UtilitasMathematica
ISSN 0315-3681 Volume 121, 2024

CARDIOALERT: PREDICTING HEART DISEASE WITH DATA-DRIVEN INSIGHTS USING MACHINE LEARNING

[1]JAGADEESWARI PILLA,[2]DR K.SELVANI DEEPTHI KAVILA
[1]M.tech Student, Department of Computer Science And Engineering(AI & ML, DATA SCIENCE), **Anil Neerukonda Institute Of Technology And Sciences(A),**Visakhapatnam, Andhra Pradesh,531162.
[2]Professor,Department of Computer Science And Engineering(AI & ML, DATA SCIENCE), **Anil Neerukonda Institute Of Technology And Sciences(A)**,Visakhapatnam, Andhra Pradesh,531162, selvanideepthi14@gmail.com,

Abstract:
In healthcare industry generates vast datasets containing valuable but often concealed information crucial for informed decision-making. This comprehensive study aims to leverage advanced techniques like data mining and some preprocessing methods to develop a robust HDPS which is Heart Disease Prediction. Utilizing sophisticated classifications algorithms such as KNN, Logistic regressor, Extreme Gradient Boost, SVM and Random Forest the HDPS predicts heart disease risk levels with high accuracy. Integrating 15 key medical parameters which includes age, gender and blood pressure etc.. the system generates prediction about the heart disease, facilitating the extraction of significant knowledge and identification of intricate relationships among medical factors. A pivotal component is the use of a neural network with back propagation method, which enhances the system's learning capabilities for refined predictions. Additionally, fine-tuning these machine learning models optimizes their performance, ensuring even greater predictive accuracy. Results demonstrate the system's success in accurately predicting heart disease risk levels. By employing diverse advanced algorithms, fine-tuning techniques, and powerful data mining methods, the study establishes a reliable HDPS, empowering healthcare professionals to understand the complex interplay of medical factors and fostering informed, proactive healthcare decision-making.
Keywords: HDPS Prediction, Data Mining, KNN, Regressor method, Extreme Gradient Boost, Random Forest, SVM, Multilayer Perceptron, Neural Network.

1 INTRODUCTION:

The healthcare industry continually generates extensive datasets containing a wealth of information Leveraging data-rich information is essential for accurate clinical decision-making and delivering high-quality patient care. However, the sheer volume and complexity of these datasets often obscure valuable insights that could enhance medical diagnoses and treatments. In recent years, the advent of advanced data mining techniques has provided powerful tools to unearth hidden patterns and relationships within these large datasets, offering significant potential for improving healthcare outcomes.

A key area where data mining can greatly influence healthcare is in predicting and managing heart disease, a leading cause of death globally.Accurate early prediction of heart disease risk can dramatically of heart disease risk can dramatically enhance patient outcomes by facilitating

timely interventions and personalized treatment strategies.Developing a prediction of heart disease where we are utilizing advanced ML is therefore critically important.
This study aims to create an HDPS utilizing a suite of advanced algorithms, including KNN, Logistic Regression, Extreme Gradient Boost, Random Forest, and SVM. These algorithms were selected for their robustness and ability to effectively manage complex, high-dimensional data. The system incorporates 15 essential medical parameters all of which are known to impact heart disease risk. A key component of this research is the implementation of abackpropagation method. This neural network architecture improves the system's capacity to learn from the data, adapting and refining its predictions to improve accuracy. By harnessing these advanced techniques, the HDPS aims not only to predict the probability of heart disease but also to provide healthcare professionals with deep insights into the underlying medical factors and their interactions.
The findings from this study demonstrate the efficacy of the HDPS in accurately getting an risk, underscoring the value of integrating diverse data mining methodologies in healthcare. This research contributes to the broader field of predictive analytics, offering a framework that can be adapted and expanded for various medical applications. Ultimately, the HDPS empowers healthcare providers with actionable insights, facilitating informed, proactive decision-making that can lead to get an better accurate.

2 LITERATURE REVIEW

Recent advancements in data mining and machine learning have significantly impacted the healthcare sector, particularly in cardiology, aiding in the early detection and prevention of heart disease, a leading cause of mortality globally [1,2,3,4,5].
Narain et al. (2016) [6] introduced a novel machine-learning-based cardiovascular disease (CVD) prediction system to enhance the precision of the Framingham risk score (FRS). Their system, employing a quantum neural network, achieved a remarkable 98.57% accuracy in forecasting CVD risk, outperforming existing methods.

Shah et al. (2020) [7] developed a model for predicting cardiovascular disease using machine learning techniques on the Cleveland heart disease dataset. Their study highlighted the K-nearest neighbor (KKN) model with 90.8% accuracy as a promising approach for disease prediction.
Drod et al. (2022) [8] aimed to identify significant risk variables for CVD in patients with metabolic-associated fatty liver disease (MAFLD) using machine learning techniques. Their model, incorporating multiple logistic regression, achieved 85.11% accuracy in identifying high-risk patients.
Alotalibi (2019) [9] investigated the utility of machine learning techniques in predicting heart failure disease. Their study on the Cleveland Clinic Foundation dataset revealed the decision tree algorithm with 93.19% accuracy as a potential tool for predicting heart failure.
Hasan and Bao (2020) [10] compared feature selection methods for predicting cardiovascular illness. Their study concluded that the XGBoost classifier with the wrapper technique provided the most accurate prediction results, achieving 73.74% accuracy.
However, prior research suffered from limited datasets, increasing the risk of overfitting. In contrast, our study utilized a larger cardiovascular disease dataset of 70,000 patients with 11 features, reducing the likelihood of overfitting and enhancing the generalizability of the models.

3 PROBLEM STATEMENT

The overall objective of my work is to get the good accurate value presence of heart disease using a minimal number variables. By focusing on key attributes, we aim to achieve efficient and precise predictions. While many additional input attributes could be considered, our goal is to enhance prediction speed and efficiency with fewer attributes. Currently, decisions are frequently guided by doctor's intuition and experience instead of utilizing the valuable, data-rich information embedded in datasets . This approach can lead to increase the accuracy

4 METHODOLOGY

The proposed work revolves around the development of a robust HDPS employing advanced techniqueslike data preprocessing and data mining a meticulous fine-tuning process. Key to this endeavor is the strategic selection of algorithms, including K-NN, Logistic Regression, Extreme Gradient Boost, Random Forest, and SVM, known for their efficacy in predictive modeling. Utilizing a publicly available heart disease dataset provides a standardized foundation for evaluation, ensuring reproducibility and facilitating comparisons.

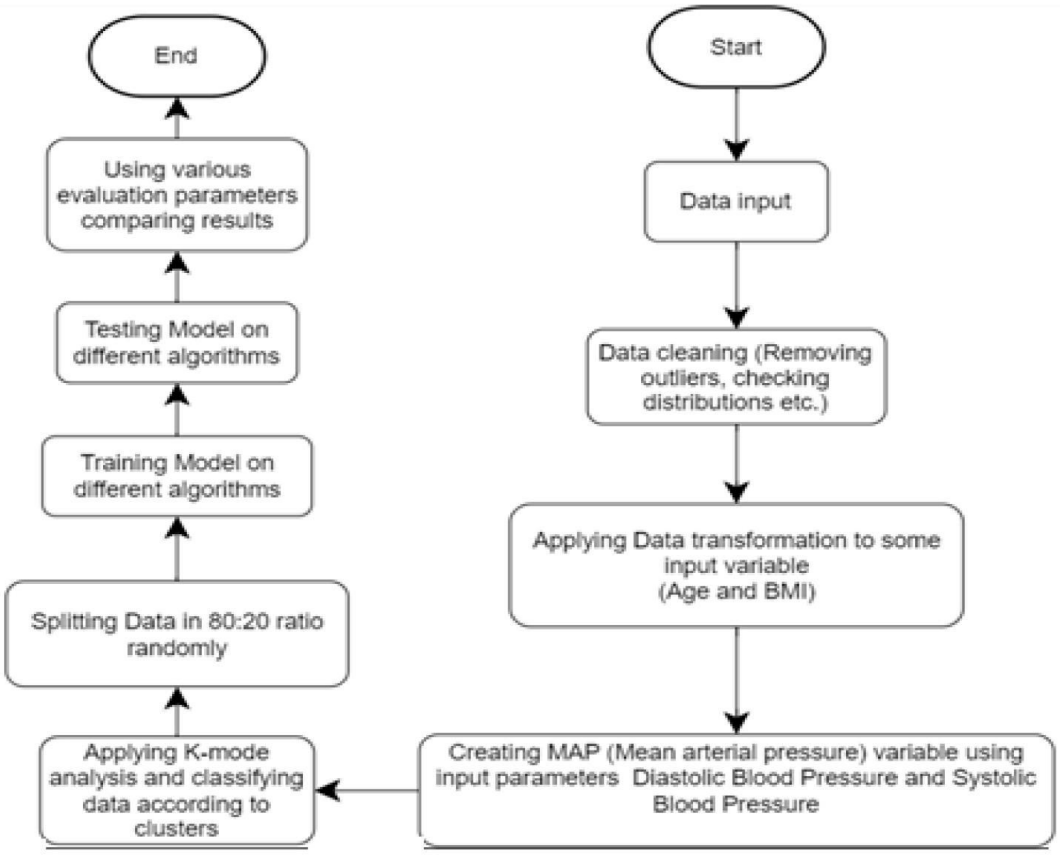

Figure 1: Proposed system

The innovation lies in the systematic fine-tuning of these algorithms to adapt to the dataset's intricacies, progressively refining basic models to capture the patterns about the disease. Central to the study is the integration of a multilayer perceptron neural network with back propagation method , which enhances the system's learning capabilities. Evaluation metrics with True positive rate false positive rate and true negative rate and false negative rate predictive

performance. Beyond accuracy, the research aims to contribute to the reproducibility and standardization of the field, providing insights into the relationships between medical parameters and heart disease. This holistic approach positions the proposed work at the forefront of data-driven healthcare research, with implications extending to the broader discourse on transparent, comparable, and interpretable predictive modeling in cardiovascular health. The proposed system of work shown in Figure 1.

4.1 Dataset Discerption:

The given data is sourced from Kaggle and pertains to heart-related health information. Comprising a total of 303 entries, the dataset is indexed from 0 to 302, with each entry corresponding to a unique record. The dataset encompasses 14 columns, each capturing distinct attributes and test results associated with patients, providing a comprehensive insight into cardiovascular health. The description of data set shown in Table 1.

Table 1: Dataset Description:

Variable	Description
age	Age of the patient in years
sex	Gender of the patient (0 = male, 1 = female)
cp	Chest pain type: 0: Typical angina 1: Atypical angina 2: Non-anginal pain 3: Asymptomatic
trestbps	Resting blood pressure in mm Hg
chol	Serum cholesterol in mg/dl
fbs	Fasting blood sugar level, categorized as above 120 mg/dl (1 = true, 0 = false)
restecg	Resting electrocardiographic results: 0: Normal 1: Having ST-T wave abnormality 2: Showing probable or definite left ventricular hypertrophy
thalach	Maximum heart rate achieved during a stress test
exang	Exercise-induced angina (1 = yes, 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment: 0: Upsloping 1: Flat 2: Downsloping
ca	Number of major vessels (0-4) colored by fluoroscopy

thal	Thalium stress test result: 0: Normal 1: Fixed defect 2: Reversible defect 3: Not described
target	Heart disease status (0 = no disease, 1 = presence of disease)

4.2 Preprocessing of the Dataset:

The preprocessing steps you have outlined are crucial for preparing a dataset for analysis and modeling. Here is an expanded explanation of each step:

Remove irrelevant features:

Identify and remove features that do not contribute meaningful information to the analysis or are redundant. This helps reduce dimensionality and can improve the efficiency of the modeling process.

Address missing values:

Missing values with statistical measures like the mean , median, or mode, removing rows or columns containing missing data, or employing advanced imputation methods

Treat outliers:

It is the process which can have a substantial effect on the performance of some models. Detection methods, such as the Interquartile Range (IQR) or Z-score, can be employed to identify outliers. Baesd on the given data, outliers can be removed or transformed to mitigate their impact.

Encode categorical variables:

Since many ML algorithms need numerical input, categorical variables must be converted into numerical formats. This conversion can be performed using techniques like one-hot encoding, or label encoding, which assign a unique numerical label to each category.In the process of encoding categorical features for the heart disease dataset, a thoughtful decision-making approach was applied based on varible nature. Nominal variables, such as chest pain type (cp), resting electrocardiographic results (restecg), and thalium stress test result (thal), were deemed suitable for one-hot encoding, given their non-ordinal characteristics. This approach prevents unintentional ordinal relationships and effectively represents the diverse categories within these features. On the other hand, binary variables were recognized as not requiring one-hot encoding, as their inherent binary nature naturally captures the two distinct categories without introducing unnecessary complexity. Additionally, ordinal variables, and the slope of the peak exercise ST segmentand , were acknowledged as not needing one-hot encoding, as their ordered nature can be adequately preserved. This comprehensive decision-making process ensures that categorical features are appropriately prepared for subsequent machine learning analyses, contributing to the robustness of the modeling pipeline.

Feature Scaling:

IT is an essential preprocessing method and step for algorithms that are sensitive to the range or scale of features. Models like SVM, KNN, and many linear models rely on distances or gradients, making them susceptible to variations in feature scales. Feature Scalling will ensure

that all features contribute equally to model's decisions, preventing those with larger magnitudes from disproportionately influencing the results.

Transform skewed features to achieve normal-like distributions:

Some machine learning models assume that the features have a normal distribution. If a feature is skewed, it might be beneficial to apply transformations, such as log transformations or Box-Cox transformations, to make the distribution more normal. Where we will improve the performance of assume normality.

These preprocessing steps important to make data normalized and the normalized data is used for the analysis and modeling. The specific techniques and methods employed in each step may vary based on the characteristics of the dataset and the requirements of the modeling algorithm being used. Preprocessing is a critical phase in the data science pipeline, and its effectiveness can impact the performance.

Box-Cox transformation is a powerful method to stabilize variance and make the data more normal-distribution-like. It's particularly useful when you're unsure about the exact nature of the distribution you're dealing with, as it can adapt itself to the best power transformation. However, the Box-Cox transformation only works for positive data, so one must be cautious when applying it to features that contain zeros or negative values.

Transforming Skewed Features & Data Leakage Concerns:

When preprocessing data, especially applying transformations like the Box-Cox, it's essential to be wary of data leakage. Data leakage refers to a mistake in the preprocessing of data in which information from outside the training dataset is used to transform or train the model. This can lead to overly optimistic performance metrics.

To avoid data leakage and ensure our model generalizes well to unseen data:

1- Data Splitting: In this processing will split the data into train and test with the ratio of 70%, 30% or 80% , 20%. and for this processes will use the sklearn library. It will separate the data and do the model performance.

2- Box-Cox Transformation: We'll examine the distribution of the continuous features in the training set. If they appear skewed, we'll apply the Box-Cox transformation to stabilize variance and make the data more normal-distribution-like. Importantly, we'll determine the Box-Cox transformation parameters solely based on the training data.

3- Applying Transformations to Test Data: Once our transformation parameters are determined from the training set, we'll use these exact parameters to transform our validation/test set. This approach ensures that no information from the validation/test set leaks into our training process.

4. Hyperparameter Tuning & Cross-Validation: Given our dataset's size, to make the most of the available data during the model training phase, we'll employ cross-validation on the training set for hyperparameter tuning. This allows us to get a better sense of how our model might perform on unseen data, without actually using the test set. The test set remains untouched during this phase and is only used to evaluate the final model's performance.

By following this structured approach, we ensure a rigorous training process, minimize the risk of data leakage, and set ourselves up to get a realistic measure of our model's performance on unseen data.

The in-depth analysis of the variable transformations provides valuable insights into the impact of these transformations on the distributions of key features.

Age: The transformation applied to the age variable has notably improved the symmetry of its distribution, aligning it more closely with a normal distribution. This transformation contributes

to meeting the assumptions of normality in statistical analyses, enhancing the reliability of results.

Trestbps: The distribution of trestbps, representing resting blood pressure, displays a more normal-like shape post-transformation. The reduced skewness suggests that the Box-Cox transformation has effectively mitigated the non-normality in the data, contributing to the validity of subsequent analyses involving this variable.

Chol: The Box-Cox transformation has successfully rendered the distribution of cholesterol levels (chol) more akin to a normal distribution. This adjustment is crucial for statistical analyses that assume normality, ensuring the robustness of inferences drawn from the data.

Thalach: Thalach, representing the maximum heart rate achieved during stress tests, initially exhibited a fairly symmetric distribution. Post-transformation, its shape remains similar, indicating that the original distribution was already close to normal. This observation suggests that the variable may not have required a significant transformation to meet normality assumptions.

Oldpeak: Although the transformation has improved the distribution of oldpeak, it is noted that the variable still does not perfectly resemble a normal distribution. This discrepancy could be attributed to the inherent nature of the data or the presence of outliers. To further enhance normality, consideration of advanced transformations like the Yeo-Johnson transformation, capable of handling zero and negative values directly, may be explored for a more comprehensive normalization of the variable.

In summary, the transformations have generally succeeded in improving the normality of the feature distributions, aligning them more closely with the assumptions necessary for robust statistical analyses. The nuanced assessment of each variable's transformation provides valuable guidance for refining the preprocessing steps and ensuring the appropriateness of the applied transformations in the context of the dataset.

4.3 MachineLearningClassifiersProposed

In our proposed approach, we applied various machine learning classifiers to our dataset, beginning with a comprehensive analysis. Firstly, we employed linear model selection using Logistic Regression to establish relationships within the data. Next, for neighbor selection techniques, we utilized the KNeighbors Classifier. Additionally, we implemented the RandomForest Classifier, a popular ensemble method, to improve model performance. To handle high dimensionality effectively, we incorporated Support Vector Machine (SVM). Furthermore, we employed the XGBoost classifier, which combines decision tree methods within an ensemble framework.

4.3.1 Random Forest Algorithm:

Random Forest, a popular supervised learning technique, operates based on ensemble learning principles. It creates a random forest by combining multiple decision trees to improve model performance. The process involves selecting random data points, building decision trees for each subset, and making predictions based on majority votes from all trees.

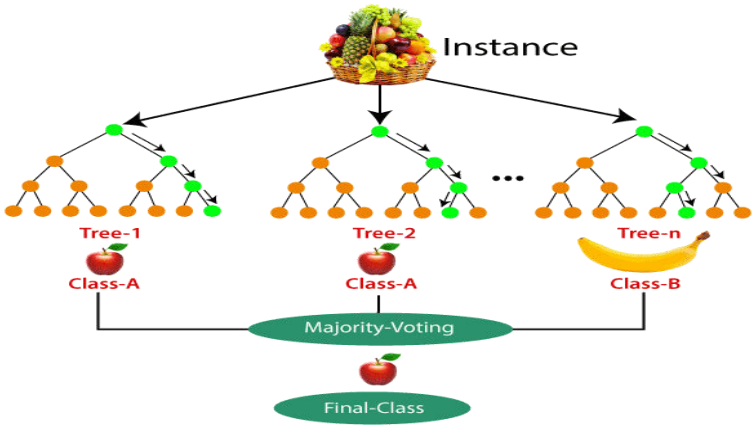

Figure 2: Random Forest Example

4.3.2 Linear Regression:

SVM is the one of the classification model which defines the relationship between two variables, typically represented as a straight line. It aims to model this linear relationship and forecast new observations. The regression equation consists of the intercept, slope, and an error term.

4.3.3 XGBoost:

XGBoost, an extreme gradient boosting algorithm, is renowned for its speed and accuracy in tree-based models. It utilizes ensemble learning and gradient descent optimization to deliver optimal results by iteratively correcting errors from previous models.

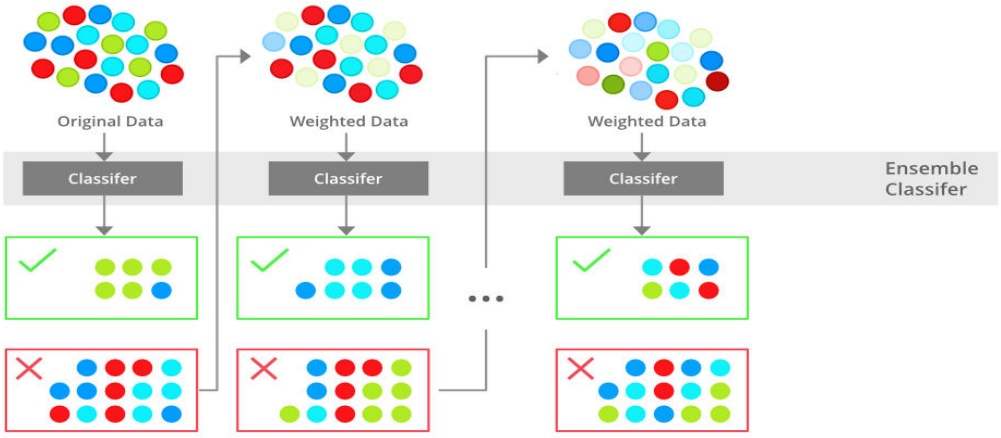

Figure 3: Boosting algorithms – working principle

4.3.4 Support Vector Machine Algorithm:

SVM is a popular choice for classifier tasks in supervised learning. It aims to create the best decision boundary, known as a hyperplane, to

segregate data points into different classes. SVM identifies support vectors, extreme points crucial for defining the hyperplane.

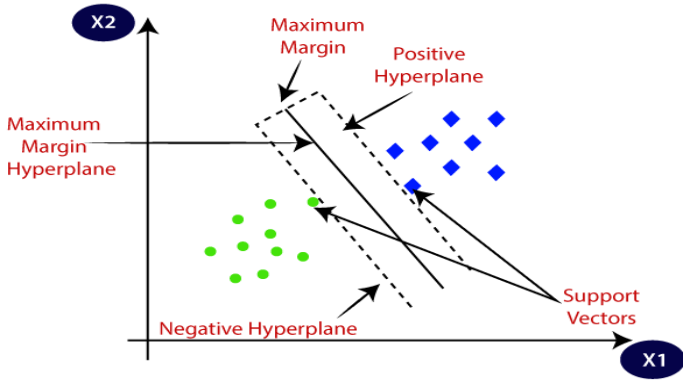

Figure 4: SVM

4.3.5 Naïve Bayes Algorithm:

Naive Bayes is a supervised learning algorithm used to mae predictions by calculating the probability of an outcome based on given data. Named after Bayes' theorem, it operates under the naive assumption that all features or variables are independent of one another. This assumption simplifies the calculation of conditional probabilities, where the probability of an event A is determined given that event B has already occurred.

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$$

Where:

1. P(A/B) is the probability of event A given that event B has occurred.
2. P(B/A) is the probability of event B given that event A has occurred.
3. P(A) and P(B) are the probabilities of events A and B independently.

Naive Bayes is a powerful classifier, known for its simplicity and effectiveness, particularly in next classification tasks. It is easy to implement.

4.4 Training process:

The heart disease prediction task involves a structured training process with two main phases.Initially, the regression model, SVM, XGBoost, RFC, and Naive Bayes are trained on the dataset without any fine-tuning.This phase serves to establish baseline performance metrics such as getting an True Positive rate with good accurate value .During this phase, the dataset is split into training and testing sets to asses the predictive capabilities of each model.

In such a way the second phase involves fine-tuning each model to better align with the specific characteristics of the dataset.This step includes adjusting hyperparameters such as learning rates,batch sizes, and the structure of neural networks(e.g., the number of hidden layers and neurons). Techniques like grid search or random search are employed to efficiently explore the hyperparameter space. After fine-tuning, the models are retrained on the dataset to capture the nuances of heart disease patterns more accurately, leading to improved prediction performance.

Best practices are followed throughout both phases, including handling class imbalances, scaling features appropriately, and using cross-validation strategies to ensure model robustness. The final evaluation compares the fine-tuned models against their initial baselines, highlighting the effectiveness of the fine-tuning process in enhancing predictive accuracy for heart disease risk.

5 EXPERIMENTAL ANALYSIS AND RESULTS

5.1 Implementation:
The implementation was carried out using google cloab on a machine equipped with an intel core i5 processor and 16 GB of RAM. The original dataset, consisting of 70,000 rows and 12 attributes, was cleaned and preprocessed, resulting in approximately 59,000 rows and 11 attributes.Outliers were removed to improve model efficiency since all attributes were categorical.The study employed several algorithms,including Random Forest, SVM and XGBoost,with performance metrices such as getting an accurate value.

5.2 **Parameters for Evaluation**
The following metrics are used for the evaluation.

- Confusion Matrix
- Accuracy

Confusion Matrix: A confusion matrix is a table used to evaluate the performance of a classification system, summarizing the output of a classification algorithm through four main properties. These properties include True Positive (TP), where the data is predicted to be positive and is indeed positive; False Positive (FP), where the data is predicted to be positive but is actually negative; True Negative (TN), where the data is predicted to be negative and is indeed negative; and False Negative (FN), where the data is predicted to be negative but is actually positive. These metrics provide a comprehensive assessment of the classifier's accuracy.
To simplify representation, the confusion matrix is depicted using shades of color within each class: lighter shades indicate larger numbers, while darker shades indicate smaller numbers. In this matrix, the false positive (4) and false negative (6) sectors are darker, signifying that the model makes very few errors and thus has more accurate predictions.

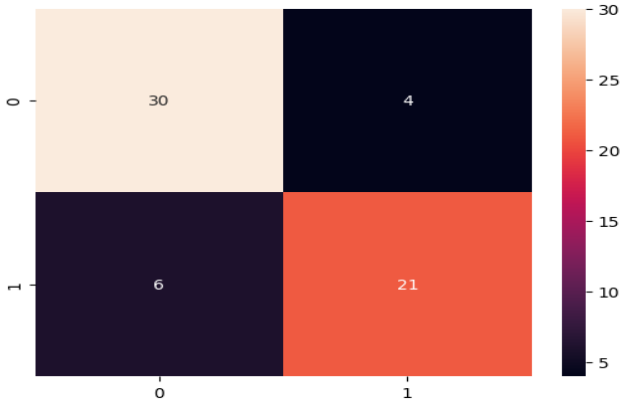

Figure 5: Confusion matrix for XGBoost_Tuned model

Precision: It is an essential metric in classification, especially when handling imbalanced datasets.It measures the accuracy of a model's which is positive predictions by calculating the proportion of true positive instances among all instances predicted as positive. The formula is:TP/TP+FP
Accuracy:Accuracy measures the overall effectiveness of a model by considering both positive and negative predictions. It is calculated using the formula:
Accuracy= TP+TN/ TP+TN+ FP+FN
Recall (Sensitivity or True Positive Rate): Recall, also known as sensitivity or the true positive rate, measures the proportion of actual positive instances that are correctly predicted. The formula for recall is: Recall =TP/TP+FN

Remember that precision and recall are often used together, especially in scenarios where class imbalance exists.

F-Measure: It is calculated using true positives(TP), false positives(FP), true negatives(TN), nad false negatives(FN)rates, proving a balanced assessment of model performance.The F1-score is derived from the formula.

F-Measure=2· Precision+Recall/Precision· Recall

Where Table 2 provides a detaild overview of the performance metrics for the ml algorithms to predict the heart disease task. Each row corresponds to a distinct model, showcasing its performance across key metrics such as True positive rate , False Positive rate and True negative rate and False Negative rate.

Regressor model and Random Forest both achieved the highest accuracy score of 89.01%, but the Random Forest model exhibited superior precision (92.11%) compared to Logistic Regression (88.10%). With Tuned Random Forest algoritham we have got the 90.11% accurate value and demonstrated high recall (88.10%) and F1 score (89.16%). The Extreme Gradient Boost model lagged behind, showing the lowest accuracy (75.82%) and R2 score (2.72%), indicating its relatively poor performance. Tuned K-Nearest Neighbour models, both manual and GridSearchCV, exhibited improvements over the basic K-Nearest Neighbour model, achieving accuracy scores of 86.81%. The models' AUC scores indicate their ability to discriminate between classes, with the Tuned Random Forest achieving the highest AUC score of 89.97%. Overall, the Tuned Random Forest model appears to be the most effective, balancing accuracy, precision, and recall.

Table 2: Performance of machine learning models

Model	Accuracy Score	Recall Score	F1 Score	AUC Score	Precision Score	R2 Score
K-Nearest Neighbour	83.52%	80.95%	81.93%	83.33%	82.93%	33.67
Logistic Regression	89.01%	88.10%	88.10%	88.95%	88.10%	55.78
Random Forest	89.01%	83.33%	87.50%	88.61%	92.11%	55.78
Support Vector Machine	84.62%	83.33%	83.33%	84.52%	83.33%	38.10
Extreme Gradient Boost	75.82%	78.57%	75.00%	76.02%	71.74%	2.72
Tuned K-Nearest Neighbour (Manual)	86.81%	85.71%	85.71%	86.73%	85.71%	46.94
Tuned K-Nearest Neighbour (GridSearchCV)	86.81%	88.10%	86.05%	86.90%	84.09%	46.94
Tuned Random Forest	90.11%	88.10%	89.16%	89.97%	90.24%	60.20

5.3 Models Comparison:

The Table 3 is focused on a simplified model comparison, specifically highlighting the algorithms which we have used to get the best accuracy value. Each row represents a model, providing a quick glance at its accuracy performance. The tuned Extreme Gradient Boost model achieved the highest accuracy at 90.11%,with the neural network closely following at 89.01%. Both Logistic Regression and Random Forest models also showed strong performance with

89.01% accuracy. The K-Nearest Neighbour (Tuned) model achieved 86.81%, reflecting significant improvements through tuning. Models like

Naive Bayes, tuned SVM scored with 84.62% accuracy. The untuned KNN model had an accuracy of 83.52%, while the Extreme Gradient Boost without tuning scored 75.82%. Interestingly, the Random Forest (Tuned) model performed the lowest with an accuracy of 60.20%, suggesting that the tuning process may have been suboptimal for this model. This table underscores the critical role of model selection and hyperparameter tuning in achieving high accuracy in predictive tasks. These accuracy-centric insights from the model comparison table are particularly useful when seeking a quick overview of each model's performance, aiding in the initial stages of model selection for a heart disease prediction system. Both tables collectively contribute to a comprehensive evaluation strategy, considering a range of metrics for a holistic understanding of model performance.

Table 3: Models comparison

Model_Name	Accuracy_Score
Extreme Gradient Boost_Tuned	90.110000
Neural Network	89.010989
Logistic Regression	89.010000
Random Forest	89.010000
K-Nearest Neighbour_Tuned	86.810000
Naive Bayes	84.620000
Support Vector Machine_Tuned	84.620000
SVM	84.620000
K-NN	83.520000
Extreme Gradient Boost	75.820000
Random Forest_Tuned	60.200000

6 CONCLUSION:

In conclusion, the development and evaluation of the disease predictionusing ml models have highlighted significant findings the risk of heart disease. Leveraging a dataset which includes the parameters like age, gender, what is the blood pressure study assessed model performance. Among the evaluated models—Extreme Gradient Boost_Tuned, Neural Network, Logistic Regression, Random Forest, K-Nearest Neighbour_Tuned, Naive Bayes, Support Vector Machine_Tuned, Support Vector Machine, KNN, and Extreme Gradient

Boost—the Extreme Gradient Boost_Tuned model achieved the highest accuracy score of 90.11%, demonstrating its superior predictive capability. The Neural Network, Logistic Regression, and Random Forest models also performed well, each achieving an accuracy score of 89.01%. In contrast, the K-Nearest Neighbour_Tuned model, while showing improvement over its untuned counterpart, achieved a lower accuracy of 86.81%. These results underscore the effectiveness of fine-tuning and optimization, as evidenced by the Extreme Gradient Boost_Tuned model's leading performance. The comprehensive evaluation, including metrics such as precision, recall, and AUC Score, provides a well-rounded understanding the conditions.

This study advances the field of health informatics by illustrating the efficacy of machine learning models in predicting heart disease risk. The findings provide a strong basic advancing research and developing an effective HDPS.Future research endeavors aim to refine and enhance prediction models, contributing to more accurate and proactive disease management in real-world healthcare settings.By addressing these areas, the goal is to improve the overall effectiveness and applicability of heart disease prediction systems.

7.REFERENCES:

[1] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. *Eur. J. Mol. Clin. Med.* 2020, *7*, 1638–1645
[2] Breiman, L. Random forests. *Mach. Learn.* 2001, *45*, 5–32.
[3] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
[4] Gietzelt, M.; Wolf, K.-H.; Marschollek, M.; Haux, R. Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. *Comput. Methods Programs Biomed.* 2013, *111*, 62–71.
[5] K, V.; Singaraju, J. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. *Int. J. Comput. Appl.* 2011, *19*, 6–12.
[6] Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.
[7] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* 2020, *1*, 345
[8] Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovasc. Diabetol.* 2022, *21*, 240.
[9] Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. *Int. J. Adv. Comput. Sci. Appl.* 2019, *10*, 261–268
[10] Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol.* 2020, *11*, 49–62
[11] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked* 2021, *26*, 100655.
[12] Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.
[13] Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. *J. Southwest Jiaotong Univ.* 2021, *56*, 220–240
[14] Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. *Int. J. Comput. Appl.* 2020, *176*, 46–54.

[15] Selvani Deepthi Kavila, Rajesh Bandaru, Tanishk Venkat Mahesh Babu Gali,2022,27-54,IGI Global,Analysis of cardiovascular disease prediction using model-agnostic explainable artificial intelligence techniques.
[16] Selvani Deepthi Kavila, Rajesh Bandaru, Tanishk Venkat Mahesh Babu Gali, Jana Shafi,Analysis of Cardiovascular Disease Prediction Using Model-Agnostic Explainable Artificial Intelligence Techniques,2022,28,10.4018/978-1-6684-3791-9.ch002,
[17] Selvani Deepthi Kavila, SVSSS Lakshmi, Rajesh Bandaru, Shaik Riyaz,2023/6/7,501-514,Lung Cancer Disease Prediction Using Machine Learning Techniques.
[18] Sivaram Kommineni; Sanvitha Muddana; Rajiv Senapati,23 August 2024,Explainable Artificial Intelligence based ML Models for Heart Disease Prediction
[19] S Saravanan, Kannan Ramkumar, K Narasimhan, V Subramaniyaswamy, Ketan Kotecha, and Ajith Abraham. Explainable artificial intelligence (exai) models for early prediction of parkinson's disease based on spiral and wave drawings. IEEE Access, 2023.
[20] Amann, J., Blasimme, A., Vayena, E., et al.: Explainability for artificial intelligence in healthcare. BMC Med. Inform. Decis. Mak. **20**, 310 (2020)
[21] Chen, Y., Qi, B.: Representation learning in intraoperative vital signs for heart failure risk prediction. BMC Med. Inform. Decis. Mak. **19**, 260 (2019)
[22] Dave, D., Naik, H., Singhal, S., Patel, P.: Explainable AI meets healthcare: a study on heart disease dataset. CoRR abs/2011.03195 (2020)
[23] Koehler, F., Koehler, K., Deckwart, O., et al.: Efficacy of telemedical interventional management in patients with heart failure (tim-hf2): a randomised, controlled, parallel-group, unmasked trial. Lancet (2018)
[24] Moreno-Sanchez, P.A.: Development of an explainable prediction model of heart failure survival by using ensemble trees. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 4902–4910 (2020)
[25] Zhang, D., Wang, W., Li, F.: Association between resting heart rate and coronary artery disease, stroke, sudden death and noncardiovascular diseases: a meta-analysis. CMAJ **188**(15), E384–E392 (2016). https://doi.org/10.1503/cmaj.160050