 72% to 84%. The comparative analysis provided in Table III demonstrates that the methodology employed in this paper outperforms earlier approaches on the same dataset.

In conclusion, the methodology presented in this research paper offers a more effective approach to heart disease prediction compared to previous studies. By leveraging machine learning techniques and optimizing various parameters, higher accuracy rates are achieved. This advancement has the potential to significantly improve the early detection and management of heart diseases, ultimately leading to better patient outcomes and healthcare outcomes.

REFERENCES:

1. Pandey, A.K. et al. (2013). Heart Disease Prediction Model using Decision Tree. IOSR Journal of Computer Engineering, 83-86.

2. Mythili, T. et al. (April 2013). Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications in Technology, 11 - 15.

3. Bashir, S. et al. (2019). Improving Heart Disease Prediction Using Feature Selection Approach. 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST), pp. 619 - 623.

4. Mohan, S. et al. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access.

5. Gavhane, A. et al. (2018). Prediction of heart disease using machine learning. 2nd International Conference on Electronics, Communication, and Aerospace Technology, pp. 1275 - 1278. IEEE.

6. Soni, J. et al. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications, 43 - 48.

7. Taneja, A. et al. (2013). Heart Disease Prediction System Using Data Mining Techniques. Oriental Journey of Computer Science & Technology, pp. 457 - 466.

8. Ishtake, S.H. et al. (2013). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International J. of Healthcare & Biomedical Research, 94-101.

9. Ziasabounchi, N. et al. (2014). ANFIS Based Classification Model for Heart Disease Prediction. International Journal of Engineering & Computer Science, pp. 7 - 12.

10. Masethe, H.D. et al. (2014). Prediction of Heart Disease using Classification Algorithms. Proceedings of the World Congress on Engineering and Computer Science.