



A STUDY ON PREDICTING HEART DISEASE WITH ENSEMBLE MODELING TECHNIQUES

Ravinder Kumar

M.Tech Computer Science, UIET MDU, Rohtak, Haryana

Cite This Article: Ravinder Kumar, "A Study on Predicting Heart Disease with Ensemble Modeling Techniques", International Journal of Multidisciplinary Research and Modern Education, Volume 4, Issue 2, Page Number 123-127,

2018.

Copy Right: © IJMRME, 2018 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

Abstract:

This study introduces a medical recommendation system designed to enhance the accuracy and efficiency of heart disease prediction and diagnostic procedures. Utilizing a soft voting ensemble classifier rooted in machine learning technologies, it aids in forecasting heart disease and suggesting appropriate treatments within the digital health framework. The machine learning algorithms employed significantly increase the precision of cardiovascular disease prediction, thereby reducing the likelihood of unnecessary medical interventions. The effectiveness and practicality of the proposed system are demonstrated through comparative analyses with existing heart disease prediction models. Building on current data mining approaches, the proposed algorithm plays a crucial role in predicting cardiovascular risks and offering relevant recommendations. Soft voting is particularly effective in the early detection of major adverse cardiovascular events in patients experiencing acute coronary syndrome, offering an improved method for managing the incidence of cardiac events. Additionally, the dataset used in conjunction with the proposed system is presumed to be devoid of noise and missing data, ensuring the reliability of the predictions..

Introduction:

The expansive and intricate medical data housed within electronic health databases has drawn research focus towards its application in medical fields, enhancing healthcare services through advanced e-health methodologies. Typically, medical applications engage in six essential activities: screening, diagnosis, treatment, prognosis, monitoring, and management, with this particular study emphasizing the classification and health risk prediction aspects. The primary aim is to broaden the scope of the prediction framework. A soft voting ensemble method is employed, which aggregates predicted probabilities for class labels and determines the class label with the highest cumulative probability. The findings demonstrate notable improvements in accuracy and performance metrics, details of which are elaborated in the results section. This research introduces an effective recommendation system that leverages a soft voting ensemble classifier and machine learning techniques for the early prediction and diagnosis of heart diseases. It evaluates disease risk factors, based on which it advises patients on necessary tests and medications.

Related Work:

Wang et al. (2011) introduced an effective diagnosis of heart diseases using neural net-works ensembles introduces a methodology that works on the basis of SAS base software for heart disease prediction and diagnosis process. It applies neural networks based ensemble methods to effectively predict and diagnose heart diseases. The ensemble model is created using the combination of the measure of the posterior probabilities. In other words, it predicts outcomes through the use of predecessor models such that efficient system models can be constructed. The experiments are conducted using the dataset taken from Cleveland heart disease database. As a result, this system provides increased performance values which, is comparatively better than the existing classification techniques. An ensemble model of cardiac prediction system of using aggregation methodology of BagMoov provides an increase level of performance is given in (Bashir et al., 2015). It provides an envisioned model of aggregation of bagging approach of voting method for diagnosis heart discomfort. It overcomes the drawbacks of the other existing classifiers through the use of basic properties of base classifier methods. This approach is experimentally evaluated with five different types of datasets taken from commonly available data repositories. The results are compared with several classifiers. The results are evaluated using 10-fold cross-validation methods. The results state that this method works with the distinct set of health attributes and provides comparatively higher prediction accuracy measures. It has shown the higher accuracy rate. This approach is experimentally satisfactory for most of the health datasets. An approach to detect and diagnose heart failure using machine learning techniques is given by (Tripoliti et al., 2017). This work emphasizes the severity of heart failure in humans. It provides a clear review of various machine learning algorithms used for the health data classification process. It assesses and presents various models that approximate the subtype, assessment of heart failure severity, the presence of destabilizations, re-hospitalizations and mortality are described in detail. It provides various aspects of existing machine learning algorithms in heart disease prediction with respective advantages and disadvantages.

Proposed Work:

We evaluate performance using by building stacked ensemble model by combining the power of best performing machine learning algorithms. This is divided into 13 major steps which are as follows:

- Data Description
- Importing Libraries & Setting Up Environment
- Loading Dataset
- Data Cleaning & Pre-Processing
- Exploratory Data Analysis
- Outlier Detection & Removal
- Training & Test Split
- Cross Validation
- Model Building
- Model Evaluation & Comparison
- Feature Selection
- Model Evaluation
- Conclusion

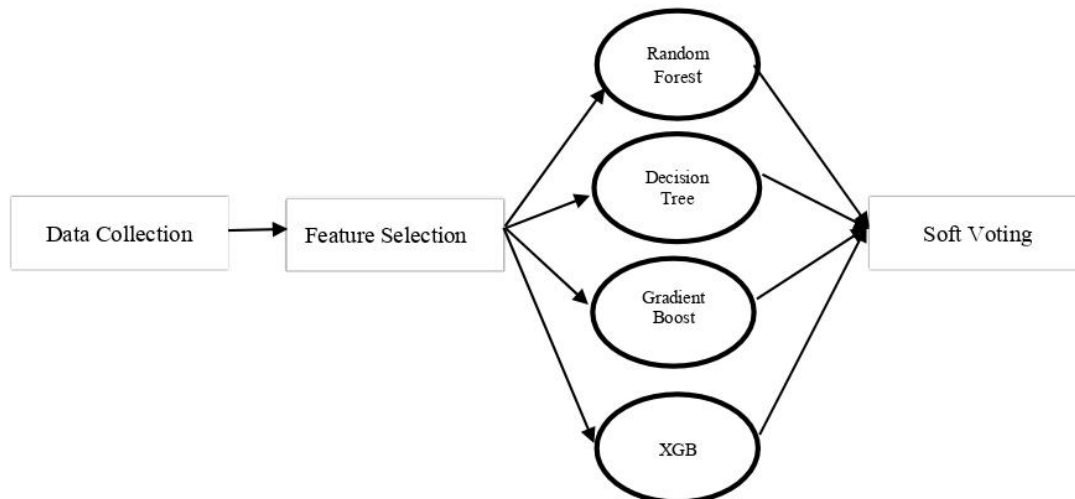


Figure 1: Proposed Architecture

We propose a soft voting ensemble classifier using machine learning algorithms such as random forest (RF), Decision tree, gradient boosting machine (GBM) and XG Boost Classifier for improving the accuracy of diagnosis and prediction of heart diseases.

Dataset Description:

UCI Dataset is used. This database contains 76 attributes, but a subset of 11 of them is selected for analysis. The details of attributes are given in table 1.

Table 1: Attributes Details

S.No	Position in Dataset	Description
1.	#3 (age)	Age in years
2.	#4 (sex)	Sex (1 = male; 0 = female)
3.	#9 (cp)	Chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
4.	#10 (trestbps)	Resting blood pressure (in mm Hg on admission to the hospital)
5.	#12 (chol)	Serum cholesterol in mg/dl
6.	#16 (fbs)	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7.	#19 (restecg)	Resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8.	#32 (thalach)	Maximum heart rate achieved
9.	#38 (exang)	Exercise induced angina (1 = yes; 0 = no)
10.	#40 (oldpeak)	ST depression induced by exercise relative to rest
11.	#41 (slope)	The slope of the peak exercise ST segment -- Value 1: up sloping -- Value 2: flat -- Value 3: down sloping

12.	Target	Diagnosis of heart disease (angiographic disease status) -- Value 0: < 50% diameter narrowing -- Value 1: > 50% diameter narrowing"
-----	--------	---

Performance Metrics:

The proposed system makes use of important performance metrics such sensitivity, specificity, Precision, F1- measure, Geometric mean and mathew correlation coefficient and finally ROC AUC curve for the performance evaluation process.

- **Mathew Correlation coefficient (MCC):** The Matthews correlation coefficient (MCC), instead, is a more reliable statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset.
- **Log Loss:** Logarithmic loss measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. A perfect model would have a log loss of 0. Log loss increases as the predicted probability diverges from the actual label. So, x predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high log loss.
- **F1 Score:** F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.701.

$$F1 \text{ Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Results & Discussion:

	Feature	Pearson	Chi-2	RFE	Logistics	Random Forest	LightGBM	Total
1	st_slope_flat	True	True	True	True	True	True	6
2	st_depression	True	True	True	True	True	True	6
3	max_heart_rate_achieved	True	True	True	False	True	True	5
4	exercise_induced_angina	True	True	True	False	True	True	5
5	cholesterol	True	False	True	True	True	True	5
6	age	True	True	True	False	True	True	5
7	st_slope_upsloping	True	True	True	False	True	False	4
8	sex_male	True	True	True	True	False	False	4
9	chest_pain_type_non-anginal pain	True	True	True	True	False	False	4
10	chest_pain_type_atypical angina	True	True	True	True	False	False	4
11	resting_blood_pressure	False	False	False	False	True	True	2

Figure 2: Feature Selection

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Soft Voting	0.897872	0.872180	0.943089	0.848214	0.906250	0.895652	3.527422	0.797405
1	Random Forest Entropy	0.902128	0.873134	0.951220	0.848214	0.910506	0.899717	3.380449	0.806549
2	MLP2	0.821277	0.813953	0.853659	0.785714	0.833333	0.819686	6.172969	0.641753
3	KNN2	0.812766	0.806202	0.845528	0.776786	0.825397	0.811157	6.466920	0.624631
4	EXtra tree classifier	0.897872	0.877863	0.934959	0.857143	0.905512	0.896051	3.527419	0.796509
5	XGB2	0.897872	0.872180	0.943089	0.848214	0.906250	0.895652	3.527422	0.797405
6	SVC2	0.808511	0.778571	0.886179	0.723214	0.828897	0.804697	6.613914	0.620202
7	SGD2	0.821277	0.804511	0.869919	0.767857	0.835938	0.818888	6.172976	0.642693
8	Adaboost	0.825532	0.825397	0.845528	0.803571	0.835341	0.824550	6.025989	0.650092
9	CART	0.876596	0.856061	0.918699	0.830357	0.886275	0.874528	4.262297	0.753998
10	GBM	0.846809	0.832061	0.886179	0.803571	0.858268	0.844875	5.291121	0.693587

Figure 3: Model Evaluation

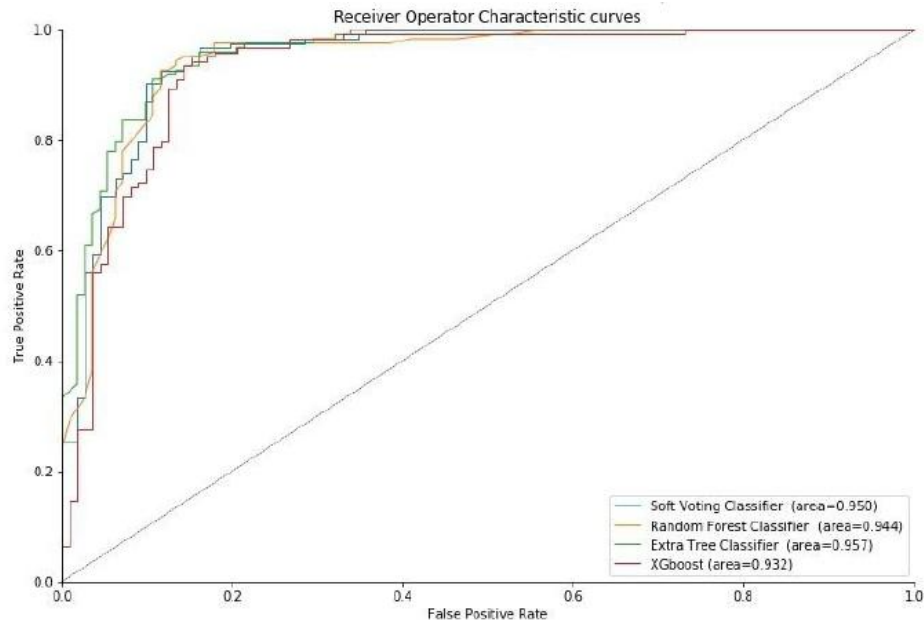


Figure 4: Receiver Operator Characteristic Curves

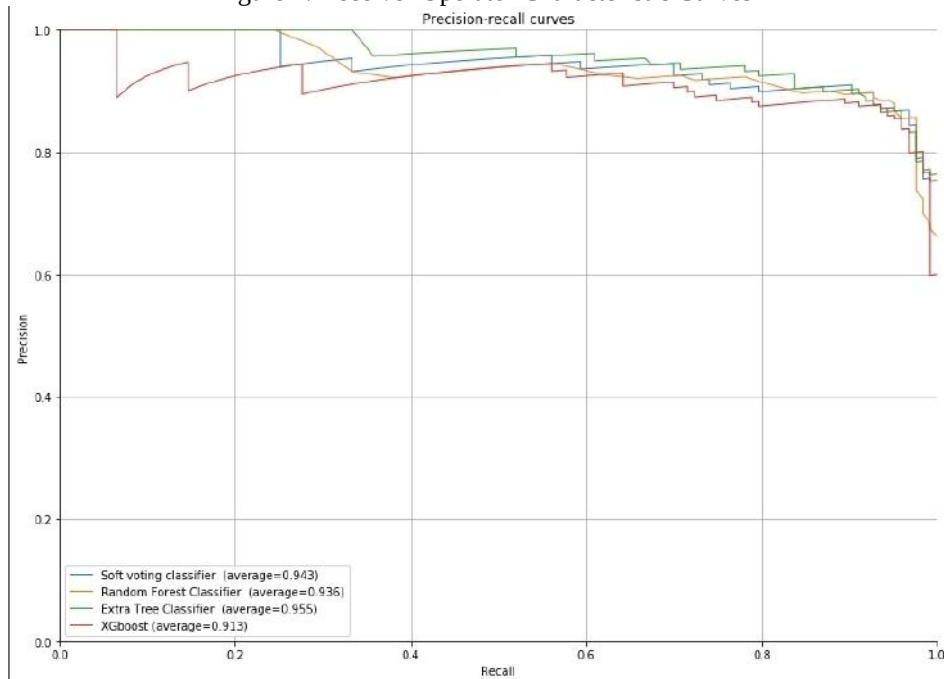


Figure 5: Precision-Recall Curves

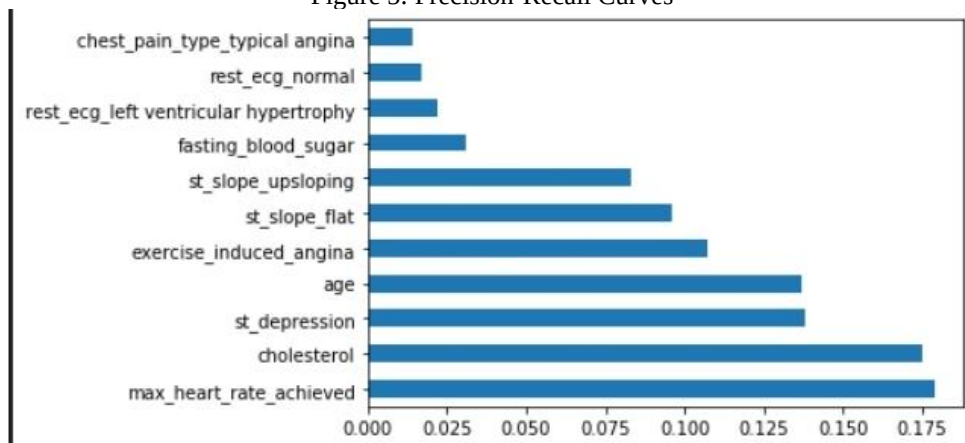


Figure 6: Feature Importance

Conclusion and Future Scope:

The utilization of a stacked ensemble of potent machine learning algorithms has demonstrated superior performance compared to any single machine learning model. Furthermore, an analysis of the second-best performing algorithm, the Random Forest, has been conducted. The five most influential features identified are:

- Maximum Heart Rate Achieved
- Cholesterol Levels
- ST Depression
- Age
- Exercise-Induced Angina

From these findings, it is evident that the recommended system not only achieves higher accuracy metrics but also stands out in efficiency for heart disease patients. It advocates for minimal testing while reducing exposure to health risk factors, thus enhancing patient care. This study contributes a novel approach to personal health record quantification using machine learning techniques and opens up new avenues for health data analysis. Looking ahead, this research suggests two primary directions for future exploration. Firstly, while the current method employs solely supervised learning techniques and focuses on predefined health datasets, there is an opportunity to expand this framework to include unsupervised learning approaches, particularly for managing electronic health records. This would allow for a broader and more nuanced understanding of health data patterns without predefined labels. Secondly, the present work concentrates on structured data; future efforts could extend to incorporate unstructured and semi-structured data, significantly enriching the dataset's complexity and the system's ability to provide comprehensive and nuanced recommendations and diagnoses. This progression towards embracing a wider array of data types and learning techniques promises to further refine and enhance the capabilities of heart disease prediction models.

References:

1. Civicioglu, P. and Besdok., E. (2013), 'A conceptual comparison of the cuckoo-search, particle swarm optimization, artificial bee colony algorithms.', *Artificial Intelligence Review*, 1, 32-38.
2. Yuehjen, Shao, C.-D. H. and Chiu, C.-C. (2014), 'Hybrid intelligent modeling schemes for heart disease classification', *Journal of Applied Soft Computing* 14, 47-52.
3. G.Purusothaman and K.Krishnakumari (2015), 'A survey of data mining techniques on risk prediction: heart disease', *Indian journal of science and technology* 8(5), 643-651.
4. Nguyen, T., Khosravi, A., Creighton, D. C. and Nahavandi, S. (2015), 'Medical data classification using interval type-2 fuzzy logic system and wavelets', *Journal of Applied Soft Computing* 30(21), 812-822.
5. Lafta, R., Zhang, J. and Tao (2016), 'An intelligent recommender system based on short-term risk prediction for heart disease patients', *Journal of web intelligence and intelligent agent technology* (12), 102-105.
6. Singh, M. and Martins (2016), 'Building a cardiovascular disease predictive model using structural equation model and fuzzy cognitive map', *Journal of fuzzy system* 02(6), 1377-1382.
7. Basiliyos Tilahun Betru, Charles Awono Onana, Bernabe Batchakui, "Deep Learning Methods on Recommender System: A Survey of State-of-the-art", *International Journal of Computer Applications* (0975 - 8887) Volume 162 - No 10, March 2017