Heart Disease Prediction

AVITI HARSHA VARDHAN REDDY
CSE, Amrita vishwa vidyapeetam,
Chennai, India
abhiharsha54@gmail.com

KARRI LIKHITH NAIDU
CSE, Amrita vishwa vidyapeetam,
Chennai, India
karrilikhith@gmail.com

THANNIRU BHANUPRAKASH
CSE, Amrita vishwa vidyapeetam,
Chennai, India
bbhanu0123@gmail.com

ADIREDDY SRINIVAS NAIDU
CSE, Amrita vishwa vidyapeetam,
Chennai, India
naidusrinvas525@gmail.com

Dr.E.Sophiya
CSE, Amrita vishwa vidyapeetam,
Chennai, India
e_sophiya@ch.amrita.edu

Abstract—**Heart problems can be predicted early. Researchers have developed model systems to assist cardiologists in improving the diagnosis process over time. However many current machine learning techniques are complex. Designed for utilising , with amounts of data. Sadly these methods are not effective when there is data for training the model. To address this issue this research proposes an efficient system that predicts heart infection using minimal records from specific datasets. This system combines Extra TreesClassifier (ETC) with a feature selection algorithm. Focuses on individuals who regularly go to the gym. Precise adjustment of hyperparameters is crucial for an implementation of any classifier. In this study min_samples_split which is considered the technique for hyperparameter optimization is used to improve ETCs hyperparameters. The interpretation of the proposed model is. Compared with classifiers to evaluate its effectiveness. The results indicate that the recommended model achieved 90.16% accuracy and an Area Under Curve (AUC) of 0.85. According to findings this model outperformed models in terms of accuracy. This research is highly relevant, for individuals who're concerned, about their well being and regularly visit gyms and fitness centers.**

I. Introduction

Nowadays around 72% of misfortunes can be attributed to heart infections, which are becoming increasingly prevalent, in both developing countries due to the lack of healthcare plans (SSPs) available, in certain regions. The current methods used to estimate heart disease are intricate and costly. However by utilizing machine learning algorithms we can simplify the process of gathering data ultimately saving lives. To reduce the number of deaths caused by heart failure it is essential to take an approach that focuses on researching contributing factors and improving prevention measures. In this article we have utilized selection methods to emphasize improvements, in estimating the level of heart contamination. Additionally we offer techniques, for tuning hyperparameters to improve the precision of models. Using our suggested example healthcare experts can evaluate a patients heart condition based on factors. Detecting complications on has the potential to reduce mortality rates by enabling intervention before further deterioration or delayed diagnoses occur.

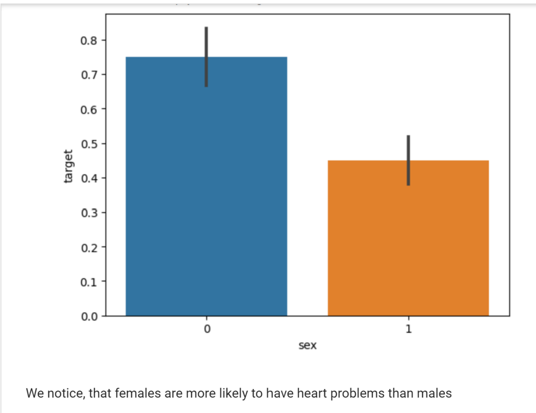
II. Literature survey

In the field of predicting heart failure, event processing (CEP) technology plays a role. This approach involves using predefined thresholds and rules to identify health parameters. To enhance prediction accuracy a study utilized three classifiers; decision tree, random forest and logistic regression. Additionally datasets, for predicting heart disease were employed. After going through results the researchers proposed a model that combines RF techniques, with linear models to improve accuracy. Another study examined the use of stacking techniques in diagnosing heart disease through the construction of level of two stacking based model that utilizes various classifiers. However it should be noted that training time can be considerable for datasets when employing stacked model architecture. In relation, to diagnosing heart disease, network (NN) models have been developed with models used for feature removal. We employed a algorithm to select relevant features. After that we used principal component analysis and a range of machine learning algorithms. To effectively and accurately predict heart disease we combined the models based on three performance indicators using a voting system. Additionally for results we incorporated an optional feature selection algorithm called AdaBoost with RFE along, with techniques such as LASSO rescue and local learning. The combination of Adaboost and support vector machines (SVM) resulted in a study focused on predicting patient survival, under conditions in which researchers performed eight models. different classifications.. Among them the additive tree classifier (ETC) performed well surpassing models in terms of performance. To address disparities in class representation they applied the Synthetic Minority Sampling Technique (SMOTE) . The study introduced a model that combines a Bayesian method with a genetic algorithm. They compared the results based on weighted rules. Found that regression models and hybrid models outperformed the other models. . While there has been progress in understanding heart disease prediction there are few gaps that are yet to be addressed. Future research should explore medicine approaches support populations and incorporate emerging technologies to enhance predictive accuracy . In machine learning competitions and real world scenarios, XGBoost (Extreme Gradient Boosting) is widely recognized as an algorithm of choice due, to its training time.

III. Dataset

In this study the researchers utilized a dataset called the UCI Cleaveland Dataset, which was created by Kaggle. This dataset consists of 303 entries, with 14 attributes that were collected from sources such as the UCI ML repository and the web. It includes features that're important in assessing the risk of heart disease .

- Age is a factor as the likelihood of heart disease generally increases with age
- Sex : Differences between males and females can influence the risk of heart disease.
- Hypertension (Blood Pressure) : High blood pressure is a contributor, to heart disease
- Cholesterol Levels : Elevated levels of cholesterol, low density lipoprotein (LDL) and low levels of high density lipoprotein (HDL) can contribute to an increased risk of heart disease
- Exercise Habits : Maintaining an exercise routine is linked to a decreased likelihood of developing heart disease. If you experience any symptoms such, as chest pain or shortness of breath it's important to take notice.
- The electrocardiogram (ECG or EKG) is a method used to measure the activity of the heart.



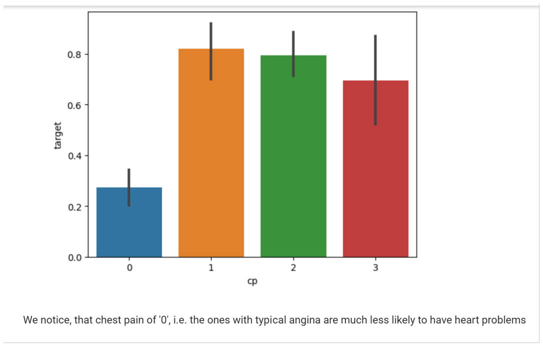
We notice, that females are more likely to have heart problems than males

Fig. 1. Barplot.

‘ ‘

A. Data preprocessing

- Step-1 Data Inspection : First lets inspect the data by examining the columns in the dataset. Next we can determine the shape of the dataset using a method to understand its dimensions. To gather insights we can count the values in the 'type' column. Lastly it's crucial to check for any missing values, within the dataset.
- Step-2 Data Cleaning and removing duplicate values : First we need to check if there are any missing values, in the dataset. If we find any we'll need to come up with a plan on how to handle them. we have an options either remove the rows with missing values fill the respective missing values using statistical measures, like mean, median or mode or use more advanced imputation techniques. The good news is that our dataset doesn't have any values.



We notice, that chest pain of '0', i.e. the ones with typical angina are much less likely to have heart problems

Fig. 2. Barplot.

- Step 3 Data Splitting : We split the dataset into a training set and testing set using the train_test_split() function. columns, from 0 to 13 are used as features, for both training and testing.
- Step 4 Machine Learning Models : To implement the model we utilized a Random Forest, Decision Trees, Extra Trees Classifier and Logistic Regression

.

B. Feature selection

- Recursive Feature Elimination (RFE) is an used feature selection technique, in machine learning. Its purpose is to enhance model performance and avoid overfitting by selecting the features. RFE operates by eliminating the significant features until the desired number of features is achieved. This approach proves beneficial when working with datasets containing features as it aids in identifying the subset of features that have the greatest impact, on the models predictive capabilities

Often these methods are employed prior, to inputting data into a model. The aim is to decrease the complexity and potentially enhance the models comprehensibility, training duration and overall performance in adapting to situations.

C. Methodology

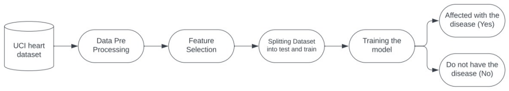

UCI heart dataset
Data Pre Processing
Feature Selection
Splitting Dataset into test and train
Training the model
Affected with the disease (Yes)
Do not have the disease (No)

Fig. 3. Flowchart.

The Extra Trees Classifier is a method of learning that creates decision trees during training and provides the most common class (for classification) or average prediction (for regression) from the individual trees. It shares similarities with Random Forests. There are some differences in how the treesre built.

Lets outline the steps involved in using the Extra Trees Classifier;

- Bootstrap Sampling to Random Forests Extra Trees builds each tree using a sample from the training data with replacement. This means that for each tree a subset of data is randomly selected for training.
- Random Feature Selection : of searching for the feature at each split in the decision tree Extra Trees randomly selects a subset of features. This randomization helps to reduce correlation between trees.
- Decision Tree Construction : Each decision tree is constructed using the sample and selected features. The tree continues to grow until it meets a predefined stopping criterion, such, as reaching depth or having a number of samples in a leaf node.
- Voting or Averaging In classification tasks all the trees predictions are combined by taking the mode (class) as the final prediction. For regression tasks an average prediction is calculated instead.

The probability $P(y|X)$ for classification tasks is often determined by the fraction of trees that predict a certain class.

While the Extra Trees algorithm involves randomization, the training process doesn't have a straightforward mathematical formulation with a set of equations like some other models. The essence of Extra Trees lies in the randomization of both data samples and features during the tree-building process.

Here's a high-level mathematical representation:

Derivations for Extra Trees specifically aren't commonly presented because the algorithm relies heavily on randomness and doesn't have a clear analytical solution. The primary focus is on empirical performance through experimentation.

Step-3; After constructing the decision trees we proceed to grow each tree by utilizing a sample and handpicked features. We continue expanding the tree until certain conditions are met, such as reaching a specified depth or having a number of samples in a leaf node.

Step-4; When it comes to classification tasks we determine the prediction by taking the mode (frequent class) of the predictions, from all the trees. For regression tasks we compute the prediction.

In classification tasks we often determine the probability $P(y|X)$ by looking at the fraction of trees that predict a class.

The Extra Trees algorithm incorporates randomization in its process. Unlike some models it doesn't rely on a specific set of equations, with a straightforward mathematical formulation. The key aspect of Extra Trees lies in the randomization of both data samples and features while building the trees.

D. Model training

- Logistic Regression is a technique employed to address classification problems wherein the result is a binary variable (0 or 1 True or False Yes or No). Its purpose is to estimate the likelihood that an instance belongs to a category. This approach has achieved an accuracy rate of 85.25%.

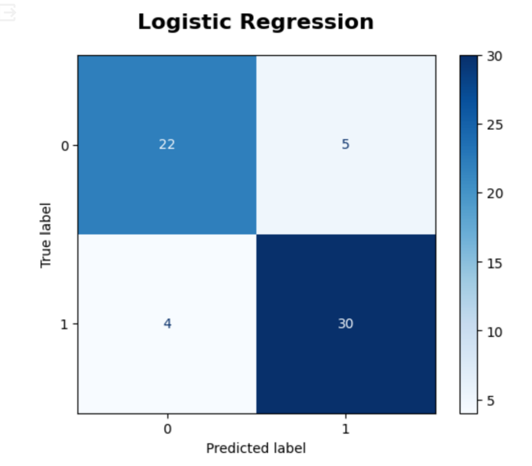


Fig. 4. Confusion matrix.

‘

- The Random Forest technique is an method that aims to prevent overfitting. It involves combining decision trees, which are constructed by evaluating features, with high information gain. In this case the Random Forest model was imported from the sci-kit library. By combining decision trees a random forest model is created the name ”ensemble method.” Hyperparameter tuning was performed using n_estimators and min_samples_split. The Random Forest model achieved an accuracy rate of 91.8% with a precision score of 91% and a recall rate of 94%.

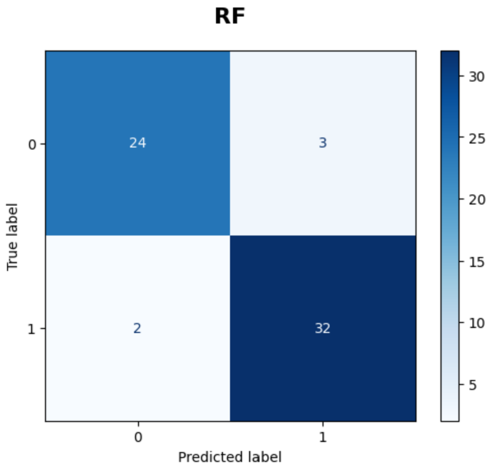


Fig. 5. Confusion matrix.

- The Extra Trees Classifier is a type of method that combines decision trees. Unlike decision trees that select the split based on a specific feature Extra Trees Classifier randomly selects features due, to its inherent randomness feature. This classifier is imported from the library. Allows for adjustment of randomness using the random state parameter. The accuracy achieved by the Extra Trees Classifier was 90.16% with precision and recall values of 0.92 and 0.85.

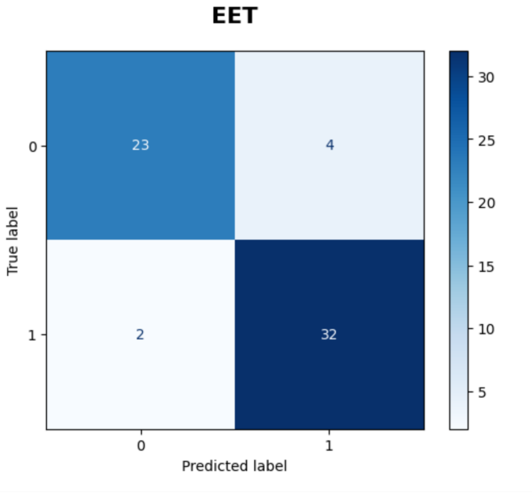


Fig. 6. Confusion matrix.

- Decision trees assess models by dividing the data using characteristics forming a hierarchical structure. The models performance is determined by how it can forecast outcomes on both the training and testing datasets. Additionally the tree structure offers insights, into the decision making procedure. Highlights the significance of various features. This specific model demonstrates an accuracy rate of 81.97%, with precision and recall values of 0.85 and 0.82 correspondingly.

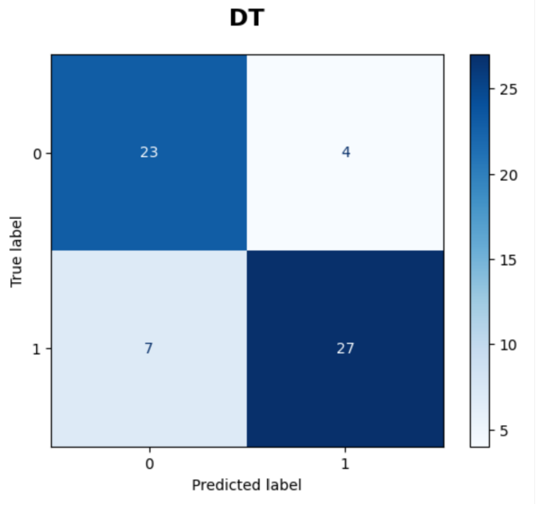


Fig. 7. Confusion matrix.

E. Results

Among all the algorithms random forest has achieved the accuracy rate of 91.8%. It is closely followed by ExtraTreesClassifier, NLP (networks) and XGboost. However random forest has managed to achieve a test accuracy of 100% by incorporating hyperparameters, like n_estimators and min_samples_split to prevent overfitting. As a result it has provided a test accuracy of 91.8% and an overall accuracy rate of 93.32%.

TABLE I
AFTER HYPERPARAMETER TUNING

S.NO	Performance metrics		
	Model	***Train ACC***	***Test ACC***
1	Random forest	93.39%	91.8%
2	Logistic regression	84.298%	85.246%
3	Naive bayes	83.47%	85.25%
4	KNN	72.31%	84.298%
5	MLP Classifier	86.89%	86.89%
6	XGBoost	85.95%	83.61%
7	ETC	90.91%	90.16%
8	Desicion tree	84.3%	81.97%
9	SVM	83.7%	81.97%

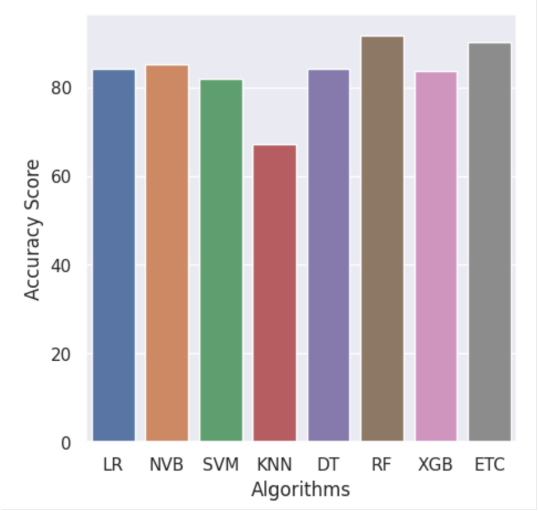


Fig. 8. Accuracy scores.

We decided to tune the hyperparameters of algorithms. Random forest is known to become overfit so we selected it for tuning. Despite having accuracy compared to forest we chose the ExtraTrees classifier as our recommended model because it fits best.

CONCLUSION AND FUTURE SCOPE

A new and effective model, for predicting heart disease has been introduced using the algorithm with six attributes. This model achieved an AUC of 0.89 and an accuracy rate of 90.16%. What sets this model apart from existing ones is its ability to demonstrate that complex machine learning algorithms or attribute selecting methods that are not necessary to achieve accuracy when working with small scale data. This proposed model has the potential to greatly assist Physician in identifying heart disease in patients at a stage enabling timely treatment. However it's important to note that creating a model is not sufficient; it's crucial to establish an ecosystem, for continuous collection of real time data in order to enhance the accuracy of the model. hence future research can explore leveraging cloud services and Application Programming Interfaces (APIs) for real time counter - deployment purposes.

References

[1] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," IEEE Access, vol. 8, pp. 133034–133050, 2020, doi: 10.1109/ACCESS.2020.3010511 .

[2] T. Vein and C. Guestrin, "XGBoost : A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939783.

[3] K. Harshini, P. K. Madhira, S. Chaitra, and G. P. Reddy, "Enhanced Demand Forecasting System For Food and Raw Materials Using Ensemble Learning," in 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), Gandhinagar, India, Sep. 2021, pp. 1–6. doi: 10.1109/AIMV53313.2021.9671005 .

[4] M.Tarawneh and O.Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," Acta Scientific Nutritional Health, vol. 3, no.7, pp.147-151, 2019 .

[5] Heart related problems and classification from AIIMS Mangalagiri.

[6] "A Stacking-Based Model for Non-Invasive Detection of Coronary Heart Disease," IEEE Access, vol. 8, pp. 37124–37133, 2020, doi: 10.1109/ACCESS.2020.2975377 .

[7] K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, vol. 19, p. 100330, 2020, doi: 10.1016/j.imu.2020.100330 .