Heart Disease Prediction Using Data Mining Techniques

Group No: 7

Adithya Raghu Ganesh
Department of Computer Science
NC State University
araghug@ncsu.edu

Meghana Ravindra Vasist
Department of Computer Science
NC State University
mravind@ncsu.edu

Karan Rakesh
Department of Computer Science
NC State University
krakesh@ncsu.edu

I. Introduction

Data Mining is the process of analysing information to build patterns, correlations and draw conclusions. There are a variety of different techniques to perform data mining, and they are used in machine learning, statistics and a variety of other fields. Parallely, the healthcare industry has been generating vast amounts of data for the past few decades and it was only fairly recently that scientists and medical professionals decided to leverage the power of data mining to enhance the understanding of diseases and cures. There have been major breakthroughs in the mitigation of diseases and even in genetic engineering through pattern matching and predictions.

Heart disease is one of the major causes of death in the past 15 years. It is said that heart attacks, strokes and other circulatory diseases account for almost 41% of deaths in the world.One person dies every 37 seconds in the United States from cardiovascular disease. Therefore it becomes extremely necessary to take precautionary measures to predict whether a person has heart disease or not before the condition gets too serious. Building prediction models to combat this disease have become a priority, so that we are in a position to actively prevent it rather than passively try to treat it. The death rates in the recent past is visualized in Fig 1.

Modern techniques [1] have proven to be effective in assisting to build robust models to predict and give doctors the insights required to combat heart disease. In this report, we have downloaded the Cleveland Heart Disease dataset from the UCI machine learning repository and have looked at the performance of various baseline techniques like linear regression and k-nearest neighbours that have been described in our curriculum and attempted at improving upon key metrics like recall and accuracy through a mixture of various novel techniques as well as some improvements to the existing techniques. We also attempt to generate insights from the results to help better interpret the results.

II. Background

Jaymin Patel et.al [3] implements variations of Decision trees called the J48 method, Logistic model tree algorithm and Random forest algorithm. In the J48 algorithm, the basic decision tree is constructed using the best splitting attribute. After extracting the decision rules from the decision tree, reduced error pruning is used to prune the obtained decision rules. Logistic model tree algorithm builds a classifier using decision trees as well. Once the complete decision tree is built, cost-complexity pruning is performed on the decision rules and logistic regression is applied on the leaf nodes in order to classify them.

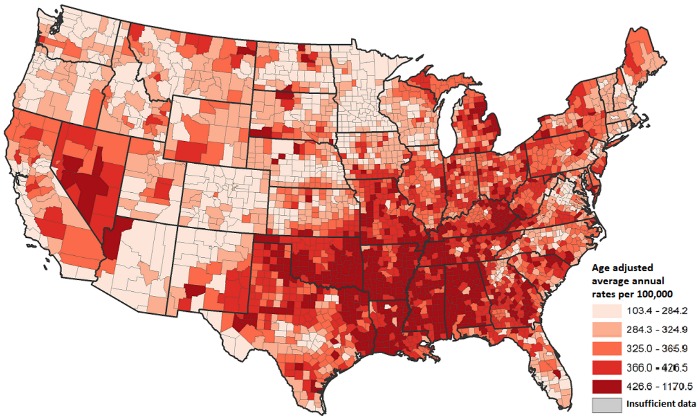

Fig. 1: Heart Disease Death Rates 2014-16 [2]

When the errors were calculated for these three algorithms it was seen that the Random forest algorithm was overfitting the data and hence, performed extremely well on the training dataset but didn't perform as well on the test data. The J48 algorithm performed worse than Random forest on the training set but performed much better on test data. Logistic model tree algorithm performed worse than both the algorithms on both training data as well as test data.

Senthilkumar Mohan et al [4] proposes a method which finds significant features by applying ML techniques to improve the accuracy of the predictive model. The proposed is a hybrid model which uses random forest with a linear model (HRFLM). HRFLM uses Vote in conjunction with a hybrid approach that uses Linear Regression and Naive Bayes algorithms. The Probabilistic Principal Component Analysis (PPCA) is used for feature selection and evaluation. PPCA is used to extract the vectors with high covariance and performs vector projection for minimizing feature dimension. The entire process of prediction starts with data preprocessing. Once that is done, the features are selected and are passed through a Decision tree with a Linear method combined with Random Forest and the predictions are obtained. This method gives an accuracy of 88.4% for the UCI Cleveland dataset. An improvement to this method is to use different feature selection techniques which can further help to increase the accuracy of the model.

To find the factors of the heart disease in the dataset, three association rule mining[5] techniques are used:
Apriori, predictive and tertius mining. Apriori and predictive rule mining are used to determine frequent itemsets from a set of transactions. Tertius mining is an inductive logic programming algorithm that considers confirmation evaluation function. This evaluation measure indicates the unexpectedness of a rule and in turn applies association rule mining. This paper also helped us derive intuition about the various features that were considered a higher priority in predicting the classes.

Methaila et al.[1] discusses using a combination of different algorithms and data mining techniques to perform prediction on the Heart Disease dataset. The algorithms used include Decision Trees, Neural Nets and Naive Bayes, along with weighted association apriori and the MAFIA algorithm. They note that the accuracy of the models improves on applying feature subset selection to obtain the optimal subset. They also find that regular apriori algorithms grow exponentially with respect to attributes and hence are not scalable, hence they use the MAFIA approach to reduce the number of subsets to be considered.

The problem with the methodology implemented in this paper is the focus on only increasing the accuracy of the dataset, whereas it is important to maximize recall in such a case like heart disease as the cost of a false positive is much higher than a false negative and accuracy does not take that into consideration. Hence we propose to create a network that also ensures that false negatives are considered while training the model.

Amita Malav et.al [6] describes a hybrid classifier that initially clusters the data using K means clustering to obtain the number of data groups and passes this information through an Artificial Neural Network to obtain higher accuracy than the individual models. This paper gave us an intuition about how to use hybrid classifiers and helped us create an ensemble architecture for this problem statement.

III. Software used

We used Python 3.7 on Jupyter Notebook to implement our project along with libraries. The libraries used include:

- pandas: To process the dataset
- matplotlib: To visualize the heatmap
- sklearn: To train the models, perform predictions and visualize the tree.
- seaborn: To generate the heatmap.
- itertools: To produce different combinations of ensembles

By implementing these new techniques we obtained a comprehensive understanding of some of the techniques used by data scientists which haven't been covered in our course.

IV. Method

In this section, we elaborate on the different methods that were implemented in past research papers [6] to simulate their accuracies and in turn improve upon the accuracy as well as the models used. After data discretization through Data Cleaning, Attribute selection and Normalization. Then, we implemented the following models.

- Preprocessing
  Data Preprocessing is the process of converting raw data into understandable format before being used.
- Data Cleaning
  Some of the data points may have human error or errors while taking the readings. Such anomalies are removed before various machine learning models can be fit on the data.
- Data Transformation
  Some of the features represented in the data might be structured differently . This makes the information represented to be falsely read by the model. Data transformation is the process of converting the data from one structure to another.
- Feature Selection
  Some of the features might hinder the performance of a model. Data reduction ensures only the essential features are used while fitting a machine learning model to ensure higher performance.

A. Approach

- K-Nearest Neighbour Algorithm:
  This algorithm assumes values belonging to the same class are clustered together. KNN captures the concept of similarity by considering a distance measure such as euclidean or manhattan to differentiate the classes.
- Linear SVC:
  The linear SVC is a classifier algorithm that returns a hyperplane that is most ideal to separate the data points into two separate classes. This algorithm uses a linear kernel. This model also uses the one vs rest strategy when dealing with multiple classes
- SGD Classifier:
  The Stochastic Gradient Descent is a classifier where the gradient of the loss that is estimated at each stage of training updates the parameters. This is also a linear model but it takes advantage of sparse data that is present. In such problem statements where high number of features are present, the SGD works in mini batches to compute results faster.
- Perceptron:
  A Perceptron is a simple Neural Network unit that uses the input features and weights them according to a threshold to determine the output class. This is primarily used for binary classification and if the dataset is large.

- Decision Tree:
  They are simple rule based classifiers that can handle both numeric and categorical data. When each feature gets considered, a set of if-then rules are applied on it at multiple levels before the decision tree classifies the given data points. It considers various metrics such as gini index and mean squared error for classification.
- Bagging Classifier:
  Such a classifier implements the concept of bagging. Here, a data set is split into multiple train sets and a base model is fit separately on all the sets of data. Once each of the models generate an output, the predictions get aggregated and a process of voting/ average is used to determine the output class.
- Gaussian Naive Bayes:
  The Gaussian Naive Bayes algorithm uses the concept of Maximum Likelihood estimation to modify the parameters of the probability distribution. This is implemented by maximizing the likelihood function which is assumed to be gaussian for this model. This uses the assumption of the presence of conditional independence amongst each pair of features and uses the Bayes theorem to determine the probabilities.
- Logistic Regression:
  Logistic Regression is a statistical technique that is used for binary classification. It provides a probability score for the observations and uses the logit function to determine the predictions from one or more combinations of the input variables.
- Random Forest Classifier:
  A Random Forest Classifier considers the average of the outputs generated from 'n' decision tree classifiers that were trained on several subsamples of the dataset to determine the prediction without any overfitting of data points.
- Extra Tree Classifier:
  In order to further prevent overfitting, the thresholds generated for splitting are randomly generated for the candidate features and of the generated thresholds, the most optimal feature is considered for splitting. Such an implementation reduces the variability and in turn increases the bias of the model.

B. Novelty

With the aim to efficiently predict whether a person has heart disease or not, we came up with the following novelty aspects in our problem definition:

- Modifying existing models in order to improve quality of predictions: After forming predictions from the baseline models, it was seen that the f1-score of the individual models was not very high. Hence, to improve this measure, we came up with an ensemble classifier which takes into account the best of these models and classifies the cases with the majority of the predictions from the individual classifiers.
- Interpreting the dataset efficiently to discover which of the features contribute more towards a positive prediction of heart disease: It is possible to visualize the dataset as a heat map which shows the correlation between the features and the target variable (prediction of whether the person suffers from heart disease or not). Based on these correlations, it is possible to determine which of the factors mainly affect the predictions.
- Finding patterns in the dataset: The data was modeled into a random forest classifier and the classification tree was visualized to get higher level interpretation of the classification model. Using this classification tree, it is possible to determine how a feature affects the prediction and hence, can assist in finding patterns in the dataset that lead to heart disease thereby aiding the medical personnel to detect the presence of the heart condition in the early stages and take necessary actions to cure it.

C. Rationale

There were various criterion that were considered for deciding the choice of models. The most critical factor was interpretability of the models since it was important to be able to draw high level conclusions from the data provided. The models chosen are a blend of faster predicting models along with more comprehensive models.

After examining the results of the models in the previous subsection, it was clear that the data had to be cleaned and transformed before fitting the data to the various models. In the dataset, we understood that features such as sex of an individual cannot be represented as ones and twos. This makes the model prioritize or weight the values differently. Thus, we incorporated data preprocessing techniques [7] such as data transformation and feature selection before feeding the data to the model. We then concluded that the following models were best suited for our problem statement:

- Decision Tree:
  This model is a single tree that is split up into various branches based on splitting criterion (GINI, entropy) and making predictions involves going down the particular branch of the tree until a leaf node is reached. It is hence a lazy learner ,i.e., training data is not required after the model has been trained and can cause significant overheads while predicting unseen data if the tree is deep. This model was chosen for its simplicity as well as its comprehensive nature i.e. it ensures the consideration of all paths and the resulting outcomes that occur and can be understood by doctors without any prior knowledge of machine learning techniques.
- Linear SVC:
  This model is a simple SVM that uses a linear kernel. It is advantageous as it is a lazy learner ,i.e., training data is not required after the model has been trained. Hence, this model is able to make quick predictions while also providing high performance. Since detection of heart disease is a time critical task, this classifier will be preferred since it makes predictions rapidly.

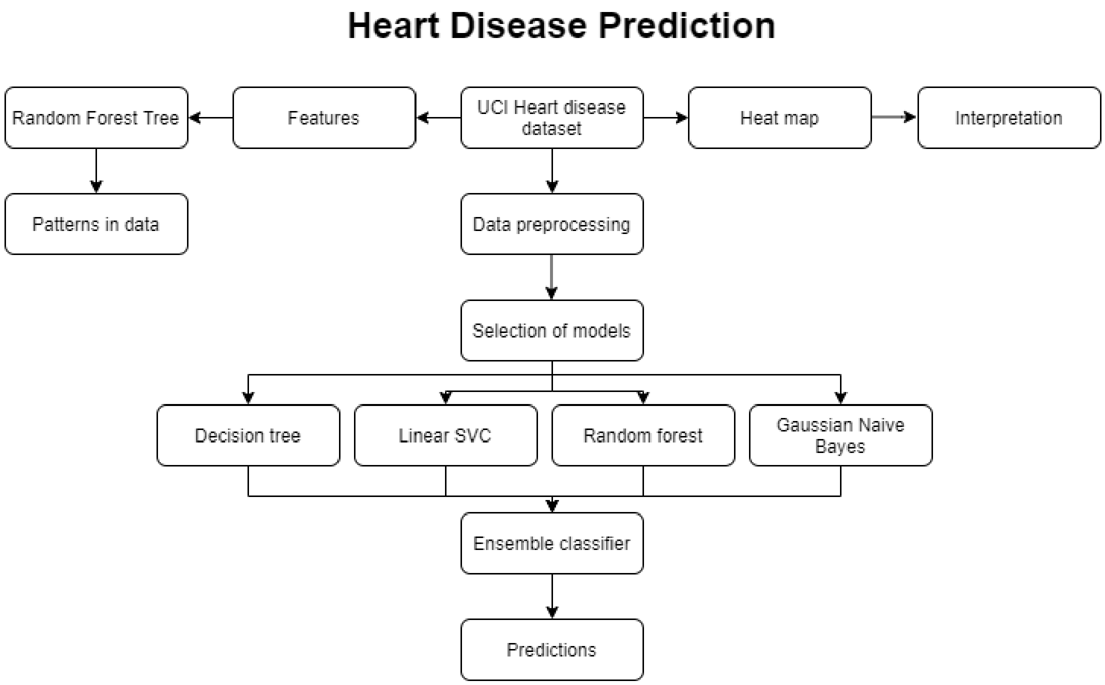

Fig. 2: Proposed architecture

- Random Forest:
  This model is a collection of various sub-trees that are collectively used to come to a particular consensus to make a prediction. Hence this model also falls unto the category of a lazy learner. This model was selected since it can be easily visualized and used to give valuable insights that help in answering more complex questions like the primary factors that contribute to the heart disease.
- Naive Bayes:
  This model is a probabilistic model used for classification. It uses the Bayes theorem to calculate conditional probabilities and takes these into account while classifying. Naive Bayes is very fast once the probabilities are calculated and these probabilities can be used without recalculation for the same training dataset. In this application, the training set does not change very frequently which makes this model suitable for prediction.

V. Experiments

A. Dataset description

The dataset named "Heart Disease Dataset" [8] was retrieved from the UCI Machine Learning repository. The data was created by a group of individuals from the Hungarian Institute of Cardiology at Budapest, University Hospitals in Zurich and Basel and the V.A Medical Center at Long Beach and Cleveland Clinic Foundation in 1988. The heart disease dataset consists of 76 attributes but most of the studies prioritize on the following 14 attributes.

- Age: Patient's age in years. (Numeric)
- Sex: Patient's gender. Male represented as 1 and female as 0. (Nominal)
- Cp: The type of chest pain - Typical angina, atypical angina, non-anginal pain, asymptotic. (Nominal)
- Trestbps: Level of blood pressure at resting mode in mm/Hg at the time of admission to the hospital. (Numeric)
- Chol: Serum cholesterol in mg/dl. (Numeric)
- FBS: Blood sugar levels on fasting $>$ 120 mg/dl; represented as 1 in case of true and 0 in case of false. (Nominal)
- Resting: Results of ECG while at rest are represented as - Nominal state as value 0, abnormality in ST-T wave as value 1, any probability or certainty of LV hypertrophy as value 2. (Nominal)
- Thali: The accomplishment of maximum heart rate. (Numeric)
- Exang: Angina induced by exercise. No depicted as 0 and yes depicted as 1. (Nominal)
- Oldpeak: Exercise-induced ST depression in comparison with the state of rest. (Numeric)
- Slope: ST segment measured in terms of slope during the peak exercise depicted as - 1 (unsloping), 2 (flat), 3 (downsloping). (Nominal)
- Ca: Fluoroscopy colored major vessels numbered from 0 to 3. (Numeric)
- Thal: Status of heart illustrated through three distinctly numbered values - Normal as 3, fixed effect as 6, reversible defect as 7. (Nominal)
- Num: Heart disease diagnosis represented in 5 values

with 0 indicating total absence and 1 to 4 representing the presence in different degrees. (Nominal)

B. Hypothesis

On studying the dataset and the prior literature on this topic, we formulated the following hypothesis that we aim to answer:

- Is accuracy the best metric to measure the performance of this problem? In all of the previous literature, the models were evaluated using accuracy solely. While accuracy can tell us how often a model makes a correct prediction, it does not measure the quality of the model with respect to the incorrect predictions. In medical applications, it is of paramount importance to maximise the number of correct predictions while also ensuring that we do not miss out classifying any positive cases incorrectly. These measures are encapsulated by precision and recall respectively. Hence, a measure that considers both these metrics is ideal, and F1-score is a measure that satisfies the requirement as it is the harmonic mean of precision and recall.
- Is it possible to deduce what factors contribute mainly in the diagnosis of heart disease? While most models can help diagnose the presence of heart disease, it is important figure out which factors can provide signs of impending heart conditions. This can help doctors take early preventive actions and avoid further complications that might arise.
- Is it possible to improve the performance on the dataset using the existing models? In order to achieve a better F1-score, we hypothesize that an ensemble model consisting of a combination of various simpler models can be used.

C. Experimental Design

1) Baseline: After completing the initial steps of data preprocessing and inference from the attributes, the models trained as follows :

- Logistic Regression, Naive Bayes classifier, Perceptron, Stochastic Gradient Descent Classifier, Ridge Regression Classifier,Bagging Classifier are trained with default attributes.
- Linear SVC is a Support Vector Machine with a linear Kernel where the dual parameter, which refers to the type of optimisation problem, is set to false since the number of samples is greater than the number of features.
- K-nearest neighbours algorithm is tuned using a grid search cross validation over the number of neighbours. 10 fold cross validation is used to compare the models. On finding the best parameters, the model is trained over the entire dataset.
- Decision Tree is implemented after tuning to find the optimal depth for the tree as well as setting a random state to control randomness and ensure reproducibility.
- Random Forest is implemented by conducting a grid search cross validation over the number of different sub trees to be considered, which is represented by the *n_estimators* parameter. Here, a 5 fold cross validation is performed to pick the best parameters before the model is trained.
- Extra Tree Classifier is trained with a fixed random state to ensure reproducibility.

2) Building a better model: :
The proposed architecture in Fig 2 has been implemented. The Data is procured from the UCI cleveland repository and intuitions are obtained by passing the features through a Random Forest visualizer and a Heat map. We interpreted the results from these visualizations to deduce patterns in the data and interpret the correlation between the features in the data respectively. Then, we preprocessed the data before splitting it into train and test. After training and testing the base models, we analysed the models side-by-side to get an understanding of the relative performance of the models. This also helped us compare the differences in using accuracy and F1 score as a metric for selection. The models are also measured on the speed of predictions since picking multiple lazy learners will significantly slow down the prediction process.

On completing the analysis, the selected models were combined into ensembles of varying sizes ranging from three to five. These ensembles were then run through a hyperparameter tuning to achieve the best boundary for the respective ensemble to classify positive and negative samples. These were put in a set and ordered to extract the ensembles with the greatest f1 score. The combination was then reverse engineered to find the classifiers that constitute them. There were then verified to ensure the correctness of the results. This was performed using the classification report functions provided in the scikit library [9].

A heatmap of the dataset was generated using the seaborn library in order to depict the correlation of the features with respect to one another. This enabled the interpretation of the dataset and helped us determine the extent to which feature is correlated to whether the person has a heart condition or not.

To find frequently occurring patterns in the dataset, random forest classification tree was used. This tree provided the visualization of the classification process. One of the decision trees of the random forest is shown in Fig 3. These trees were visualized using *export_graphviz* [10] module in the sklearn.tree library. Using this tree, patterns in the dataset were determined which enables the medical personnel to predict the heart disease faster and provide medication at the early stages.

VI. Results

The baseline models were trained and the results were tabulated as shown in the table 1. The results are sorted in ascending order of the F1-Score. It shows the model name, the accuracy of the model on the test dataset and the F1 score of the model. The last line of the table shows the performance of the ensemble model created. After taking into the consideration the results of the baseline models as well as the requirement for speed of predictions, a set of ensembles are assembled and tested to achieve the highest performance.

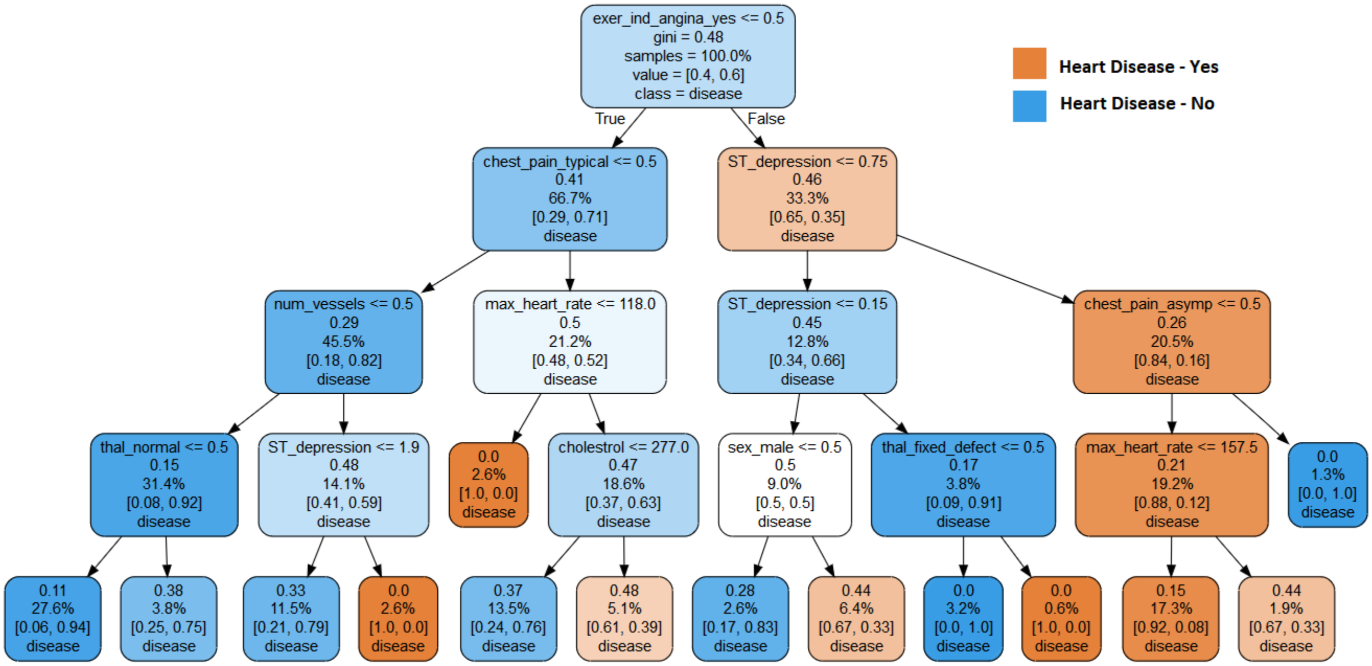

Fig. 3: Sample tree visualization for Random Forest

Model	**Accuracy**	**F1-score**
K Nearest Neighbor	0.64	0.67
Perceptron	0.56	0.72
Support Vector Classifier	0.69	0.76
Stochastic Gradient Descent	0.72	0.8
Bagging Classifier	0.8	0.81
Decision tree	0.8	0.82
Logistic Regression	0.85	0.87
Gaussian Naive Bayes	0.85	0.87
Linear SVC	0.85	0.87
Ridge Classifier	0.85	0.87
Extra Tree Classifier	0.87	0.88
Random Forest	0.89	0.9

Table 1: Baseline models with their accuracies and F1-scores

The models chosen were the Decision Tree model, Naive Bayes model, Random Forest model and the Linear Support Vector Classifier. On performing hyperparameter tuning on the ensemble for the best decision boundary, a boundary of 2.5 was obtained i.e. 3 or more positives constituted a positive prediction. More discussion about this result is presented in the following sections.

A. *Analysis*

Upon examination of the predictions made by the baseline models, it can be seen that the F1-scores for the models are not the same as the accuracies for the same models. As it can be observed from Table 1, K nearest neighbor gives a higher accuracy than Perceptron model, but when the F1-scores are examined, the Perceptron model performs better than the K nearest neighbor. This validates our hypothesis that accuracy alone is not a strong metric to measure the quality of the model.

For the proposed classifier, the models used in the ensemble are shown in the table below.

Model	**Accuracy**	**F1-score**
J48 [3]	0.84	-
HRFLM [4]	0.88	0.9
Augmented Naive Bayes [11]	0.80	-
Proposed Classifier	**0.92**	**0.93**

Table 2

These models provided the ideal combination of strong performance as shown in the table 1 as well as providing interpretable results as well as quick predictions on the data as there are only two lazy learners. The boundary chosen for the classifier ensures that the model only makes positive predictions when it has an absolute majority i.e. atleast 3 positive votes. This reduces the number of misclassified positive instances. Since it is not an extreme boundary, it doesn't unnecessarily classify correct instances as false, thus not affecting recall adversely.

Table 2 compares the accuracies of previous state of the art architectures with our proposed ensemble classifier and it can be seen that our model performs better than the previous models. In order to determine which factors contribute mainly to the prediction of the disease, a heatmap as shown in Fig 4. The values in the heatmap indicate the magnitude of correlation between the features. A value of 1 and -1 indicates high positive correlation and high negative correlation respectively. Positive values indicate positive correlation and negative

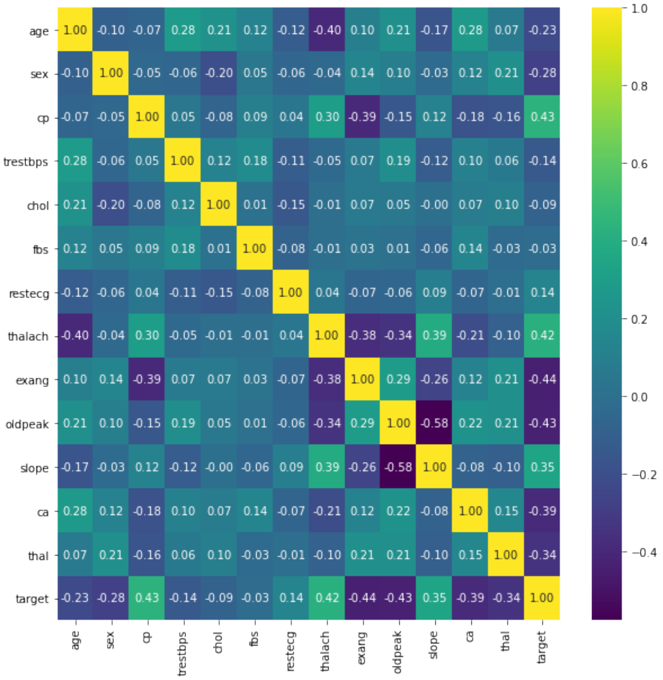

Fig. 4: Heatmap

values indicate negative correlation. In order to determine the factors that are highly correlated with the target value, the last column or the last row is checked. In this case, exang i.e; exercise induced angina has the highest correlation with the target value with a value of -0.44. This is interpreted as, exang is negatively correlated to the prediction, i.e; if exang has a high value, then the possibility of the person having heart disease is low.

Fig 3, each node contains data about the split in the following order:

- First row: Indicates the attribute with the value of the attribute along with the inequality which jointly form the condition based on which the attribute is split.
- Second row: Indicates the criterion chosen to determine the best splitting attribute. In this case, the criterion chosen is Gini index.
- Third row: Indicates the number of samples out of the whole number of samples on which the condition is applied.
- Fourth row: Indicates the probabilities of disease and no disease in the samples considered under the node.
- Fifth row: Indicates the class for which the node is constructed. In this case, all the nodes are constructed for the class ”disease” where each condition determines the split for whether the attribute contributes to heart disease or no heart disease.

In the tree, all the samples that satisfy the condition move to the left of the node and all the samples that do not satisfy the samples move to the right of the node. One of the patterns that can be extracted from the sample random forest classification tree in Fig 3 is: If the person does not have exercise induced angina less than 0.5, has the ST depression on the ECG plot greater than 0.15 and lesser than 0.75 and has fixed defect thalassemia less than 0.5, then 0.91 portion of the samples that come under this node do not have heart disease. This portion contributes to 3.2% of the total data samples passed through the node. Note that this portion is very small since the image depicts only a part of the random forest classifier.

VII. Conclusion

From the analysis, it can be concluded that accuracy is not a sufficient metric to evaluate model performance, and a more robust metric like F1 Score must be used. It is also evident that the techniques used to generate the proposed classifier have resulted in an improvement of the F1-Score from 0.9 (maximum from the baseline models) to 0.93. Meaningful interpretations were also derived with regard to the dataset apart from predictions.This also helps in understanding the causes for the problem to ensure that early measures are taken to prevent it from occurring at all.

Future work includes using more sophisticated feature engineering techniques, developing a custom metric catered to heart disease prediction and researching advanced techniques to use for the ensembles to achieve even higher performance.

A. *Roles of Team Members:*

Karan Rakesh:

- Implementation: Logistic Regression, K Nearest Neighbors, Gaussian Naive Bayes, Ensemble Classifier
- Report: Introduction, Rationale, Experimental Design, Results, Conclusion

Adithya Raghu Ganesh:

- Implementation: Bagging Classifier, Perceptron, Ensemble Classifier, Visualizations
- Report: Approach, Rationale, Dataset Description, Figures, Tables

Meghana Ravindra Vasist:

- Implementation: Extra Tree Classifier, Decision Tree, Random Forest Classifier, Ensemble Classifier
- Report: Background, Rationale, Novelty, Hypothesis, Analysis

Code can be found at: https://github.com/adithyarganesh/CSC522_P07_Heart_Disease

VIII. Virtual Meetings

We had virtual meetings on Zoom on the following dates at the following times:
04/18/2020 - 12:00PM to 2:00PM
04/19/2020 - 10:00AM to 12:00PM
04/21/2020 - 11:00AM to 12:30PM
04/22/2020 - 3:00PM to 10:00PM
04/23/2020 - 11:30AM to 6:00PM
04/24/2020 - 11:00PM to 6:00PM

REFERENCES

[1] Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, et al. Early heart disease prediction using data mining techniques. *Computer Science & Information Technology Journal*, pages 53–59, 2014.

[2] https://www.cdc.gov/heartdisease/facts.htm.

[3] Jaymin Patel, Samir Tejpalupadhyay, and Samir Patel. Heart disease prediction using machine learning and data mining technique. 03 2016.

[4] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554, 2019.

[5] Jesmin Nahar, Tasadduq Imam, Kevin S Tickle, and Yi-Ping Phoebe Chen. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Systems with Applications*, 40(4):1086–1093, 2013.

[6] Amita Malav, Kalyani Kadam, and Pooja Kamat. Prediction of heart disease using k-means and artificial neural network as hybrid approach to improve accuracy. *International Journal of Engineering and Technology*, 9(4):3081–3085, 2017.

[7] S Palaniappan, TV Rajinikanth, and A Govardhan. Spatial data analysis using various tree classifiers ensembled with adaboost approach. In *Emerging Trends in Electrical, Communications and Information Technologies*, pages 165–174. Springer, 2017.

[8] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[10] Braxton Fitts, Ziran Zhang, Massoud Maher, and Barry Demchak. dotapp: a graphviz-cytoscape conversion plug-in. *F1000Research*, 5, 2016.

[11] K Srinivas, B Kavihta Rani, and A Govrdhan. Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, 2(02):250–255, 2010.