Defining The Best-Fit Machine Learning Classifier Prediction Model For Diagnosis of Heart Disease

Debarati Dey Roy (debaratidey24@gmail.com)
B. P. Poddar Institute of Management & Technology
Debashis De
Maulana Abul Kalam Azad University of Technology

Research Article

Keywords: Heart disease, Machine learning, Classification learner, Confusion matrix.

Posted Date: December 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1152876/v1

Additional Declarations: No competing interests reported.

Abstract

Cardio vascular disease or alternatively heart disease is the primitive cause of death all around the world. Last few decades, it was observed that maximum death cases occurred due to heart failure. The heart failure death cases are associated with many risk factors for example high blood pressure, cholesterol level, sugar level etcetera. Therefore, it is advisable that regular and early diagnosis of these factors may reduce the risk of heart failure and hence achieve prompt disease management service. A commonly used technique to process these enormous medical data is called data mining, which help the researchers in health care domain. Several machine learning algorithms are used to analyses these data and help to design the best-fit model for early detection of heart diseases. This research paper contributes various attributes related to heart mal functioning and build the best-fit model using supervised learning algorithm such as various tree (fine tree, medium tree etc), Gaussian Naïve Bayes, Coarse KNN, Medium Gaussian SVM algorithms. In this paper, we used the data set from Kaggle.com. These data set comprises with total 732 instances along with 5 attributes. All these 5 attributes are to be considered for testing purpose and also to find out the best fit model for prediction of heart disease. In this research article we also compare the various classification models based on supervised learning algorithms. Based on the performance and accuracy rate we therefore, choose 'Medium Tree' model as the best-fit model. Maximum accuracy is obtained for 'Medium Tree' model. The confusion matrix for each model are calculated and analyze.

Introduction

Heart disease is the one of the leading cause of death in the world for the last few decades. One of the published report of WHO, it was mentioned that over 17.9 million deaths occurred due to heart malfunctioning every year [1, 2]. However, there are several professional and genetic, habits are responsible for heart disease but still some health factors are associated for cardio vascular malfunctioning. For example high blood pressure, sugar level, cholesterol level, maximum heart rate also important factors for the early detection of heart disease. Data mining is an essential area for medical diagnosis. For the effective early detection of heart disease, a best-fit model is highly appreciated using supervised machine learning algorithms. The accuracy rate, validation rate are important factors to choose the best-fit model. To get an efficient model, it is important to have huge and authentic data set. If the number of attributes are increased then the efficiency of such models are also increased. To assist heath care professionals, data mining and machine learning tools are the two most significant fields of research. The machine learning algorithms are powerful tools to build an efficient best-fit model for early detection of heart diseases. Data mining helps to extract the required information from a large datasets. Feature extraction plays an important role for this stage of model building domain. Machine learning is generally categorized in four different categories namely, supervised learning, Semi-supervised, un-supervised learning and reinforcement learning. This supervised learning helps the machine to diagnosis properly from its previous experience and act accordingly. In artificial intelligence domain, machine learning is one of the most challenging and growing tool for detection of predicting models. Machine learning algorithms are useful to deal with large amount of data set. It is useful for routine prediction modelling approach where a computer gains and understands from its previous experiences. Building up of high accuracy levels for this prediction models depend on the previous experiences. More the experiences, more accuracy level. Machine learning algorithms are exploring large datasets to extract required information to remove errors from the prediction models and factual outcomes [2, 3]. These prediction models help to analyze future decisions accurately. Various supervised machine-learning algorithms are Decision Trees, Naïve Bayes, random forest, K-nearest neighbor etc. Among these various algorithmic models, a comparative analysis has introduced. The data sets that we have used in this paper collected from Kaggle.com. This research paper demonstrates the followings:

- Most efficient prediction model.

- Comparative analysis among the most significant prediction models.
- Calculation of confusion matrix and its analysis.
- Calculation of precision, prevalence, null error rate, sensitivity of the best-fit model.

Background

Cardio vascular illness affect billions of people every year worldwide. Therefore, proper medical diagnosis should be efficient, effective, reliable and accurate with the help of data mining and machine learning algorithm tools. Data mining is a software technology by means of which computers are able to shape and categorize various attributes correctly. This research article use classification-learning method to find out best-suited prediction models to predict heart disease with high accuracy. In fig.1, the various steps for supervised machine learning is shown using block diagram. Among these steps data preprocessing and data cleaning are the two most important processes, which are associated with supervised machine learning prediction model preparation. In this research article, we have used the data set from Kaggle.com website. It comprises a real data set of 732 instances with 5 attributes for example cholesterol level, blood pressure, sugar level etc. We have used four classification algorithms to create a prediction model with maximum possible accuracy. The real life data contains some missing and noisy data. Before using this data set, we need to remove those noisy data and missing data, which is known as data pre- processing. Figure 1 explains the sequential steps for the various parts of our proposed model. Cleaning is the process by which noise removal is being completed. Noisy data should be removed and the missing values should be fulfilled before these data are applied for processing. ***Transformation*** is useful for changing the data format in a comprehensive manner. Smoothing, normalization and aggregation are the parts of this process. ***Integration*** is the process to associate various data sets from different source of data. ***Reduction*** is useful for complex data and it is required to be formatted to achieve results that are more accurate. Therefore, the data are categorized and isolate into two parts, namely training data set and test data set to build a best-suited supervised prediction model.

Results And Discussion

In this paper, we propose a best-fit model for the early detection of heart disease using five attributes. The data collected from 'Kaggle.com'. We propose the model using classification-learning algorithm. We tested for various Tree algorithms. Among these algorithms, the best accuracy model has selected for the nomination of best-fit model for early detection of heart disease. Cross-validation is also to be considered for the modeling purpose and by modulating the cross-validation value, the validation accuracy has also been tested for the chosen best-fit algorithmic model. Cross-validation is essential whenever a best –fit model is need to choose among the competitive models. This process is also known as re-sampling of data or rotation estimation or out-of-sample testing, which removes the repetitive errors, which might happened during the consideration of training data set. This process is a resampling method that uses diverse slices of the data to test and train a model on dissimilar iterations. In this settings, where the goal is prediction, and accuracy then this cross-validation models play an important role.

Table 1: Comparison table of various machine-learning algorithms proposed for this problem analysis:

Machine Learning algorithms	Accuracy (Validation)	Prediction speed	Total cost(Validation)	Maximum number of splits	Training time
Decision Tree (cross-validation=5)					
Fine Tree	72.8%	~5400 obs/sec	201	100	13.901 sec
Medium Tree	93.7%	~38000 obs/sec	146	20	1.3995 sec
Coarse Tree	79.0%	~60000 obs/sec	154	4	0.55792 sec
Optimizable Tree	79.0%	~23000 obs/sec	156	2	72.875 sec
Decision Tree (cross-validation=6)					
Fine Tree	74.0%	~15000 obs/sec	190	100	7.9149 sec
Medium Tree	78.6%	~21000 obs/sec	157	20	2.189 sec
Coarse Tree	79%	~9400 obs/sec	154	4	2.0803 se
Optimizable Tree	79.5%	~9300 obs/sec	150	13	65.883 sec
Decision Tree (cross-validation=7)					
Fine Tree	73.9%	~20000 obs/sec	191	100	6.5226 sec
Medium Tree	77.9%	~19000 obs/sec	162	20	2.3796 sec
Coarse Tree	79%	~19000 obs/sec	154	4	2.2469 sec
Optimizable Tree	79.6%	~11000 obs/sec	149	6	46.378 sec
Decision Tree (cross-validation=8)					
Fine Tree	72.7%	~20000 obs/sec	200	100	2.439 sec
Medium Tree	80.1%	~19000 obs/sec	154	20	2.2201 sec
Coarse Tree	79.2%	~17000 obs/sec	152	4	2.5031 sec
Optimizable Tree	80.7%	~2700 obs/sec	141	15	37.555 sec

Table 1 shows that various machine –learning algorithms are proposed for this specific problem to diagnosis early detection of heart problems. In this best-fit model prediction, various decision tree algorithms are considered. Namely, fine tree, medium tree, coarse tree, optimizable tree algorithms are considered for modelling. Different parameters are

chosen for comparative study among them. High accuracy rate helps to choose more accurate best-fit model. For different numbers of cross-validation, we get higher range of accuracy mostly for medium tree algorithmic approach. Therefore, this medium-tree model has suggested as the best-fit model for the predictive model of early stage detection of heart disease. Decision tree is a classification learner algorithm that works on definite and mathematical data. It provides tree-like structures to handle huge medical database. However, there are many classification algorithms are available, but due to the simplicity of this decision tree it attracts researchers. There are three main nodes are available for decision tree; root node: main node on which all other nodes depend; interior node: which handles various characteristics and finally leaf node: that represents the outcome of each test.

Table 2: Comparison of performance analysis of various decision tree algorithms:

Machine Learning algorithms	Accuracy	Precession	Misclassification Rate/ Error Rate	True Positive Rate/ Sensitivity" or "Recall	False Positive Rate:	True Negative Rate	Prevalence
Fine Tree	72.5%	0.759	0.274	0.764	0.327	0.672	0.575
Medium Tree	93.71%	0.826	0.1994	0.826	0.234	0.765	0.575
Coarse Tree	78.27%	0.824	0.217	0.79	0.22	0.771	0.575
Optimizable Tree	78.6%	0.808	0.213	0.824	0.263	0.736	0.575

Table 2, shows the different mathematical parameters for these algorithms. In this table, various mathematical parameters have calculated. Accuracy mention that how often classifier is correct. Higher the value of accuracy, prediction is more accurate. Precision is an important parameter for this statistical classifier which when predicts ‘yes’, then it also predicts how often it is correct. Misclassification rate determines how often this prediction is wrong. It is also known as “Error rate”. True positive rate is the parameter which is determined by the ratio of ‘true positive values’ and actual ‘yes’. This is also known as ‘sensitivity’ or ‘recall’. False positive rate is calculated by false positive value by actual number. True negative rate true negative number by actual number. True negative rate is also known as “Specificity”. From this table , it is seen that medium tree has highest accuracy among all. Therefore, medium tree is selected for the best-fit model.

ROC (receiver operating characteristic curve) for this best-fit model is analyzed in Fig. 2. This graph is evaluated the performance of a classifier model at all classification thresholds. It is generated by plotting the True Positive Rate (TPR, y-axis) against the False Positive Rate (FPR, x-axis) as the threshold varies for transmission of observations to a given class. The optimum threshold for both TPR and FPR are 0.77 and 0.17 respectively. Here the area under this curve is accommodate 82% area under positive class. TPR 0.77 indicates that the current classifier (medium tree) assigns 77% of the observations correctly for the positive class. However, FPR of 0.17 indicates that the 17% observations are incorrect for the positive class. For ROC curve in Fig.2, if the shape of the graph is a proper right angle, then there is no misclassified points and it is said to be perfect. If the curve holds 45° angle then poor result is obtained for the classifier. The Area Under Curve (AUC) measures the overall quality of the classifier. Larger AUC indicates better performance of the classifier. ROC curve compares the performances of the classes and the trained models. In these predictive models, we obtained 82% and 80% areas for the trained and test models respectively. These results shows the high performance rate of the classifier.

The scatter plot of this medium tree classifier, examine the classifier results. After train the classifier, scatter plot displays the data as a predicting model. Using cross-validation, these prediction models represent the various class against various attributes. In Fig.3 (a-e), these prediction models show the four classes representation against five

attributes. The decision tree algorithm divide the data into two or more equivalents sets based upon the most crucial attribute. The entropy of each indicator is therefore, calculated with predictions. The predictors having maximum information with minimum entropy. The most desirable thing for this algorithm is that the results can be obtained easily and above all these results are interpretable. However, only one attribute is tested at a time and the data may be over classified [2-5].

Finally, the comparative table 3 shows that the various existing prediction models versus the proposed prediction model and its novelty.

Table 3: Comparison table of the existing best-fit supervised model and the proposed best-fit model for early heart disease detection.

References	Learning algorithms	Accuracy	Best Technique Found	Dataset	Precession	Sensitivity	Tool
Shah et al [2]	Naïve Bayes K-NN Decision tree Random forest	88.157% 90.789% 80.263% 86.84%	KNN	UCI Machine learning repository	Not mentioned	Not mentioned	WEKA
Parthiban et al. [6]	Naïve Bayes	74%	Naïve Bayes	Not mentioned	97.52%	Not mentioned	WEKA
Vembandasamy et al. [7]	Naïve Bayes	86.419%	Naïve Bayes	A diabetic research institute in Chennai	Not mentioned	Not mentioned	WEKA
Kumar Dwivedi [8]	Naïve Bayes Classification tree K-NN Logistic regression SVM ANN	83% 77% 80% 85% 82% 84%	SVM	UCI Machine learning repository	Not mentioned	89%	Not mentioned
Proposed model	Fine tree Medium tree Coarse tree Optimizable tree	72.5% 93.71% 78.27% 78.6%	Medium tree	Kaggle	75.9% 82.6% 82.4% 80.8%	76.4% 82.6% 79% 82.4%	MATLAB

Conclusion

The aim of this paper is to making decision using best-fit model for early cardio-vascular decease. Efficient and effective results have obtained using supervised classification learning algorithms. The decision tree learner provides high accuracy rate along with good efficiency for test data set. Among the four decision tree algorithms, medium tree provides high accuracy rate, i.e.93.7 %. However other tree algorithms also provides high accuracy rate. The comparison table shows that medium tree fulfills the entire criterion to being a best-fit model for the early level cardio-vascular disease detection model.

Declarations

Availability of data and materials

All the data and material are available in the manuscript.

Competing interests

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding

Not applicable.

Authors' Contributions

All the authors have equally contribution to prepare the manuscript.

Acknowledgements

Not applicable.

Ethics

On behalf of all authors, the corresponding author states that there is no conflict of Ethics

Consent to participate

All the authors have consent

Consent for publication

All the authors have consent for publication

References

1. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol.2011; 3:67.
2. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), 1-6.
3. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

4. Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015,2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.
5. Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p.p.482–86.
6. Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. Int J Appl Inf Syst, (IJAIS). 2012;3(7):25–30.
7. Vembandasamy K, Sasipriya R, Deepa E. Heart diseases detection using Naive Bayes algorithm. Int J Innov Sci Eng Technol. 2015;2(9):441–4.
8. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29(10):685–693

Figures

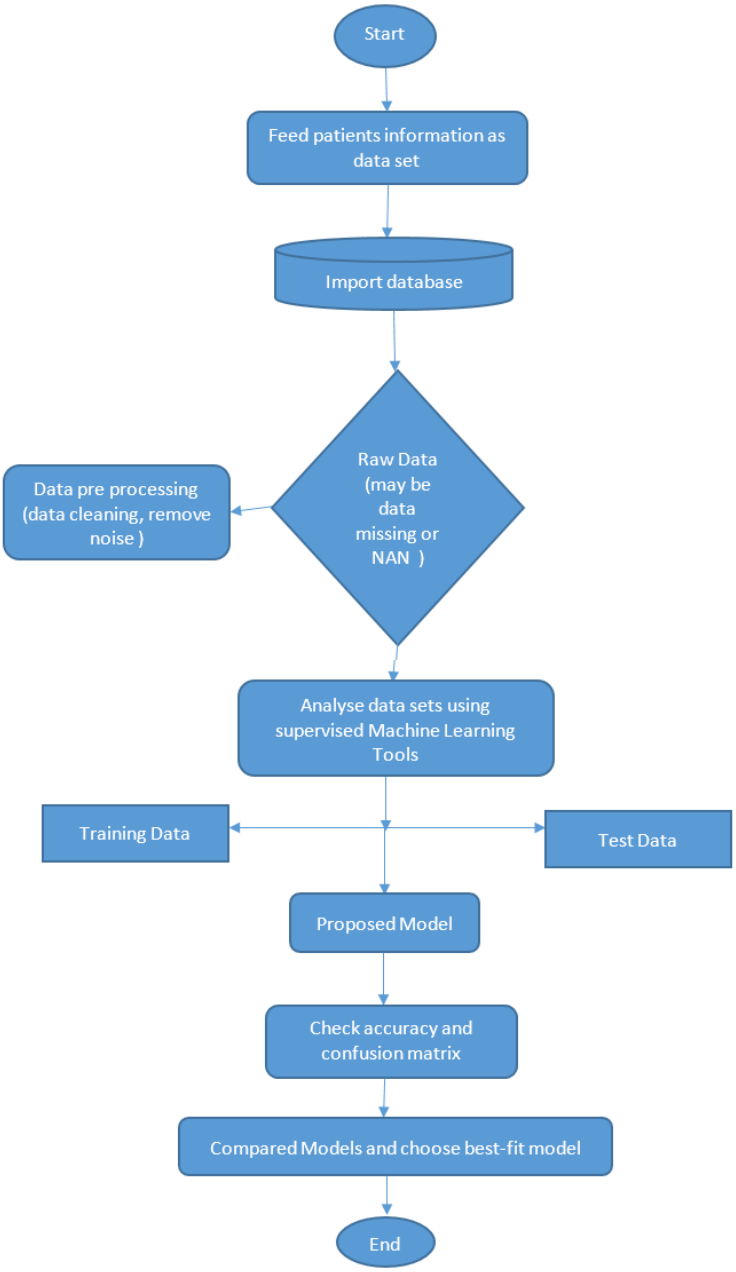

Figure 1

Step-by-step block diagram of proposed machine learning model

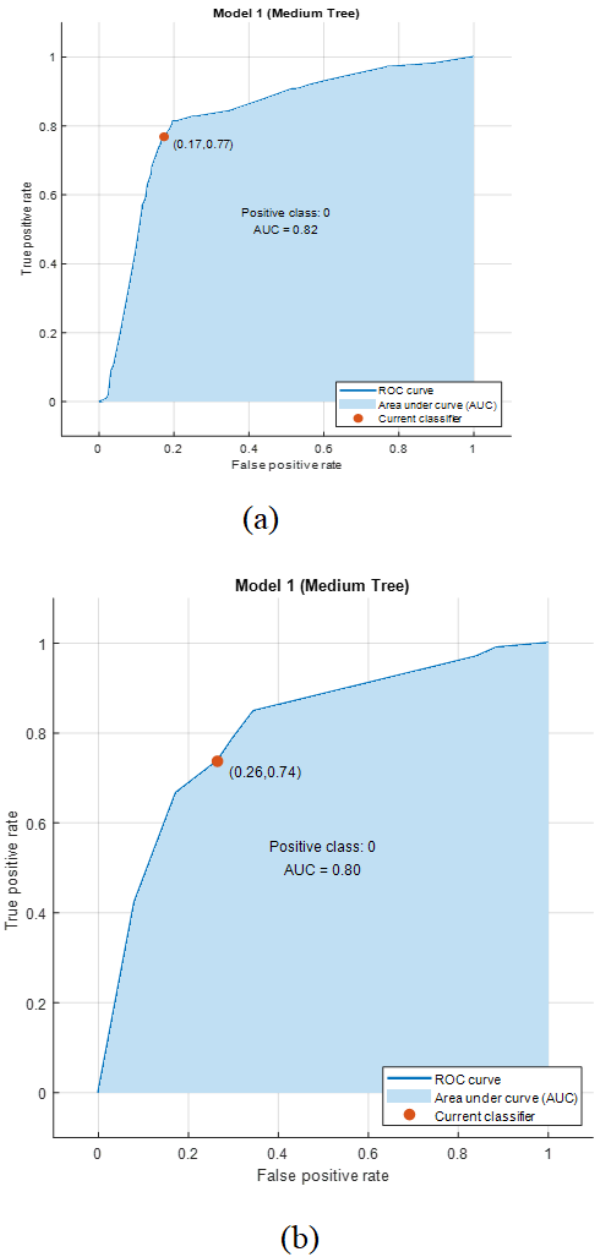

Figure 2

ROC graph of best-fit model of proposed machine learning model. a. Training data b. Test data

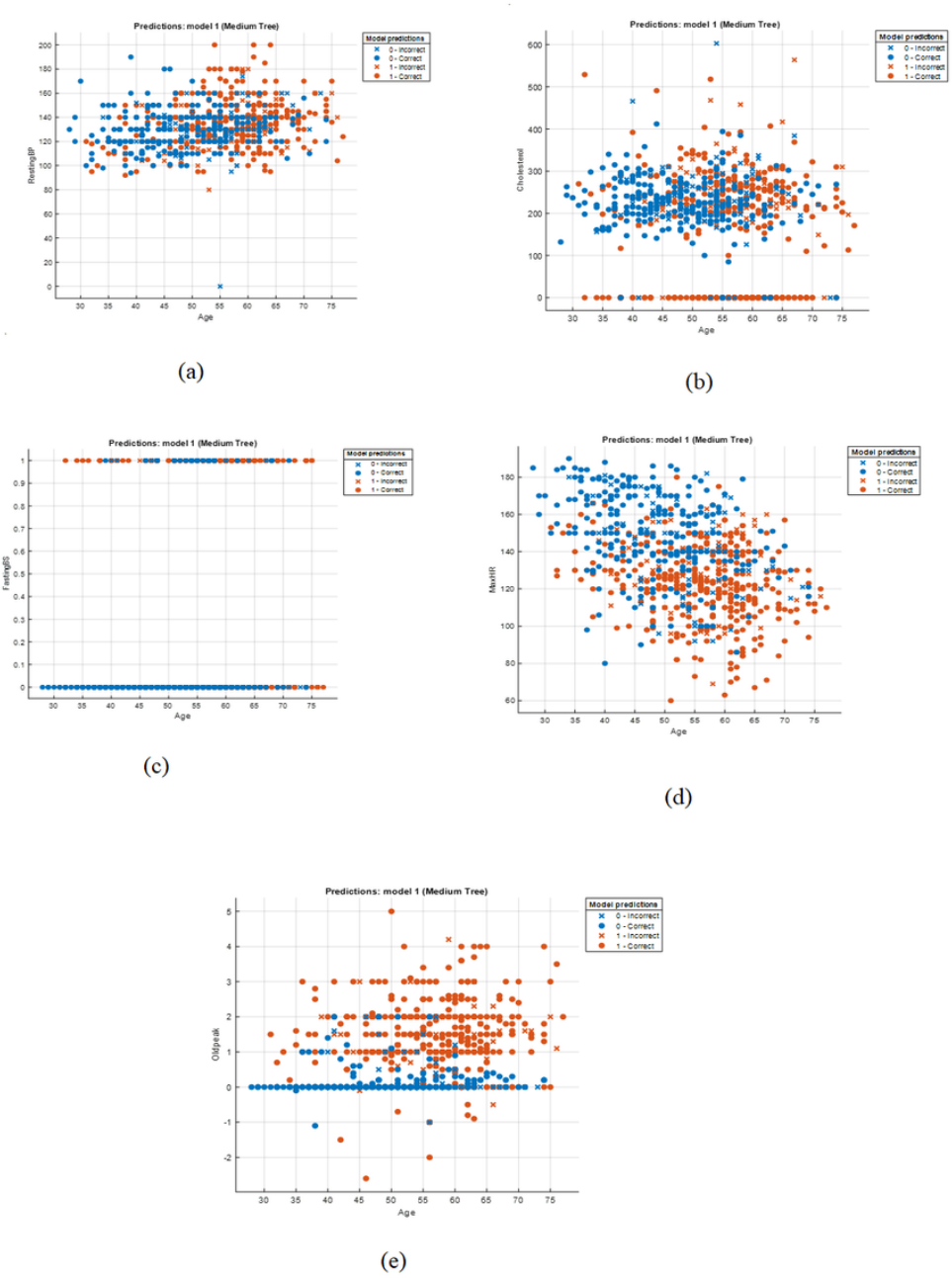

Figure 3

Best-fit prediction models for different classes. a. Resting B.P. Vs. Age b. Cholesterol Vs. Age c. Fasting sugar Vs. Age d. Maximum heart rate Vs. Age e. Old peak Vs. Age