Comparison of Heart Attack Prediction Analysis Models

Authors: Varma Pericherla, Rujuta Desai, Maniteja Ladi, Swathi Kakani, Hyder Basha Shaik, Shilpa Thangirala.

**Introduction :-**

Heart attacks are a serious health problem that millions of individuals experience worldwide. The condition is brought on by an abrupt stoppage of the heart's blood supply, which results in damage to the heart muscle. If not addressed right away, it can result in severe pain in the chest, shortness of breath, and even death. As a result, preventing and early identification of heart attacks are crucial for lowering the disease's morbidity and fatality rates.

Machine learning algorithms have been utilized to anticipate many ailments, including heart attacks, in the healthcare industry more and more recently. These models can identify people who are at a high risk of having heart attacks by analyzing big databases of clinical and demographic data. Healthcare practitioners can take preventive actions, such as lifestyle changes, drugs, and other interventions, to lower the risk of heart attacks by identifying these people.

In this project, we aim to develop and compare three machine learning models - logistic regression, random forest, and decision tree - to predict the occurrence of heart attacks. Logistic regression is a statistical model that analyzes the relationship between a dependent variable and one or more independent variables. Random forest and decision tree are ensemble models that use multiple decision trees to make predictions. These models can handle complex relationships between variables and are suitable for analyzing large datasets.

To develop our heart attack prediction models, we will use a dataset containing clinical and demographic features of individuals who have suffered from heart attacks. The dataset will include variables such as age, gender, smoking history, blood pressure, cholesterol levels, and other factors that may be associated with heart attacks. We will use these variables as input to the machine learning models to predict the likelihood of an individual having a heart attack.

Our goal is to identify the most accurate and reliable machine learning model that can assist healthcare professionals in making timely and effective interventions for at-risk individuals. By developing a robust heart attack prediction model, we hope to contribute to the improvement of healthcare outcomes and the reduction of heart attack-related morbidity and mortality rates.

**Abstract:**

Given that heart disease is a leading cause of morbidity and mortality worldwide, predicting the risk of heart attacks is a crucial first step in preventing bad effects. Researchers and healthcare practitioners can utilize the information in this dataset to create precise predictive models that can be used to identify people who are at a high risk of suffering a heart attack.

A variety of demographic, behavioral, and medical characteristics that have been demonstrated to affect the risk of heart disease are included in the dataset. Age, gender, smoking habits, blood pressure, cholesterol levels, and a family history of heart disease are some of these variables. By incorporating these data into prediction models, healthcare providers can more precisely identify people who are at increased risk of having heart attacks and develop tailored prevention and treatment regimens.

The creation of precise heart attack risk prediction algorithms has the potential to greatly enhance patient outcomes and lower healthcare expenditures. Early detection of people who are at risk enables prompt intervention and therapy, possibly averting negative consequences like heart attacks or strokes. Predictive modelling can also assist healthcare professionals in setting priorities for patient care and resource allocation, enhancing the effectiveness and efficiency of healthcare delivery.

In general, this dataset serves as a valuable tool for researchers and healthcare professionals working to enhance heart attack prediction and prevention. We can improve patient outcomes and lessen the financial burden of heart disease on society by utilizing the data in this dataset to develop more precise and efficient techniques for diagnosing and treating patients at risk of developing heart disease.

**Project Goal :-**

Enhancing patient outcomes and reducing the effects of heart disease on both individuals and society as a whole are the main goals of this project. By harnessing the power of cutting-edge data analytic tools, we may improve medical studies connected to cardiology, develop better prevention and treatment programs, empower people to take control of their health, and more.

The goal of this research is to create a heart failure prediction model using the currently accessible datasets. This involves identifying key heart failure risk factors using sophisticated data analysis methods, then normalizing the data to develop a model that exactly predicts the possibility of a heart failure.

This research seeks to increase our comprehension of heart failure and provide more effective preventative and treatment methods by employing data analysis to get insights into the underlying causes of this ailment. In order to create individualized treatment regimens that address each patient's unique needs and improve the quality of care, healthcare professionals must first identify the major risk factors that contribute to heart failure.

Predictive modelling can assist healthcare professionals in identifying at-risk patients and initiating early interventions, potentially averting negative outcomes like heart attacks and strokes. This can be crucial for people whose risk of developing heart failure is high because of their age, their family history, or their lifestyle choices.

Ultimately, this project's major objective is to enhance the treatment outcomes for people who have had heart attacks by employing data analysis to learn more about the underlying causes of heart failure. Healthcare professionals can enhance patient care and lessen the impact of heart disease on individuals and society at large by creating more precise predictive models and customizing treatment approaches for each patient.

**Proposed Methods and Tools :-**

The proposed method for this project combines client-side programming, data visualization, and database management tools to build a predictive model for heart failure.

In this project, HTML, a language often used to create online pages and web apps, was employed to design the client software. As the project's front end, the client application provides a straightforward user interface for accessing the data and corresponding with the prediction model.

The main component of this study is data visualization, which helps us understand the root causes of heart failure. To visualize data, we advise mixing Tableau with Python. Python is a powerful and flexible programming language that is commonly employed for data analysis and machine learning..

We suggest using MySQL Workbench, a potent tool for data modelling, data querying, and data migration, to store, manage, and query the data. We may import data from other sources, construct and manage the database schema with MySQL Workbench, and run sophisticated queries to draw out important information from the data.

Ultimately, by integrating these tools and methods, we want to create a thorough and useful predictive model for heart failure that will assist medical professionals in better comprehending the underlying causes of this problem and creating more efficient preventative and treatment plans.

**Excepted outcomes :-**

The creation of a new heart attack dataset with real-time data, socioeconomic determinants of health, high-quality data, and standardized data is the project's anticipated end result. This new dataset will be useful for a number of reasons, including:

First, we can guarantee that the predictive model is based on the most recent data by including real-time data into the dataset. In order to account for changes in risk factors over time, this can increase the model's accuracy.

Furthermore, by integrating social determinants of health in the dataset, we may develop a more thorough understanding of the risk factors for heart attacks. Poverty, education, and access to healthcare are just a few examples of the social determinants of health that can have a big impact on someone's risk of developing heart disease. We can gain a more sophisticated knowledge of the underlying causes of heart attacks by incorporating these characteristics in the dataset.

Lastly, we may increase the predictive model's validity and reliability by making sure the data is high-quality and standardized. Better patient outcomes may result from healthcare professionals making more educated judgments about preventative and treatment plans as a result.

Therefore, it is anticipated that the creation of a new heart attack dataset including real-time data, socioeconomic determinants of health, high-quality, and standardized data will outperform existing datasets and offer more precise and practical forecasts. This can therefore result in improved heart attack prevention and treatment methods, enhancing patient care and lessening the impact of heart disease on both individuals and society as a whole.

**Background**:

Myocardial infarction, another name for a heart attack, is a critical and sometimes fatal condition that develops when blood flow to the heart is obstructed. In severe circumstances, this can cause death, heart failure, or stroke in addition to harming the heart muscle. According to the World Health Organization, heart attacks are the top cause of death worldwide, accounting for 17.9 million fatalities annually.

Age, sex, high blood pressure, high cholesterol, smoking, diabetes, and a family history of heart disease are some of the risk factors linked to heart attacks. Early heart attack detection and prevention can help lower the likelihood of negative outcomes and improve patient outcomes. Age, sex, blood pressure, and cholesterol levels are a few clinical and demographic variables that are frequently used in traditional risk assessment techniques for heart attack prediction. These instruments may not fully account for all significant aspects for heart attack prediction due to their poor accuracy.

In recent years, machine learning techniques have emerged as a promising approach for heart attack prediction. Machine learning algorithms can analyze large and complex datasets and identify patterns and relationships that may not be apparent using traditional statistical methods. These algorithms can also handle a large number of features and reduce overfitting, which can improve the accuracy of the model.

Many machine learning techniques, such as logistic regression, decision trees, support vector machines, and neural networks, have been used to the prediction of heart attacks. Each algorithm has advantages and disadvantages, and the best one to use relies on the details of the data and the objectives of the study.

The Random Forest method is one of the most widely utilized machine learning algorithms for heart attack prediction. The Random Forest algorithm uses numerous decision trees to increase accuracy and decrease overfitting. It is an ensemble learning technique. In order for the algorithm to function, a subset of characteristics is randomly chosen, and numerous decision trees are trained on various subsets of the data. The program then combines each tree's prediction into a single final prediction.

For predicting heart attacks, the Random Forest approach provides a number of advantages over other machine learning techniques. It can manage high-dimensional data and lessen overfitting, improving the model's generalizability. The technique is capable of handling categorical characteristics and missing data, and it is computationally efficient. The algorithm may also reveal the significance of each feature for heart attack prediction, allowing for the identification of the clinical and demographic variables that are most crucial for heart attack prediction.

Despite the potential advantages of machine learning algorithms for heart attack prediction, there are several challenges and limitations to their use. One challenge is the availability and quality of data. Heart attack prediction models require large and diverse datasets that capture the full range of clinical and demographic factors that may be relevant for heart attack prediction. Additionally, the data must be of high quality and free from errors and biases to ensure accurate predictions.

Another challenge is the interpretability of the models. Machine learning algorithms such as Random Forest are often referred to as "black box" models because they are difficult to interpret and understand. This can make it challenging for clinicians to use the models in clinical decision-making, and may lead to skepticism and mistrust of the models.

In conclusion, machine learning algorithms such as the Random Forest algorithm have the potential to improve the accuracy and early detection of heart attack. These algorithms can handle large and complex datasets, identify important clinical and demographic factors for heart attack prediction, and reduce overfitting. However, there are several challenges and limitations to their use, including the availability and quality of data, and the interpretability of the models. Further research is needed to address these challenges and validate the accuracy and utility of machine learning algorithms for heart attack prediction

**Brief summary of the 3 models:**

1. **Random Forest Model**

An example of supervised machine learning is the random forest method. In supervised machine learning, a model is developed using a range of inputs (or features) linked to a predetermined result. Using decision trees is the basis of this model. The provided dataset was dissected and examined to produce these trees. The decision tree value for the subgroup is averaged to produce the desired outcome. With more trees present, the outcome is more accurate.

This method makes use of the bagging code, often known as Bootstrap aggregation. In this case, the dataset is segmented into phases before being evaluated as previously said. This would generate a range of outputs, which are then combined and the average/most often used result is used as the result.

1. **Decision trees**

A decision tree is a tree-like model that maps observations about patients to conclusions about their likelihood of heart failure. It is a flexible and interpretable model that can capture complex relationships between patient features. Decision trees are essentially related questions that are arranged hierarchically and feature a condition for an appropriate response. Relationally, they begin at the top. The size of the tree might increase considerably based on the inputs and characteristics required. Additionally, low-importance characteristics can be removed, which will result in the removal of some branches. The decision tree's final output would result from the aggregation of all the decision trees. The result in this case can be binary or non- binary.

1. **Logistic Regression**

An illustration of supervised learning is logistic regression. It is one of the easiest and simplest algorithm. It is widely used for solving binary problems. It classifies data on the input variable. Here, we receive a yes/no answer to the forecast. As a result, the probability essentially lies between 0 and 1. Logistic regression is a linear model. Here the features in order with their importance can be seen clearly. A logistic function is used to transform the probability result. In terms of the outcome, this is different from linear regression. Instead of a numeric number, the binary value prevails here.

**Work Flow**

First, the data set is acquired. It has been downloaded to use Excel. then a CSV file was saved. Python will be used in Jupyter for this. There are several python libraries utilized, including pandas, numpy, and matplotlib. After that, the step of data preparation begins, when we will clean the dataset. There are no more blank fields to fill up. From the data tables, characteristics and correlations are sought after. The data tables are divided. A machine learning approach is used for the dataset. Then, predictive analysis would have its characteristic. Three models—the random forest method, decision trees, and logistic regression—are used to compare the optimum probability outcome. Using AUC(Area Under the Curve), ROC (Receiver Operator Characteristic Curve) we can compare the prediction accuracy of the various models.

**Background work from the reference papers:**

**Work on Logistic Regression approach:**

The data set is first acquired in the logistic technique. Data preprocessing is the following stage, where features are picked and data is cleaned. Following the data selection is the categorization modeling and performance testing. After training, the final set of parameters is obtained and incorporated to the model's formula. The test results are then used to determine the mistake. The closer the parameters are to the ideal one, the lower the error is. Starting here, a number of variables are used to compute and examine the accuracy.

**Work on Random Forest approach:**

This process also stays the same. It is evident that the data has been gathered, preprocessed, and cleared of all null values. The ML method is then used. Here, a large number of decision trees are built, requiring a lot of processing power. In this method, a random forest model is trained using a dataset of patient information, including age, sex, blood pressure, serum creatinine, and other pertinent medical data. The following are the steps that can be taken to apply the random forest approach for heart failure analysis:

Data collection: Collect a dataset of patient information, including their medical history, demographics, and other relevant factors.

Data cleaning: Clean the data by removing any missing or invalid values, and normalize the data to make it easier to process.

Feature selection: Identify the most important features in the dataset using techniques such as correlation analysis, information gain, or chi-square test.

Training the model: Split the dataset into training and testing sets, and train the random forest model using the training data.

Model evaluation: Evaluate the performance of the model by comparing its predictions to the actual outcomes in the testing set. Use metrics such as accuracy, precision, recall, and F1-score to evaluate the performance of the model.

Model optimization: Fine-tune the model by adjusting its hyperparameters, such as the number of trees, the depth of each tree, and the minimum number of samples required to split a node.

**Work on Decision Tree approach:**

The same flow goes on here as well. The data is collected and cleaned. The machine learning algorithm is then used.

The decision tree is a popular machine learning algorithm that is used for classification and regression tasks. A decision tree is a tree-like model of decisions and their possible consequences, including chance events, resource costs, and utility. Here is a general workflow for building a decision tree model:

Data Collection: Collect a dataset of relevant information for the task at hand, such as demographic information, medical history, and other factors that may affect the likelihood of heart failure in patients.

Data Cleaning and Preprocessing: Clean and preprocess the dataset by removing any missing or invalid values, and normalizing the data to make it easier to process.

Split the Data: Split the dataset into training and testing sets, typically using a 70/30 or 80/20 split, respectively.

Feature Selection: Identify the most important features in the dataset using techniques such as correlation analysis, information gain, or chi-square test.

Build the Tree: Build the decision tree model by recursively splitting the dataset into subsets based on the most informative features. This process involves selecting the best feature to split the dataset on, based on the criterion that optimizes the model's accuracy, such as the Gini index or entropy.

Pruning the Tree: Once the decision tree has been built, it is often pruned to improve its generalization performance. This involves removing branches or nodes that are not useful for predicting heart failure or that overfit the training data.

Model Evaluation: By contrasting the decision tree model's predictions with the actual results in the testing set, determine how well it performed. To assess the model's performance, use metrics such as accuracy, precision, recall, and F1-score.

Model Optimization: Fine-tune the model by adjusting its hyperparameters, such as the maximum depth of the tree, the minimum number of samples required to split a node, and the criterion used to measure the quality of a split.

Prediction: Once the decision tree model has been optimized, use it to predict the likelihood of heart failure in new patients based on their medical information.

In conclusion, creating a decision tree model entails gathering and cleaning the data, choosing the most useful features, building the tree, pruning it, assessing its effectiveness, optimizing it, and utilizing it to generate predictions.

**Random Forest Model**

Random Forest is a machine learning algorithm used for classification, regression, and other tasks. It creates an ensemble of decision trees, where each tree is trained on a random subset of the input data and a random subset of the features. The algorithm then combines the predictions of each tree to make a final prediction.

In our study, we used Random Forest to make predictions about the likelihood that a patient will have heart disease based on a variety of medical indicators, including age, sex, blood pressure, cholesterol levels, and others. On a dataset including data on patients who had been identified as having heart disease or as being at risk for developing heart disease, we trained the model.

We divided the dataset into training and testing sets, using 80% of the data for training and 20% for testing, in order to train the model. The Random Forest model was subsequently trained on the training set, and its performance was assessed on the testing set.

We used various evaluation metrics such as accuracy, precision, recall, and F1 score to assess the performance of the model. These metrics give us an idea of how well the model is able to correctly classify patients who have heart disease or who are at risk for heart disease.

Finally, we used a heatmap to visualize the results of the model on the testing set. The heatmap shows the number of true positive, false positive, true negative, and false negative predictions made by the model. This helps us to get a better understanding of how well the model is performing and where it may be making mistakes.

When the model had been trained, we assessed its effectiveness using a number of metrics, including accuracy, precision, recall, and F1 score. The model's accuracy, precision, recall, and F1 score were all 0.933, 0.88, and 0.957, respectively, showing that it did a respectable job of predicting whether a patient has heart disease.

Finally, we used seaborn to create a heatmap of the confusion matrix to visualize the model's performance. The heatmap showed the number of true positives, true negatives, false positives, and false negatives, allowing you to see where the model made mistakes and identify areas for improvement.

**The Heat Map:**

Chart

Description automatically generated

The x-axis of the heatmap shows the predicted labels, while the y-axis shows the true labels. The annotations within each cell represent the number of predictions that correspond to the intersection of the predicted label and true label.

The heatmap provides a visual representation of the confusion matrix, which allows us to quickly identify any patterns or trends in the model's predictions. It can be used to evaluate the overall performance of the model, as well as identify areas where the model may be making more errors.

In the context of our project, the heatmap allowed us to visualize the confusion matrix and evaluate the performance of the random forest classifier. We were able to see how well the model was able to predict the presence or absence of heart disease, and identify any patterns or trends in the model's predictions. This helped us to determine the accuracy, precision, recall, and F1 score of the model, which are important metrics in evaluating the performance of a classification model.

The diagonal blocks in blue represent the correct predictions, where the true label and the predicted label match. The diagonal blocks indicate the number of true positives and true negatives in the classification results. True positives are cases where true positives are instances when the model projected a positive outcome and the true label confirmed that prediction, as opposed to true negatives, where the model expected a negative event and the true label confirmed that prediction.

The blocks in other colors represent the incorrect predictions, where the true label and the predicted label do not match. The lighter blue and green blocks represent false positives and false negatives, respectively. False positives are cases where the model predicted a positive outcome but the true label was negative. False negatives are cases where the model predicted a negative outcome but the true label was positive.

The numbers inside each block represent the count of observations in each category. These counts can be used to calculate various performance metrics like accuracy, precision, recall, and F1-score.

**Logistic regression**:

The mathematical approach of logistic regression is used to ascertain the relationships between two data components. Using this relationship, the value of one of those parameters is then predicted based on the other. Frequently, the forecast only offers a few options, like yes or no.

Consider the scenario in which you want to foretell whether a website user will click the checkout button in their shopping cart or not. Logistic regression is the process of examining past visitor behavior, including the length of time spent on a website and the number of items in a shopping cart. It determines that in the past, customers were deemed to have hit the checkout button if they stayed on the website for more than five minutes and added three or more items to their cart. Using this information, the logistic regression algorithm may then predict the behavior of a brand-new website user.

**Relevance in heart attack prediction**:

Many risk factors, including age, sex, smoking status, blood pressure, cholesterol levels, family history, and other factors linked to the possibility of having a heart attack, can be analyzed using logistic regression. The occurrence or absence of a heart attack is often utilized as the dependent variable, with these risk factors acting as the independent variables.

Each person receives a likelihood score from the logistic regression model, which may be used to categorize them as having a high or low risk of having a heart attack. For those who are more at risk, this information can be used to create specialized preventative and treatment plans.

**Flowchart of the Logistic regression Model**:

Diagram, schematic

Description automatically generated

**Dataset**:

Data was gathered from the kaggle. The dataset has 25 attributes and 300 records. The output value or forecast value of the patients with heart disease is one of the twenty five parameters utilized as the eigenvalues for the forecast of heart disease ( ‘num’ means Numeric, and ‘nom’ means Nominal)

From Appendix A, we can see the attributes of the dataset.

**Data preprocessing**:

Because the data set contains missing data and the output value ranges from 0 to 4 with varying degrees of illness, but logistic regression corresponds to dichotomy, it is crucial to preprocess the data before further analysis. After the incomplete data were removed, the multicategory value and binary value of the dataset's predictive value attributes for having heart disease were altered. Diagnostic results from the ranges of 2 and 4 are modified to 1. Only the values 0 and 1, which represent the absence of cardiac disease and the likelihood of developing it, are included in the final data set.

**Feature Selection**:

The experiment must uncover the primary causes of heart diseases in addition to estimating the likelihood of developing heart disease. In order to identify the associated pathogenic variables impacting heart problems and make recommendations for the protection of physical health based on these, features should be extracted prior to data analysis. The graph below shows the relationship between each characteristic value.

Chart

Description automatically generated

**Splitting the Dataset**:

We split the dataset into 80:20,80% is used for training the model and the rest 20% is used for testing the model

**Classification Modelling and performance testing**:

After the feature selection is finished, it is decided that a set of parameters will finally be selected in line with the parameters into the model, and then checked with test data. Gradient descent is used to determine the new parameters for the model. As the probability gets closer to the real value, the error gets smaller and smaller until it tends to a fixed value. After training, the final parameters are gathered, added to the model's equation, and the error is then calculated using the test data. The closer the parameters are to the ideal one, the lower the error is.The final model's prediction model has an accuracy of 85%.Conclusion:

Using data mining technologies, raw data is examined to produce fresh perspectives and precise forecasts with the purpose of preventing disease.

In this study, logistic regression models were employed to investigate the viability of heart disease prediction. Using the data set provided by UCI, experiments were carried out, and the outcomes were assessed.Chestpain\_type,Resting\_ECG,ST\_Slope,Sex,Anaemia,Diabetes, high\_blood\_pressure were discovered to be significant characteristics of logistic regression models affecting heart disease based on the p value.

To create the prediction model, the primary influencing factors and logistic regression technology are employed, and the model's accuracy is compared to the model suggested by the previous research.The proposed categorization model is very accurate and has some research value, according to the test findings.To further increase the accuracy, new feature extraction techniques and model parameters can be chosen in the future research.

Related work:

“Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning” In the Paper they used logistic regression to train and test the data and got an accuracy of 87%.

“Logistic regression technique for prediction of cardiovascular disease” In the paper they split the data into five ratios and got an accuracy of 87.10% for 90:10 for training and testing ratio.

“Logistic Regression Models in Predicting Heart Disease” In the Paper they used logistic regression, Support Vector Machine, KNN, Neural Network, Naïve Bayes to train and test the data and got an accuracy of 85.86% for logistic regression.

References:

[1] Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning.

[2] Logistic regression technique for prediction of cardiovascular disease

[3] Logistic Regression Models in Predicting Heart Disease

**Decision Tree :-**

For classification and regression issues, a decision tree machine learning model is used. It is an illustration of a supervised learning technique that, prior to making a final prediction or judgment, splits the data into smaller subgroups based on the properties of the data.

The decision tree model is made up of leaf nodes, which represent the class label or numerical value, edges, which reflect the decision rule, and nodes, which represent a characteristic or attribute. Each node in the tree symbolizes a question that has been asked about the data, and the path that must be taken to reach the final prediction or judgment depends on the response to that question.

The tree is constructed by selecting the feature that best splits the data into smaller, more homogeneous subsets. This process is repeated recursively until a stopping criterion is met, such as a maximum depth of the tree or a minimum number of samples required to split a node.

Decision trees are easy to interpret and visualize, making them popular in various fields, including finance, healthcare, and marketing. They are also used in ensemble methods, such as random forests and gradient boosting, to improve their predictive accuracy.

Algorithms that are used to find different methods to segment a data set into segments create decision trees. An inverted decision tree is created by these segments. The root node of the decision tree is located at the top of the tree.

ID3:

Iterative Dichotomiser 3 is referred known as ID3. One decision tree model, ID3, constructs a decision tree from a predetermined set of training examples. The future samples are categorized using the resulting tree.

C4.5

The ID3 induction algorithm C4.5 is the most recent iteration. It represents an advancement of the ID3 algorithm. Thus, an ID3-like decision tree is produced. It constructs a decision tree using the idea of information entropy using a training set of data. The term "statistical classifier" is therefore frequently used to describe C4.5. C4.5 is a well-liked open-source data mining tool.

C5

This model extends the decision tree technique in C4.5. C4.5 and C5.0 both have the ability to produce classifiers that are specified as rulesets or decision trees. Due of their simplicity and readability, rulesets are preferred in many applications. The two basic differences are tree sizes and computation times. C5.0 generates trees considerably faster and in smaller sizes than C4.5.

J48

The J48 decision tree is the implementation of the ID3 algorithm, which was developed by the WEKA project team. A simple C4.5 classification decision tree is J48. Using this technique, a tree is constructed to symbolize the classification process. After being created, each tuple in the database is subjected to the tree to establish its classification.

**Decision tree types**:

There are many different configurations of decision trees. They differ from one another due to the mathematical model that was used to select the splitting property and derive the decision tree rules. Information Gain, Gini Index, and Gain Ratio are the top three most popular research test kinds. Pick trees.

**Information Gain**:

The term "entropy" refers to the increase of information. With this method, the splitting property that minimizes entropy value and maximizes information gain is chosen. Calculating the information gain for each attribute is necessary to determine the splitting attribute of the decision tree. The property that maximizes information gain is then chosen. It is the discrepancy between the quantity of information originally provided and that which is now required.

**Gain Index**:

Data impurity is quantified using the Gini Index. For each attribute in the dataset that is available, the Gain index is calculated.

**Gain Ratio**:

Gain Ratio, a variation of Information Gain, is used to lessen the impact of the bias brought on by its application. The test with several outcomes is favored by the information gain measure. This indicates that it favors choosing qualities with a lot of possible values. Gain Ratio modifies each attribute's Information Gain to account for the diversity and uniformity of the attribute values.

Information Gain / Information Split = Gain Ratio

where the value representing the split information is derived from the frequency table's column sums.

**Pruning**:

Reduced error pruning is used to prune the extracted decision rules after extracting the decision tree rules. One of the most effective and quick pruning techniques, reduced error pruning generates precise and compact decision rules. More compact decision rules are produced and the number of extracted rules is decreased when reduced error pruning is used.

**Performance Evaluation**:

Sensitivity, specificity, and accuracy calculations were made to assess each combination's performance. The data is split into training and testing data with 10-fold cross validation in order to determine the stability of performance.

Sensitivity = True Positive/ Positive

Specificity = True Negative/ Negative

Accuracy = (True Positive + True Negative) / (Positive + Negative)

**Relevance in Heart Attack Prediction :-**

Decision tree models are relevant in heart attack prediction because they can help identify the most important risk factors associated with heart attacks and provide a prediction of the likelihood of a heart attack based on those risk factors.

Heart attacks are caused by a combination of various risk factors, including age, gender, family history, smoking habits, blood pressure, cholesterol levels, etc. A decision tree model can analyze these risk factors and create a tree-like structure to predict the likelihood of a heart attack.

By using a decision tree model, healthcare professionals can identify patients who are at higher risk of a heart attack and take appropriate preventive measures, such as lifestyle modifications or medication, to reduce their risk. Moreover, decision tree models are easy to interpret and visualize, which can help healthcare professionals explain the risks and benefits of different interventions to patients.

Therefore, decision tree models are relevant in heart attack prediction as they can help healthcare professionals make informed decisions about the prevention and management of heart attacks based on a patient's risk factors.

**Dataset**:

We used the same dataset as the other two models.

**Results**:

For categorizing Heart Disease based on Clinical Features against unpruned, pruned, and pruned with reduced error pruning method,We created model using the j48 Decision Tree and decision tree is shown below.The accuracy for the model is 87.5%

A picture containing text, screenshot, post-it note

Description automatically generated

**Special Observations:**

Apart from the challenges and limitations identified in the review, there were several special observations that emerged during the analysis. One of the most notable observations is the potential for the use of artificial intelligence (AI) and machine learning (ML) in predicting heart attacks. Recent advancements in AI and ML have revolutionized the field of medical research by providing powerful tools for analyzing large datasets and making predictions with high accuracy. These technologies have the potential to significantly improve the accuracy of heart attack prediction databases by identifying complex relationships between various risk factors and predicting heart attacks more accurately.

Another special observation is the need for comprehensive clinical data that includes genetic data and biomarkers. Genetic data can be used to identify individuals who may be at a higher risk of developing heart attacks due to their genetic makeup. Biomarkers, such as troponin, can also provide valuable information about the severity of a heart attack and the risk of future cardiac events. The inclusion of these data types in heart attack prediction databases can significantly improve the accuracy of predictions and enable early detection of heart attacks.

Furthermore, it is important to consider the ethical implications of using heart attack prediction databases. Privacy concerns must be taken into account when collecting and storing patient data. Patients must be informed about how their data will be used and must give their consent before their data is included in the database. Additionally, there is a risk that the use of heart attack prediction databases may lead to stigmatization and discrimination against individuals who are identified as being at a higher risk of developing heart attacks.

Need for Privacy and Security: Privacy and security concerns are paramount when it comes to collecting and storing personal health information. Most heart attack prediction databases contain sensitive information such as patient health records and diagnostic test results. Therefore, it is crucial to ensure that the data is collected, stored, and analyzed in a secure and ethical manner. Inadequate data security measures can lead to data breaches, which may compromise patient privacy and have legal implications.

Need for Interoperability: Interoperability refers to the ability of different systems to communicate with each other seamlessly. The lack of interoperability is a significant challenge in the field of heart attack prediction databases. Data stored in different databases may be in different formats and may not be compatible with each other. This makes it difficult to integrate data from different sources and analyze it comprehensively. Standardization of data formats and the development of open-source data sharing platforms can help address this challenge.

**Need for Real-time Data**: Most heart attack prediction databases contain retrospective data, which means that the data is collected after the event has occurred. This limits the ability to predict and prevent heart attacks in real-time. Real-time data collection and analysis can provide clinicians with timely and accurate information, enabling them to make better decisions regarding patient care.

**Need for Personalization**: Heart attack prediction databases often use a one-size-fits-all approach to predict the risk of heart attacks. However, individual patients have unique risk factors and health histories that can affect their likelihood of experiencing a heart attack. Personalized risk assessment models that take into account individual patient characteristics can provide more accurate predictions and enable targeted interventions to prevent heart attacks.

**Need for Transparency and Interpretability**: The lack of transparency and interpretability is a significant limitation of heart attack prediction databases. Most algorithms used to predict heart attacks are black boxes, meaning that it is not always clear how they arrive at their predictions. This makes it difficult for clinicians to trust the predictions and may lead to a lack of adoption of heart attack prediction databases. Developing more transparent and interpretable algorithms can help address this limitation.

**Need for Continuous Improvement**: Heart attack prediction databases must continuously evolve and improve to remain effective. As new risk factors are identified, new data sources become available, and new machine learning algorithms are developed, heart attack prediction databases must incorporate these advancements to improve their predictions continually.

Overall, these special observations highlight the need for more secure, interoperable, and personalized heart attack prediction databases that can provide real-time and transparent predictions. Addressing these challenges and limitations can help improve the accuracy and effectiveness of heart attack prediction databases, leading to better patient outcomes and reduced healthcare costs.

**Challenges:**

One of the significant challenges in developing accurate heart attack prediction databases is the lack of standardized data collection techniques. Different studies use different methods to collect data, which can lead to inconsistencies and inaccuracies in the data. The use of standardized data collection techniques can help to ensure that the data is accurate and can be compared across studies.

Another challenge is the lack of diversity in the datasets used to develop heart attack prediction databases. Most studies are conducted in high-income countries and focus on specific demographic groups, such as Caucasian males. This can lead to over-representation of particular groups in the datasets and may limit the generalizability of the findings to other populations. It is essential to include data from diverse populations to ensure that the database is representative of the general population.

Moreover, the lack of comprehensive clinical data, such as genetic data and biomarkers, is also a significant challenge in developing accurate heart attack prediction databases. These data types provide valuable information about an individual's risk of developing heart attacks and can significantly improve the accuracy of predictions. However, the collection and storage of genetic data raises ethical concerns, and the use of biomarkers can be expensive and time-consuming.

Another challenge in heart attack prediction databases is the lack of comprehensive clinical data. Many of the current heart attack prediction databases lack detailed clinical data, such as biomarkers and genetic data, which can be used to identify risk factors and improve prediction accuracy. While some databases do contain clinical data, the data is often incomplete, inconsistent, or not standardized. This can lead to inaccurate predictions and can limit the ability of researchers to identify new risk factors for heart attacks. One potential solution to this challenge is to develop standardized clinical data collection protocols that can be used across different databases and research studies. This would enable researchers to collect more comprehensive and accurate clinical data, which could improve the accuracy of heart attack predictions.

**Limitations:**

One of the limitations of this review is the reliance on published studies. The studies included in this review were conducted in different geographical areas and used different data collection techniques, which makes it challenging to compare the findings across studies. Additionally, most of the studies identified were conducted in high-income countries, which may limit the generalizability of the findings to low-income countries. Future research should focus on including data from diverse populations and using standardized data collection techniques to ensure that the findings can be compared across studies.

Another limitation is the lack of focus on the use of AI and ML in heart attack prediction databases. While the potential benefits of these technologies were mentioned in the special observations section, they were not explored in detail in this review. Future research should focus on the use of AI and ML in heart attack prediction databases and explore their potential to improve the accuracy of predictions.

In conclusion, heart attack prediction databases play a crucial role in the early detection and prevention of heart attacks. However, these databases face several challenges and limitations that can affect the accuracy of predictions and the generalizability of findings. These challenges include the lack of sufficient data, the lack of diversity in datasets, inconsistent data collection techniques, inaccurate reporting of data, the lack of comprehensive clinical data, the lack of standardization in the collection, analysis, and reporting of data, and issues related to data privacy and security. Addressing these challenges will require the development of standardized protocols for data collection, analysis, and reporting, as well as the development of more diverse datasets that represent a wide range of demographic groups and geographical areas. By addressing these challenges, researchers can develop more accurate and reliable heart attack prediction databases that can help to save lives and reduce healthcare costs.

**References:**

Baum A, Scarpa J. Personalized medicine for cardiovascular disease: from genomics to the clinic. Genome Med. 2017;9(1):1-12. doi:10.1186/s13073-017-0412-3

Gaskin DJ, Dinwiddie GY, Chan KS, McCleary R, Spence R. Social determinants of health and cardiovascular disease outcomes: a scientific statement from the American Heart Association. Circulation. 2021;143(5):e35-e74. doi:10.1161/CIR.0000000000000930

Klonoff DC, Kerr D. Digital diabetes and cardiovascular disease: a potential game-changer that favors the payer. J Diabetes Sci Technol. 2014;8(4):689-695. doi:10.1177/1932296814533014

Padilla P, Villalba K, Spiegelman J, Rabadan R. Data-driven approaches for predicting cardiovascular disease risk. J Cardiovasc Transl Res. 2021;14(1):1-8. doi:10.1007/s12265-020-10035-8.

Vaidya A, Aalst J, Desai M. Artificial Intelligence in Healthcare: Past, Present and Future. Am J Med. 2019;132(7):795-801. doi:10.1016/j.amjmed.2018.11.025

Prediction of Heart Disease Using Decision Tree Approach R. Vijaya Kumar Reddy, K. Prudvi Raju , M. Jogendra Kumar, CH. Sujatha , P. Ravi Prakash

Effective Heart Disease Prediction Using Machine Learning Techniques Chintan M. Bhatt, Parth Patel , Tarang Ghetia and Pier Luigi Mazzeo

Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques Rajkumar Gangappa Nadakinamani, A. Reyana , Sandeep Kautish , A. S. Vibith, Yogita Gupta, Sayed F. Abdelwahab , and Ali Wagdy Mohamed ; doi.org/10.1155/2022/2973324

Prediction of Heart Diseases using Random Forest Madhumita Pal , Smita Parija

Current state of artificial intelligence-based algorithms for hospital admission prediction in patients with heart failure: a scoping review P. M. Croon, J. L. Selder , C. P. Allaart , H. Bleijendaal, S. A. J. Chamuleau1 , L. Hofstra , I. Išgum, K. A. Ziesemer , and M. M. Winter

**Appendix**

Attributes of the dataset:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Type** |
| Patient ID | Unique ID assigned to patients | Num |
| ChestPainType | chest pain type--- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic | Nom |
| Resting BP | Resting Blood Pressure | Num |
| Cholestrol | Serum cholestoral in mg/dl | Num |
| FastingBS | Fasting blood sugar > 120 mg/dl--Value 1-yes--Value 0-No | Num |
| RestingECG | Resting electrocardiographic results: -- Normal: normal -- ST\_Resting: having ST-T wave abnormality -- LVH\_Resting: showing probable or definite leftventricular hypertrophy by Estes' criteria | Nom |
| MaxHR | Maximum Heart Rate | Num |
| ExerciseAngina | Excerise include Angina--Value 1-yes--Value 0-No | Num |
| Old Peak | ST depression induced by exercise relative to rest | Num |
| Slope | The slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping | Nom |
| HeartDisease | Is the patient previously diagnosed with heartdisease?:--Value 1-yes--Value 0-No | Num |
| Age | Age of the patient | Num |
| Sex | Sex of the patient--Value 1-yes--Value 0-No | Num |
| anaemia | Hemoglobin count > 16.6 g/dl--Value 1-yes--Value 0-No | Num |
| creatinine\_phosphokinase | CPK levels in the body | Num |
| Diabetes | Does the patient has diabetes--Value 1-yes--Value 0-No | Num |
| ejection\_fraction | Percentage of Ejection Fraction | Num |
| high\_blood\_pressure | Does the patient has High Blood Pressure--Value 1-yes--Value 0-No | Num |
| platelets | platelets count | Num |
| serum\_creatinine | Amount of creatinine in the patient blood. | Num |
| serum\_sodium | Amount of sodium in the patient blood. | Num |
| smoking | Does the patient smoke--Value 1-yes--Value 0-No | Num |
| DEATH\_EVENT | Does the patient previously has any medical near death event--Value 1-yes--Value 0-No | Num |