**HEART DISEASE PREDICTION**

**Sai Karthik Nagabhairava, Ram Sai Repalle, Bharath Kumar Pakam, Deepshika Yedla**

**Abstract:**

A disease is a particular abnormal condition that negatively affects the structure or function of all or part of an organism, and that is not due to any immediate external injury. Disease can affect both physically and mentally and disease can alter the affected person’s perspective on life. heart diseases are mostly unpredictable. There are various heart diseases and in those main and most common disease are cardiovascular diseases and heart attack. We use Machine learning techniques which are very efficient in providing accurate results. ML algorithms has shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health industry. The dataset we have taken is a public dataset UCI heart disease dataset, Cleveland heart disease dataset. In this project, we use different ml techniques like linear model, Random Forest, Decision tree, k nearest neighbor, for the heart disease prediction to get better accuracy. Although, there are many different combinations of ML algorithms are published in papers, there are mostly efficient, and there are different combinations of algorithms exist, such as linear model along with regression model, etc. which produce excellent accurate results when trained with huge amount of data. In this project we will be looking at Some hybrid algorithms so that we will tend to get the result with much more efficiency than the previous hybrid algorithms used.

**Introduction:**

Data mining (DM) is the extraction from large data sets of useful information that results in the data being predicted or represented using techniques such as classification, clustering, association, etc. In the healthcare industry, data mining has found broad applicability, such as classifying optimal methods of treatment, predicting risk factors for illness, and identifying appropriate cost models for patient care. Data mining models have been used to study illnesses such as diabetes, asthma, cardiovascular diseases, AIDS, and so on. Different data mining techniques have been used to build models in healthcare research, such as naïve Bayesian classification, artificial neural networks, support vector machines, decision trees, logistic regression.

Every year, an estimated 17 million people die from cardiovascular diseases (CVD). Although these diseases can be managed, their early prognosis and the assessed risk of a patient are important to reduce their high mortality rates. Coronary heart disease, cardiomyopathy, hypertensive hearing impairment, heart failure, etc. are common cardiovascular disorders Smoking, diabetes, lack of physical activity, hypertension, a high cholesterol diet, etc. are common causes of heart disease.

Study using data mining in the field of cardiovascular diseases has been an ongoing endeavor involving high precision prediction, treatment, and risk score analysis. Several CVD surveys have been carried out with the Cleveland Heart Clinic data collection being the most popular one. As such, the Cleveland Heart Disease Database (CHDD) has been dubbed the de facto heart disease research database.

Owing to many contributory risk factors such as diabetes, high blood pressure, high cholesterol, irregular pulse rhythm, and several other factors, it is difficult to detect heart disease to find out the seriousness of heart disease among humans, different techniques have been used in data mining and neural networks. Based on various methods such as K Nearest Neighbor Algorithm (KNN), Decision Trees (DT), and Naïve Bayes, the seriousness of the disease is graded (NB). The essence of heart disease is complicated, and the disease must therefore be carefully treated. Failure to do so may affect the heart or cause premature death. The medical research and data mining perspective is used for the identification of different forms of metabolic syndromes. In the prediction of heart disease and data investigation, data mining with classification plays a major role.

In general, neural networks are seen as the best method for the detection of diseases such as heart disease and brain disease. The suggested approach we use has 13 characteristics for the prediction of heart disease. The findings indicate an increased degree of productivity in works such as opposed to the current methods. In the medical sector, carotid artery stenting (CAS) has also become a prevalent mode of care in recent years. The CAS supports the frequency of significant adverse cardiovascular events (MACE) in elderly patients with heart disease. Their assessment is very critical. We generate results using the ANN Artificial Neural Network, which produces strong output in heart disease prediction. Methods of the neural network are applied, integrating not only posterior probabilities but also expected values from various predecessor techniques.

The rules were implemented randomly with encoding techniques that resulted in an overall increase in accuracy. Based on symptoms, including pulse rate, sex, age, and many others, heart disease is expected.

In the field of medicine and science, we also present Computer Assisted Decision Support System (CADSS) experiments used to classify the characteristics of a machine here. In previous work, the use of data mining techniques in the healthcare industry has been shown to take less time with more reliable outcomes for disease prediction. Using GA, we suggest the diagnosis of heart disease. For tournament selection, crossover and mutation, this approach uses successful association rules inferred from the GA, resulting in the new proposed fitness function. We use the well-known Cleveland data set, which is gathered from a UCI machine learning repository, for experimental validation. Later, as compared to some of the known supervised learning methods, we can see if our findings prove to be prominent. Particle Swarm Optimization (PSO) is the most effective evolutionary algorithm, and some heart disease rules are developed. The rules were implemented randomly with encoding techniques that resulted in an overall increase in accuracy. Based on symptoms, including pulse rate, sex, age, and many others, heart disease is expected.

**Tasks and goals:**

1. **Data collection:**

The data collection process involves the selection of quality data for analysis. we used heart disease dataset taken from uci.edu site for machine learning implementation. The job of a data analyst is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.

1. **Data visualization:**

A large amount of information represented in graphic form is easier to understand and analyze. As Some companies specify that a data analyst must know how to create slides, diagrams, charts, and templates. In our approach, the heart disease rates is shown as data visualization part.

Chart, bar chart, waterfall chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. **Data preprocessing:**

The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data helps data scientist to get more precise results from an applied machine learning model. These techniques include data formatting, cleaning, and sampling.

1. **Dataset splitting:**

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets. Training set: A data scientist uses a training set to train a model and define its optimal parameters it must learn from data. Test set: A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model’s ability to identify patterns in new unseen data after having been trained over a training data. It’s crucial to use different subsets for training and testing to avoid model overfitting, which is the incapacity for generalization we mentioned above.

1. **Model training:**

After a data scientist has preprocessed the collected data is split it into train and test can proceed with a model training. This process entails “feeding” the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

**6. Model evaluation and testing:**

The goal of this step is to develop the simplest model able to formulate a target value fast and well enough to get accurate values. A data scientist can achieve this goal through model tuning. That’s the optimization of model parameters to achieve an algorithm’s best performance.

**Goal:**

Our team goal is to provide a hybrid model which increases accuracy than any other existing model, so we had implemented a hybrid model which combines Random Forest, SVM, Logistic Regression.

**Assessment:**

To fulfill the need the proposed system is to predict heart disease through an automated medical diagnosis system based on machine learning. For the prediction system, a hybrid model is used. For the heart disease prediction program, the Cleveland database is used. Since the Cleveland database is the ML researchers most widely used database.

The dataset includes 303 instances and 76 attributes, but all published studies reference only 14 of them. The ‘goal’ area of varying values from 0(absence) to 1 indicates whether the patient has heart disease. Since the Cleveland database studies centered on differentiating absence (value 0) from presence (value 1).

The proposed heart disease prediction system has been designed as a Hybrid Model. Patients can go for treatment based on the report generated helps patient to take preventive measures in advance.

**Data Set:**

The data collection process involves the selection of quality data for analysis. we used heart disease dataset taken from uci.edu site for machine learning implementation. The job of a data analyst is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques. The dataset collected with attributes age, sex, cp, treetops, Chol, fibs, resting, thalach, exang, oldpeak, slop, ca, thalach , reattribute.

**Methods and Models:**

The modules included in our implementation are as follows:

□ Logistic Regression

□ K-Nearest Neighbor (KNN)

□ Naïve bayes

□ Decision Tree

□ Support Vector Machine (SVM)

□ Random Forest

We used some advanced methods while executing models to overcome Overfitting, Data imbalance… those are cross validation and Hyper parameter tuning.

For Example:

For Logistic Regression -Backward and Forward Propagation

For K-Nearest Neighbor- Feature Selection and Cross-Validation

For Decision Tree- Cost Complexity Method

For Random Forest – Cross Validation and Hyper Parameter Tuning

**Feature Selection:**

Feature selection is the process of selecting a subset of the most important features (or variables) from a dataset to be used for building a model. In k-nearest neighbors (k-NN), feature selection refers to choosing the most relevant features to use in calculating distances between instances, which is a critical step in determining the nearest neighbors.

Choosing the right features can have a significant impact on the performance of a k-NN model. The selection process typically involves evaluating the relevance of each feature to the target variable and selecting the ones that contribute the most to the model's accuracy while avoiding overfitting. One common method for feature selection in k-NN is to use correlation coefficients or mutual information scores to assess the strength of the relationship between each feature and the target variable.

Once the relevant features are selected, they can be used to calculate the distances between instances, and the k-NN algorithm can be applied to classify new instances based on their nearest neighbors.

**Cost Complexity Method:**

Cost complexity pruning is a technique used in decision trees to prevent overfitting by reducing the size of the tree. The method works by penalizing the tree for having too many nodes or branches, thereby encouraging the tree to have fewer nodes and a simpler structure.

The cost complexity method involves adding a tuning parameter to the decision tree algorithm, which controls the trade-off between tree complexity and accuracy. The parameter, known as the complexity parameter or alpha, determines the strength of the penalty for having too many nodes in the tree.

The cost complexity pruning process involves building a series of decision trees with varying values of alpha, starting with a fully grown tree and progressively reducing the number of nodes by removing the least important ones. At each step, the tree's performance is evaluated using cross-validation or another validation technique, and the value of alpha that results in the best performance is chosen.

The final pruned tree is the one that results from using the chosen value of alpha to build a decision tree that balances accuracy and simplicity. This method can improve the performance of the decision tree model by reducing overfitting, making it more robust and less likely to generalize poorly to new data.

**Cross Validation:**

Cross-validation is a commonly used technique for evaluating the performance of machine learning models, including random forests. The goal of cross-validation is to estimate how well a model will generalize to new data that it has not seen before.

In random forest, one common form of cross-validation is called k-fold cross-validation. In k-fold cross-validation, the data is split into k roughly equal-sized folds or partitions. The random forest model is trained on k-1 of these folds and the remaining fold is used for testing the performance of the model. This process is repeated k times, with each fold being used for testing exactly once.

At the end of the k-fold cross-validation process, the results from each fold are combined to give an estimate of the model's performance on new, unseen data. One commonly used metric for evaluating the performance of random forest models is the mean squared error (MSE) or mean absolute error (MAE) of the predictions on the test set.

K-fold cross-validation can be repeated multiple times with different random splits of the data into folds to get a more reliable estimate of the model's performance. This technique is called repeated k-fold cross-validation.

Using cross-validation to evaluate the performance of a random forest model can help to prevent overfitting and provide a more realistic estimate of how well the model will perform on new data. It also allows for hyperparameter tuning, which involves adjusting the parameters of the random forest algorithm to optimize its performance.

**Hyperparameter Tuning:**

Hyperparameter tuning is an important step in optimizing the performance of a machine learning algorithm, including random forest. Hyperparameters are the configuration settings of the algorithm that cannot be learned from the data but need to be set before training, such as the number of trees in the forest, the maximum depth of each tree, and the minimum number of samples required to split a node.

Grid search: In grid search, a set of values for each hyperparameter is specified, and all possible combinations of these values are tried to find the combination that results in the best performance.

We implement a technique in this work that we call the Hybrid Model. The main objective of this research is to enhance the accuracy of heart disease prediction in terms of results. Many studies have been carried out that limit the selection of features for algorithmic use. The Hybrid process, by contrast, uses all features without any feature selection constraints. Here with a hybrid approach, we perform experiments used to classify the features of a machine learning algorithm. The findings of the experiment indicate that our proposed hybrid approach has a greater potential to predict heart disease relative to current techniques.

**Accuracies of Models:**

These are the accuracies of the all the models except our hybrid model:

1)Logistic Regression: 80.42%

2)K-Nearest Neighbor: 82.86%

3)Support Vector Machine: 85.66%

4)Naive Bayer’s: 81.12%

5)Decision Tree:83.22%

6)Random Forest:90.91%

**Hybrid Model:**

Ensemble learning is a powerful technique that can help improve the accuracy and stability of machine learning models. The basic idea behind ensemble learning is to combine the predictions of multiple models trained on the same data to produce a final prediction. There are different ways to combine the predictions, including majority voting, weighted voting, and stacking.

In this implementation, the ensemble consists of a voting classifier that uses the soft voting method. Soft voting is a weighted average of the predicted class probabilities, with the weights determined by the performance of each base model on the training data. This allows the voting classifier to take into account the confidence of each base model in its prediction, leading to a more accurate final prediction.

To train the base models, grid search with cross-validation is used to search for the best hyperparameters. Grid search is a technique that systematically searches a hyperparameter space to find the best combination of hyperparameters that leads to the highest performance on the validation data. Cross-validation is a technique that splits the data into multiple subsets and trains and evaluates the model on each subset, leading to a more robust estimate of performance.

The four base models used in this implementation - XGBoost, Random Forest, SVM, and Logistic Regression - are all different types of classifiers that use different algorithms and techniques to learn from the data. By using multiple types of classifiers, the hybrid model is able to capture different aspects of the data and reduce the risk of overfitting to a specific model or algorithm.

Finally, the performance of the hybrid model is evaluated on a test set using the accuracy score. The accuracy score is a simple metric that measures the proportion of correct predictions. While accuracy is not always the best metric to evaluate model performance, it is a commonly used metric that provides a quick and intuitive measure of how well the model is able to predict the correct class labels.

**Results:**

We developed a hybrid model for classification of our dataset using an ensemble approach, combining the predictions of several base models. The base models were XGBoost, Random Forest, Support Vector Machine, and Logistic Regression.

The hybrid model achieved an accuracy of 97.20% on the test set, outperforming all individual base models. The confusion matrix showed that the model performed well on all classes, with a high precision and recall for each class. Additionally, the model had a high F1-score, indicating a good balance between precision and recall.

We also evaluated the performance of each individual base model on the test set. XGBoost and Random Forest achieved the highest accuracy, followed by SVM and Logistic Regression. However, combining the predictions of all models using the voting classifier improved the accuracy significantly, demonstrating the effectiveness of the ensemble approach.

Overall, our hybrid model achieved excellent performance on the classification task, outperforming individual models and demonstrating the benefits of combining multiple models. The model could be used in a real-world application to classify similar datasets with high accuracy.

**Models Accuracies:**

Text, letter

Description automatically generated

**Hybrid Model Accuracy:**

Text

Description automatically generated with medium confidence

**Comparison Of Accuracies:**

Chart, bar chart

Description automatically generated

**Conclusion:**

The main objective of this research is to enhance the accuracy of heart disease prediction in terms of results. Many studies have been carried out that limit the selection of features for algorithmic use.

Here we implemented multiple models and shown their accuracy. We are worked on Random forest and Hybrid model based on random forest which produces more accuracy than any existing models.

We had used the models which will work both on large and small datasets because we also taken care about overfitting, underfitting and all the problems related to small datasets.

Comparison between few models also be done and shown with their accuracy. We came up with our hybrid model which produces more accuracy than any other model.

**Roles:**

Sai Karthik Nagabhairava, Ram Sai Repalle-Developed a hybrid Model, Working on Models and Accuracy, Comparison between models.

Bharath Kumar Pakam-Literature Survey, Data collection (data set collection), Data pre-processing, Data Splitting, Data visualization, Helping in model.

Deepshika Yedla-Research on past papers, Model Testing, Helping in model Accuracy.

**Bibliography:**

**1.) Reference notebook:**

<https://www.kaggle.com/code/andls555/heart-disease-prediction>

<https://github.com/MuntahaShams/Heart-Disease-Prediction-using-Neural-Networks>

**2.) Reference papers:**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8767405/#:~:text=The%20random%20forest%20algorithm%20was,the%20Kaggle%20dataset%20were%20considered>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6836338/>

**3.) Some related articles:**

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519796/>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6811630/>