I.J. Intelligent Systems and Applications, **2025, 1, 88-97**
Published Online on February 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2025.01.06

Data Transformation and Predictive Analytics of Cardiovascular Disease Using Machine and Ensemble Learning Techniques

J. Cruz Antony*
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
E-mail: jcruzantony@gmail.com
ORCID iD: https://orcid.org/0000-0003-0318-1292
*Corresponding Author

E. Murali
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
E-mail: emurali88@gmail.com
ORCID iD: https://orcid.org/0000-0003-2509-4499

D. Deepa
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
E-mail: deepa21me@gmail.com
ORCID iD: https://orcid.org/0000-0003-1543-5368

R. Vignesh
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
E-mail: vigneshramamoorthy438@gmail.com
ORCID iD: https://orcid.org/0000-0002-1355-7276

S. Hemalatha
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
E-mail: hemasira27@gmail.com
ORCID iD: https://orcid.org/0000-0002-0738-9501

Umme Fahad
Astrophysics Research Institute, Liverpool John Moores University, Liverpool, UK
E-mail: umme.phd.1997@gmail.com
ORCID iD: https://orcid.org/0009-0004-3293-3654

Received: 05 July 2024; Revised: 19 September 2024; Accepted: 01 November 2024; Published: 08 February 2025

Abstract: About one person dies every minute from cardiovascular disease; consequently, it has almost surpassed war as the largest cause of death in the twenty-first century. In cardiology, early and accurate diagnosis of heart illness is a cornerstone of effective healthcare. Predictive analytics, which involves machine-learning algorithms, can be a great option for contributing towards the early detection of cardiovascular disease. This study evaluates the data preprocessing techniques involved in building machine learning models to predict cardiovascular disease and identify the features contributing to the cardio attack. A novel data transformation technique named the superlative boundary binning method was proposed to enhance machine learning and ensemble learning classification models for predicting cardiac illness based on independent physiological feature parameters. The results revealed that the ensemble learning classifier AdaBoost using the superlative boundary binning method has performed well with a classification accuracy of 93% when compared with the other data transformation and machine learning classifier models.

Index Terms: Cardiovascular, Data Transformation, Ensemble Learning, Heart Disease, Machine Learning.

1. Introduction

All the organs in the body, including the brain and kidneys, will suffer if the heart is not working correctly. The term "heart disease" refers to a group of illnesses that disrupt normal cardiac function. Currently, the main cause of mortality is cardiovascular disease (CVD). Approximately one person dies every minute from heart disease; consequently, heart disease has surpassed war as the most common cause of death in the twenty-first century. This includes both male and female demographic information, and the sex ratio may fluctuate depending on the age bracket and sex. Nothing here suggests that persons of a particular age are immune to CVD. According to the World Health Organization, 17.9 million people are lost annually due to cardiovascular conditions, cardiovascular disease, coronary disease, heart attack, or heart-related disorders [1]. In developed countries, CVD has a higher mortality rate and is the primary cause of 43% of deaths. This disease is influenced by risk factors such as hypertension, smoking, diabetes, obesity, and physical inactivity. The prevalence of CVD is further exacerbated by ageing populations, unhealthy diets, and sedentary lifestyles, all of which contribute to the increasing incidence of the disease. The CVD's impact is not only limited to physical health but also poses significant economic burdens, with costs associated with healthcare, loss of productivity, and long-term disability [2,3]. Knock is a kind of cardiac disease that may be triggered by high cholesterol levels or by the tightening, blockage, or thinning of blood vessels that provide oxygen to the brain. In today's healthcare system, maintaining a great infrastructure is the biggest problem. The quality of care provided to patients depends on their ability to obtain an accurate diagnosis and appropriate treatment. The devastating results of an incorrect diagnosis are rejected. The quantity of clinical information or data is vast, yet it comes from a wide variety of sources. Doctors' interpretations are a crucial part of this information.

Due to the potential for noise, incompleteness, and inconsistency in real-world data, data preparation will be essential for directives to fill in the null values within the databases. Although CVD has been identified as a leading cause of mortality worldwide, it has also been declared to be one of the most preventable and treatable conditions. Correct and complete illness care depends on an accurate and timely diagnosis of the condition. There seems to be a critical need for a reliable and systematic technique for identifying individuals at high risk, and data mining and machine learning techniques help in the rapid investigation of heart infection [4]. Heart disease symptoms may vary from person to person. However, the most common complaints include aches and pains in the back, shoulders, arms, and jaw as well as digestive issues and difficulty breathing. Cardiac arrest, strokes, and coronary artery diseases are just a few of the many heart-related conditions that may affect a person. Doctors specializing in the heart maintain a large and thorough database of patient information. In addition, this approach provides a fantastic opportunity for extracting useful information from large databases [5].

Classifying the underlying causes and diseases associated with the cardiovascular system has become more difficult in recent years. Intense research is being conducted to identify the factors that put individuals at risk for CVD, with researchers using a wide range of statistical methods and data mining software. A family history of heart disease, high BP, high cholesterol levels, diabetes, advanced age, tobacco use being overweight, and not engaging in enough exercise are recognized as risk factors for CVD, as are high BP and high blood glucose levels. It is crucial to understand the many forms of heart disease to provide better treatment for those at risk and to prevent the illness from occurring first.

2. Literature Review

Data mining was used to conduct a survey of systems for predicting CVD incidence. The report addressed the use of data mining in pattern development to uncover patterns in patient data. This method bolstered and improved their preexisting health care plan. The algorithms they described have a limited scope and were designed to predict CVD incidence. They developed an association rule based on a real informational index and the patient's cardiac health records, which resulted in a high rate of accuracy. The recommended calculations that they performed address not only the problem of a large number of principles but also the proper approval of guideline backing. To facilitate determination as a preliminary task in therapeutic database characterization, kernel F score feature selection was performed. The suggested KFFS strategy first converts healthcare information to binary capacity by increasing BFF and LINEAR skills. Second, in the F score equation, the piece capacities of a non-directly separable set of data are transformed into a straight distinct element space, allowing for the determination of treatment datasets [6,7].

An IHDPS prototype was constructed with the use of data mining methods such as DT, NB, and NN methods. The outcomes demonstrated the distinct value that each method brings to the table in accomplishing the predetermined mining goals. Whereas conventional decision support systems are unable to address such nuanced "what if" concerns, the IHDPS is well suited to do so. The risk of developing heart disease should be predicted based on demographic information, including age, sex, BP, and glucose levels. It permits the establishment of important information, such as patterns and correlations between medical variables associated with heart disease [8].

A study was conducted to detect CVD using machine learning techniques such as multilayer perceptron and K-nearest neighbor methods. Input data were collected from the University of California Irvine repository, which consists of 76 features and 303 sample tuples. The performance of the ML algorithm was evaluated using statistical performance matrices. As a result, a classification accuracy of 82.47% and an area under the curve of 86.41% were obtained from the

MLP model, which performed better than the KNN model [9].

Researchers analyze vast amounts of intricate medical data using a variety of data mining and machine learning approaches, assisting medical personnel in making predictions about cardiac disease. A study was proposed in which 303 cases and 17 attributes of the Cleveland Heart Disease dataset, which were sourced from the UCI machine learning repository, were utilized to generate the data for this study. Numerous supervised classification techniques, including naïve Bayes, decision trees, random forests, and k-nearest neighbors (KKNs), were used. The KKN model achieved an accuracy of 90.8%, which was the best level of precision compared to that of the other supervised models. This study underscores the potential value of machine learning methods for predicting CVD incidence and stresses the significance of choosing the right models and methods to obtain the best outcomes [10].

An attempt to enhance the precision of weaker methods by merging numerous classifiers and improving the reliability of forecasting the risk of heart disease using ensemble classification techniques was conducted. The potential of the ensemble technique for enhancing prediction accuracy in cardiovascular illnesses was investigated using a comparative analytical strategy. Conclusions from these studies suggest that ensemble methods such as bagging and boosting may significantly increase the prediction performance of underperforming classifiers and provide respectable results regarding gauging the likelihood that a certain individual will develop CVD. With the aid of ensemble classification, the accuracy of weak classifiers improved by as much as 7%. The prediction accuracy was significantly increased once a feature selection implementation was added to the procedure [11].

The Fast Fourier Transformation is used to analyze time series and was suggested as part of a MATLAB-based system for recommending treatments for cardiovascular illness. Together, we use an ensemble model that combines bagging, a synthetic neural network, ordinary least support vector machines, and a naïve Bayes classifier [12]. Machine learning models are increasingly vital in understanding and addressing the prevalence of CVD and have been instrumental in the early detection and stratification of high-risk patients, enabling timely treatment. This predictive capability is crucial for implementing preventive measures and reducing the global burden of CVD [13,14]. A novel bio-system reliability approach for prediction of cardiovascular disease was proposed [15]. Dataset was collected from 195 countries, which record death due to annual cardiovascular disease. The method used for prediction of CVD is a multidegree of freedom biosystem. Hence, a reliable high-dimension non-linear spatio-temporal health system was developed, but only a stimulation model was developed.

A study has been conducted on the death of cardiovascular disease patients for a period of 20 years [16]. The different machine algorithms used for prediction are support vector machine, eXtreme gradient boosting, adaptive boosting, logistic regression. As a result, the XGB has obtained higher accuracy than other models. However, the accuracy was found to be less.

For an accurate forecast of cardiovascular disease, have conducted a study using a support vector machine model such as linear SVM and polynomial SVM[17]. The dataset was collected from UCI repository online, which consists of 303 samples and various parameters such as age, sex, chest pain, BP, cholesterol, FBS, RestingECG, Heart rate, EscerciseAngina, STSlope etc. Hence, SVM polynomials have an accuracy of 93% more than linear SVM. The classification method must be improved.

3. Materials and Methods

The dataset consists of asymmetric binary CVD data extracted and loaded into the Python Data frame [18] along with influencing features such as systolic blood pressure (SBP), diastolic blood pressure (DBP), cholesterol level, glucose level, age, height, weight, sex, alcohol intake, and smoking habits. The feature ranking methods, viz., the information gain [19], the gain ratio [20], Gini index [21], and chi-square [22], were applied to the CVD dataset, which includes 12 features, including the target feature CVD occurrence data. These feature ranking methods can acquire relationships among the features that are dependent on each other, can determine the discrimination level of the features in a target feature, and can evaluate the gain of every feature in the context of the target feature to identify effective features for classification. Figure 1 visualizes [23] that systolic blood pressure (SBP), diastolic blood pressure (DBP), cholesterol level, glucose level, age, and weight play vital roles in contributing to the classification of CVD.

The proposed architecture (Figure 2) incorporates the data transformation technique using the superficial boundary binning method to transform continuous features (systolic blood pressure, diastolic blood pressure, age, height, and weight) into discrete buckets. The pseudocode for the superlative boundary binning method is given in Table 1. Data discretization is the process of translating a continuous numerical feature into a set of two or more discrete buckets, thus improving the predictive modelling task [24]. The discretized features are easier to use and understand and are also closer to a knowledge-level representation that can improve model performance than the same models trained on the original dataset without performing data discretization [25]. Data discretization is commonly achieved using binning methods, such as equal width and equal frequency. The equal-width method partitions the dataset into ‘n’ intervals, with each interval containing the same number of data points. The number of bins required is represented by 'n', and the width 'W' of each bin is calculated as

$$W = (\max - \min)/n$$

Where ‘max’ and ‘min’ refer to the highest and least values in the dataset. The range of each bin bracket is defined as [min + w], [min + 2w] …. [min + kw]. On the other hand, the equal-frequency method partitions the dataset into *k* bin brackets where each bin bracket contains approximately an equal amount of data points.

	Info. gain	Gain ratio	Gini	χ^2
N SBP	0.151	0.080	0.097	355.554
N DBP	0.114	0.065	0.075	248.217
C cholesterol	0.038	0.036	0.025	192.444
N age	0.035	0.017	0.024	123.235
N weight (kg)	0.027	0.013	0.018	89.824
C glucose	0.008	0.010	0.005	40.524
N height (cm)	0.001	0.001	0.001	0.195
C physical_active	0.001	0.001	0.001	0.882
C smoking	0.000	0.001	0.000	1.242
C alcohol	0.000	0.000	0.000	0.044
C gender	0.000	0.000	0.000	0.000

Fig.1. Feature ranking of the CVD dataset

Since the physiological features are independent, the data points of each feature vary widely. Therefore, using the same interval for all features is unfair in the data discretization process as it might lead to incorrect model classification. To address this issue, the superlative boundary binning method was proposed. It sorts each feature separately to calculate the maximum difference between data points, and boundaries are drawn to achieve discrete brackets for individual features.

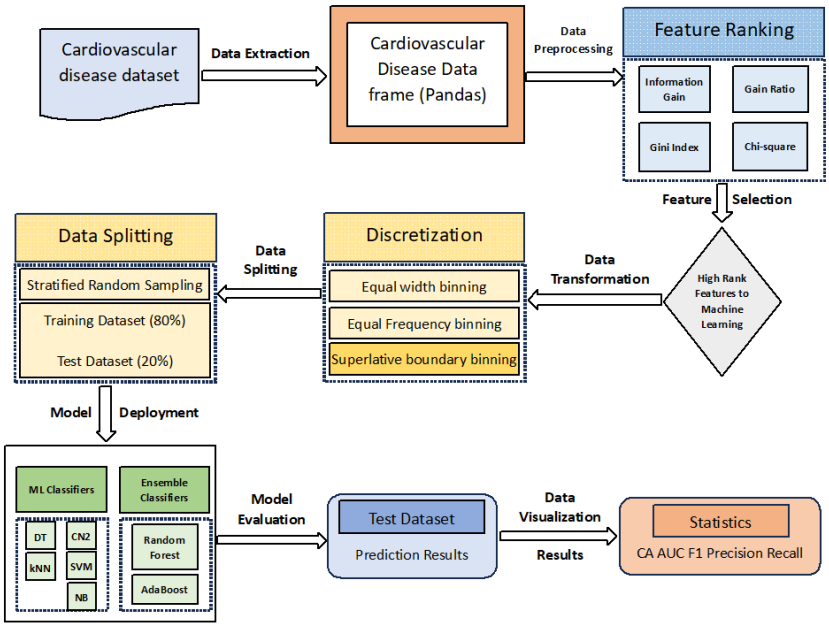

Fig.2. Architecture diagram of the proposed study

After the completion of the proposed preprocessing steps, the dataset was split into training and test datasets using the stratified random sampling technique. In predictive analysis, the data sampling techniques involved in data splitting have a compelling effect on the performance of the model [26]. Finally, machine learning algorithms (decision tree, CN2 rule induction, support vector machine, k nearest neighbour, and naïve Bayes), ensemble learning algorithms (random forest), and boosting algorithms (AdaBoost) were built upon the CVD training dataset to determine whether a given patient has this condition with the help of the test dataset. Machine learning algorithms have the potential to classify various issues pertaining to agriculture, chronic diseases, the stock market, etc. Among the various machine learning algorithms, decision tree (C4.5), support vector machine, k nearest neighbour, naïve Bayes, and CN2 algorithms are popular in breast cancer prediction [27] and diabetic prediction [28]. In the present scenario, the enormous surge in data has devastated traditional machine learning algorithms when handling high-volume, noisy, and imbalanced data [29,30]. To overcome and improve the performance of machine learning classifiers, multiple classification methods using a hybrid

or ensemble learning approach have been adopted [31]. The basic idea of ensemble learning is to merge two or more machine learning classifiers to be trained on the dataset to support and reduce bias and variance and thus increase the prediction accuracy and performance of the model [32].

Table 1. Pseudocode for the superlative boundary binning method

Procedure superlative boundary (info: no. of tuples, divergence: array of difference between tuples, stance: position of max. difference between tuples)

```
//Sorting in Ascending order
d=length(info)
repeat
nd=0
for x=0 to d-1 inclusive do
if info[x-1] > info[x] then
swap(info[x-1],info[x])
nd = x
end if
end for
d=nd
until d=0
//Max. divergence
for y=1 to d inclusive do
divergence[y-x]=info[y]-info[y-1]
end for
//Position of Max. divergence
d=length(divergence)
max=divergence[0]
for z=0 to bin-1 inclusive do
for l=0 to d-1 inclusive do
if divergence[l]>max then
max=divergence[l]
stance[z]=1
end if
end for
end for
end procedure
```

4. Results and Discussion

After modelling the CVD dataset with various machine learning and ensemble learning algorithms, extensive simulations were carried out to identify an efficient model for predicting whether a patient with given symptoms is likely to experience a CVD attack. The proposed ML and Ensemble models were evaluated with the aid of a confusion matrix [33] and five important performance metrics, which included the Classification Accuracy (CA), Area Under Curve (AUC), F1, Precision, and Recall scores [34].

The model is classified into two categories. In the first category, the models were built using machine learning classifiers based on the traditional discretized dataset (equal-frequency and equal-width binning methods) and the proposed superficial boundary binning method dataset; the results are displayed in Figures 3, 4, and 5. In the second category, the models were built using ensemble classifiers on the same discretized dataset discussed above, and the corresponding performance results are displayed in Figures 6, 7, and 8.

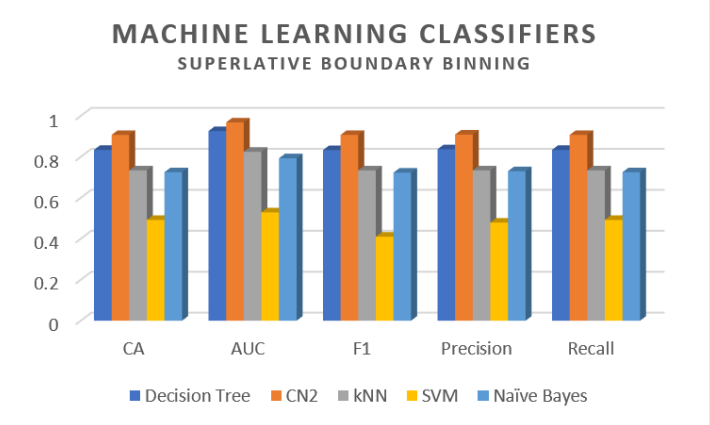

Fig.3. Performance of the ML models built on the superlative boundary binning dataset

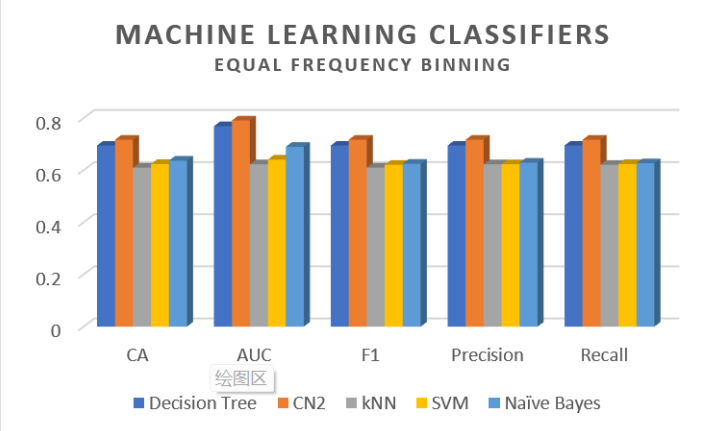

Fig.4. Performance of the ML models built on the equal-frequency binning dataset

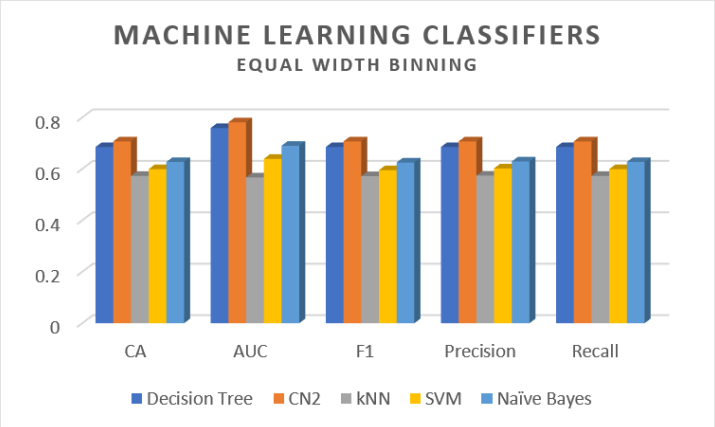

Fig.5. Performance of the ML models built on the equal width binning dataset

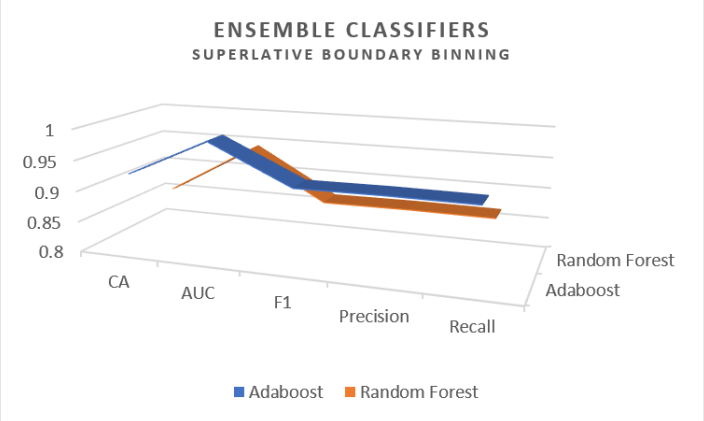

Fig.6. Performance of ensemble models built on superlative boundary binning dataset

The performance matrices in Figures 3, 4, and 5 clearly indicate that the CN2 machine learning classifier model built upon the cardiovascular dataset discretized using the superlative boundary binning method (91% classification accuracy) outperformed the other ML models that used the traditional binning method for discretization. The CN2 ML classifier is a learning classification algorithm for implementing a rule induction model. The CN2 model generates IF-THEN rules, which are articulated and precise for humans [35]. The IF-THEN rules derived from the CN2 model built on the superlative boundary binning dataset are given in Table 2.

The IF-THEN rules given in Table 2 clearly indicate that patients are vulnerable to cardiovascular attack when their systolic blood pressure (SBP) is greater than 139 mmHg, their weight is between 69.5 kg and 75.5 kg, and their age is between 46.5 and 52.5 years with high cholesterol and glucose levels. There is no risk for patients when their SBP is less than 120 mmHg, their weight is less than 64 kg, their diastolic blood pressure (DBP) is less than 80 mmHg with a normal glucose level and low cholesterol, or no smoking habits.

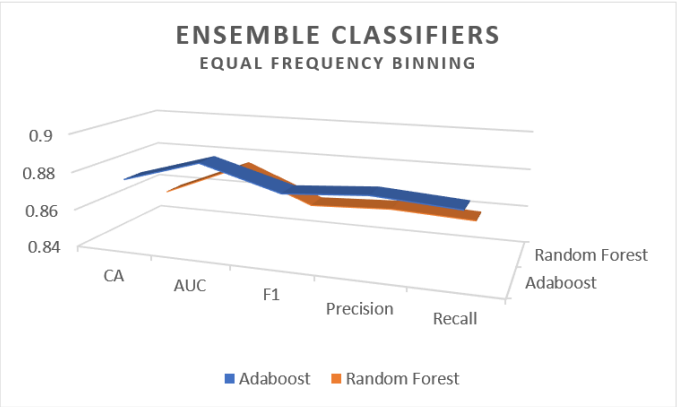

Fig.7. Performance of ensemble models built on an equal-frequency binning dataset

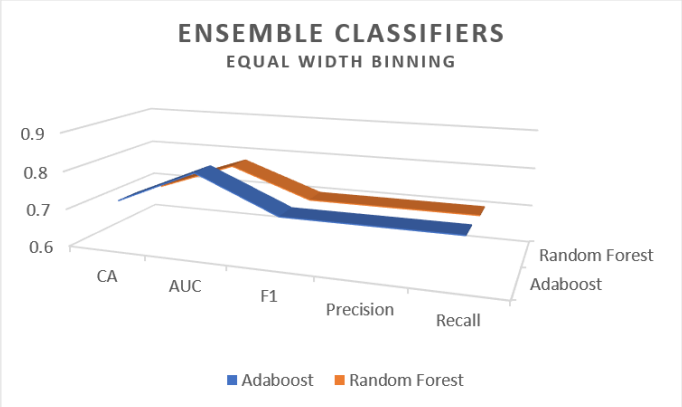

Fig.8. Performance of the ML models built on the equal width binning dataset

Table 2. IF-THEN rules (the CN2 model) generated using the superficial boundary binning method

Sno	Antecedent	Consequent (Class)
1	SBP >= 145.5 AND Weight = 69.5 – 75.5 AND Cholesterol = 2 – 3	1
2	SBP >= 145.5 AND Weight = 69.5 – 75.5 AND DBP = 79.5 – 80.5	1
3	SBP >= 145.5 AND Gender = 1 AND Glucose = 3	1
4	SBP = 138.5 – 145.5 AND Age = 46.5 – 52.5 AND Height = 167.5 – 171.5 AND Gender = 2	1
5	SBP <=119.5 AND Weight <= 63.5 AND Age = 46.5 – 52.5 AND DBP <=79.5	0
6	SBP <=119.5 AND Weight <= 63.5 AND Height >=171.5 AND Smoking = 0	0
7	Age <= 46.5 AND Height = 158.5 – 163.5 AND Glucose = 2 AND Cholesterol = 1	0

The performance matrices in Figures 6, 7, and 8 indicate that the AdaBoost ensemble model built with the superlative boundary binning dataset (93% classification accuracy) outperformed the RF classifier model implemented with the traditional binning methods for discretization. In ensemble learning, both the AdaBoost and random forest models perform well in classification, but the AdaBoost model is considered to be insensitive and has dominance over the random forest model when dealing with imbalanced datasets [36,37]. The ANOVA tests were conducted to determine the statistical significance of performance differences between machine learning and ensemble models. Under the null hypothesis, it was assumed that all the models were equivalent. Thus, rejecting this hypothesis suggests differences in the performance of the models implemented [38]. Subsequently, a post-hoc test using Turkey's HSD[39] was conducted to determine if the control or proposed model displays statistical differences compared to the other methods in the comparison.

According to the ANOVA results from Table 3, the p-value of the CN2 machine learning model using equal frequency, equal width, and superlative boundary binning methods is less than 0.05 (i.e. 9.856933350251298e-10). Therefore, we reject the null hypothesis, indicating that at least one model's performance is significantly different from the others. Furthermore, Turkey's HSD test shows that there are no statistically significant differences between the equal width and equal frequency methods. However, there is a substantial difference between the equal frequency and superlative boundary binning methods, as well as between the equal width and superlative boundary binning methods.

Table 3. ANOVA and HSD test results for the CN2 machine learning models

F-statistic: 49.2283702213281, p-value: 9.856933350251298e-10						
Group1	Group2	mean diff	p-adj	Lower	Upper	Reject
Eq-freq	Eq-width	-0.012	0.3558	-0.0333	0.0093	False
Eq-freq	Sup-bound	0.067	0.0	0.0457	0.0883	True
Eq-width	Sup-bound	0.079	0.0	0.0577	0.1003	True

The p-value (Table 4) of the AdaBoost ensemble learning model using equal frequency, equal width, and superlative boundary binning methods is less than 0.05 (i.e. 2.816016551276829e-12). Therefore, the null hypothesis is rejected again and the Turkey's HSD test shows that all the model pairs, equal frequency, equal width and superlative boundary binning method have statistically significant differences.

Table 4. ANOVA and HSD test results for the AdaBoost ensemble learning models

F-statistic: 83.30913642052562, p-value: 2.816016551276829e-12						
Group1	Group2	mean diff	p-adj	Lower	Upper	Reject
Eq-freq	Eq-width	-0.026	0.0061	-0.0451	-0.0069	True
Eq-freq	Sup-bound	0.07	0.0	0.0509	0.0891	True
Eq-width	Sup-bound	0.096	0.0	0.0769	0.1151	True

5. Conclusions

Early detection and prediction of CVD are crucial in healthcare to save human lives. Machine learning models are better than data mining and statistical models because they can handle large imbalanced datasets and can be used for personalized prediction. Machine learning models are compelling for the treatment of chronic diseases. Hence, in this study, an AdaBoost model with a superficial boundary method was incorporated to predict chronic cardiovascular disease. People with high systolic blood pressure, diastolic blood pressure, and cholesterol levels older than 50 years, and obesity are at risk of cardiovascular attack. We evaluated the ability of ensemble and machine learning classifiers to predict CVD incidence, and the AdaBoost ensemble model achieved an accuracy of 93% when the dataset was preprocessed using the superlative boundary binning method. Thus, this study contributes to the early warning of CVD, and the proposed model can also be used to predict other chronic diseases. However, the proposed model can be further enhanced by adding more patient data with physiological features including diabetes, lipid levels, sodium intake, smoking habit, ambient air pollution, and ECG. By adding these features, the ML model is expected to perform well in real world scenario without any bias and overfitting issues.

Compliance with Ethical Standards

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Conflict of Interests: The authors declare that there is no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors

References

[1] Wacker-Gussmann, Annette, and Renate Oberhoffer-Fritz, “Cardiovascular risk factors in childhood and adolescence,” Journal of Clinical Medicine, Vol. 11, no. 4, pp. 1136, 2022.

[2] Roth, Gregory A., et al. "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study." Journal of the American college of cardiology 76.25 (2020): 2982-3021.

[3] Oleg Gaidai, Yu Cao, Stas Loginov, “Global Cardiovascular Diseases Death Rate Prediction”, Current Problems in Cardiology, Volume 48, Issue 5, 2023, 101622, ISSN 0146-2806, https://doi.org/10.1016/j.cpcardiol.2023.101622

[4] K. Vanisree, S. Jyothi, “Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks,” International Journal of Computer Applications, Vol. 19, no. 6, pp. 6-12, 2011.

[5] S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, “Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data,” PLOS, Vol. 1, no.12, 2017.

[6] M. Thiyagaraj, G. Suseendran, “Survey on heart disease prediction system based on data mining techniques,” Indian Journal of Innovations and Developments, Vol. 6, no. 1, pp.1-9 2017.

[7] C.S. Dangare, S.S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,” International Journal of Computer Applications, Vol. 47, no. 10, pp.44-48, 2012

[8] S. Palaniappan, R. Awang, “Intelligent heart disease prediction system using data mining techniques,” IEEE/ACS International Conference on computer systems and Applications, pp.108-115, 2008.

[9] Pal, Madhumita, Smita Parija, Ganapati Panda, Kuldeep Dhama, and Ranjan K. Mohapatra, “Risk prediction of cardiovascular disease using machine learning classifiers,” Open Medicine, Vol. 17, no. 1, pp. 1100-1113 2022.
[10] Shah, Devansh, Samir Patel, and Santosh Kumar Bharti, “Heart disease prediction using machine learning techniques,” SN Computer Science. Vol. 1, pp.1-6, 2020.
[11] C.B.C. Latha, S.C. Jeeva, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques,” Informatics in Medicine, pp. 100203, 2019.
[12] Raid Lafta, Ji Zhang, Xiaohui Tao, Yan Li, Xiaodong Zhu, Yonglong Luo and Fulong Chen, “Coupling a Fast Fourier Transformation with a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment,” IEEE, pp. 10674-10685, 2017.
[13] K Tsarapatsani, V Pezoulas, A Sakellarios, V Tsakanikas, W Marz, M Kleber, L Michalis, D Fotiadis, Prediction of all-cause mortality in cardiovascular patients by using machine learning models, European Heart Journal, Volume 43, Issue Supplement_2, October 2022, ehac544.1185, https://doi.org/10.1093/eurheartj/ehac544.1185
[14] Rahmanul Hoque, Masum Billah, Amit Debnath, S. M. Saokat Hossain and Numair Bin Sharif, Heart Disease Prediction using SVM, International Journal of Science and Research Archive, 2024, 11(02), 412–420
[15] Oleg Gaidai, Yu Cao, Stas Loginov, “Global Cardiovascular Diseases Death Rate Prediction”, Current Problems in Cardiology, Volume 48, Issue 5, 2023, 101622, ISSN 0146-2806, https://doi.org/10.1016/j.cpcardiol.2023.101622.
[16] K Tsarapatsani, V Pezoulas, A Sakellarios, V Tsakanikas, W Marz, M Kleber, L Michalis, D Fotiadis, Prediction of all-cause mortality in cardiovascular patients by using machine learning models, *European Heart Journal*, Volume 43, Issue Supplement_2, October 2022, ehac544.1185, https://doi.org/10.1093/eurheartj/ehac544.1185
[17] Rahmanul Hoque, Masum Billah, Amit Debnath, S. M. Saokat Hossain and Numair Bin Sharif, Heart Disease Prediction using SVM, International Journal of Science and Research Archive, 2024, 11(02), 412–420
[18] David, Andrew, “A. Introducing Python programming into undergraduate biology,” The American Biology Teacher, 2021.
[19] Stiawan, Deris, Mohd Yazid Bin Idris, Alwi M. Bamhdi, and Rahmat Budiarto, “CICIDS-2017 dataset feature analysis with information gain for anomaly detection,” IEEE Access 8, pp. 132911-132921, 2020.
[20] Cilia, Nicole Dalia, Claudio De Stefano, Francesco Fontanella, and Alessandra Scotto di Freca, “A ranking-based feature selection approach for handwritten character recognition,” Pattern Recognition Letters, Vol. 121, pp. 77-86, 2019.
[21] Liu, Haoyue, MengChu Zhou, Xiaoyu Sean Lu, and Cynthia Yao., “Weighted Gini index feature selection method for imbalanced data,” 2018 IEEE 15th international conference on networking, sensing and control (ICNSC), pp. 1-6, 2018.
[22] Trivedi, Shrawan Kumar, “A study on credit scoring modeling with different feature selection and machine learning approaches,” Technology in Society, pp. 101413, 2020.
[23] E. Murali and S. Margret Anouncia, "Visualization of Multiple Ontology Agro Knowledge Mining Model", International Journal of Reliability, Quality and Safety Engineering, Vol. 29, No. 05, 2022.
[24] M. Pratheepa, and J. Cruz Antony, "Outlook of various soft computing data preprocessing techniques to study the pest population dynamics in integrated pest management.", in Soft Computing for Biological Systems, H. Purohit, V. Kalia, R. More, Eds., Singapore: Springer, 2018, pp. 187-200.
[25] Tsai, Chih-Fong, and Yu-Chi Chen, “The optimal combination of feature selection and data discretization: An empirical study,” Information Sciences, pp. 282-293, 2019.
[26] May, Robert J., Holger R. Maier, and Graeme C. Dandy, “Data splitting for artificial neural networks using SOM-based stratified sampling,” Neural Networks, pp. 283-294, 2010.
[27] Bayrak, Ebru Aydındag, Pınar Kırcı, and Tolga Ensari, “Comparison of machine learning methods for breast cancer diagnosis,” 2019 Scientific meeting on electrical-electronics & biomedical engineering and computer science (EBBT), pp. 1-3, 2019.
[28] Faruque, Md Faisal, and Iqbal H. Sarker, “Performance analysis of machine learning techniques to predict diabetes mellitus,” 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1-4, 2019.
[29] Behar, Nishant, and Manish Shrivastava, "A Novel Model for Breast Cancer Detection and Classification." Engineering, Technology & Applied Science Research Vol. 12, no.6, pp. 9496-9502, 2022.
[30] Dong, Xibin, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma, “A survey on ensemble learning,” Frontiers of Computer Science, pp. 241-258, 2020.
[31] Matloob, Faseeha, Taher M. Ghazal, Nasser Taleb, Shabib Aftab, Munir Ahmad, Muhammad Adnan Khan, Sagheer Abbas, and Tariq Rahim Soomro, “Software defect prediction using ensemble learning: A systematic literature review,” IEEE Access 9, pp. 98754-98771, 2021.
[32] Fitriyani, Norma Latif, Muhammad Syafrudin, Ganjar Alfian, and Jongtae Rhee, “Development of disease prediction model based on ensemble learning approach for diabetes and hypertension,” IEEE Access, pp. 144777-144789, 2019.
[33] Zeng, Guoping, “On the confusion matrix in credit scoring and its analytical properties,” Communications in Statistics-Theory and Methods, pp. 2080-2093, 2020.
[34] Ajagbe, Sunday Adeola, Kamorudeen A. Amuda, Matthew A. Oladipupo, F. AFE Oluwaseyi, and Kikelomo I. Okesola, “Multiclassification of Alzheimer disease on magnetic resonance images (MRI) using deep convolutional neural network (DCNN) approaches,” International Journal of Advanced Computer Research, Vol. 51, 2021.
[35] J. Cruz Antony, and M. Pratheepa, “Study of population dynamics of soybean semilooper Gesonia gemma Swinhoe by using rule induction model in Maharashtra, India”, Legume Research-An International Journal, pp. 369-373, 2017.
[36] Tang, Jiayi, Alex Henderson, and Peter Gardner, “Exploring AdaBoost and Random Forests machine learning approaches for infrared pathology on unbalanced datasets,” Analyst, pp. 5880-5891, 2021.
[37] Shanmugasundar, G., M. Vanitha, Robert Čep, Vikas Kumar, Kanak Kalita, and M. Ramachandran, “A comparative study of linear, random forest and AdaBoost regressions for modeling nontraditional machining,” Processes. 2021.
[38] Abdi, Hervé, and Lynne J. Williams. "Tukey’s honestly significant difference (HSD) test." Encyclopedia of Research Design 3.1 (2010): 1-5.
[39] García, S., Fernández, A., Luengo, J., & Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Computing, 13, 959-977.

Authors' Profiles

Dr. J. Cruz Antony is currently working as an Associate Professor in the Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology, Chennai, India. He has 11+ years of experience working in academic and Govt. R&D institutions. His areas of interest include Machine Learning, Deep Learning, and Natural Language Processing. He is the founder and a designated partner of BENX Solutions LLP recognized as a startup by the Department of Promotion of Industry and Internal Trade, GoI. His startup focuses on building AI-powered automated solutions in the EdTech industry.

Dr. E. Murali is working as an Associate Professor in the Department of Computer Science and Engineering at Sathyabama institute of Science and Technology, Chennai, India. He received his Bachelor's degree in B.E Computer Science and Engineering from Anna University Affiliated College and Master's in M.Tech Computer Science and Engineering from Dr. MGR Research and Educational Institute. He received his Ph.D in Computer Science and Engineering from Vellore institute of Technology. As a part of research work an ontology based incremental mining model was developed. He has published papers in National, International journals and Conference. He has got teaching experience of 12 years and his areas of interest include data mining, ontology mining and Knowledge engineering.

Mrs. D. Deepa is working as an Assistant Professor at Sathyabama institute of Science and Technology, Chennai, India. She received her Bachelor's degree in B.E Computer Science and Engineering from Rajalakshmi institute of science and technology and Master's in M.E Big data analytics from College of Engineering, Anna University, Guindy Campus. She is pursuing her Ph.D at Sathyabama Institute of Science and Technology and her area of research includes Machine learning, IoT, Big data, Blockchain. 7+ years' experience in teaching.

Mr. R. Vignesh graduated from Anna University with Bachelor's Degree in Computer Science and Engineering in 2015 and from Valliammai Engineering college affiliated to Anna University with Master's Degree in Computer science and Engineering in 2017. Currently he is pursuing Ph.D Degree in in Computer Science and Engineering at Sathyabama Institute of Science and Technology .He has more then 7 years of teaching Experience in Department of Computer Science and Engineering at Sathyabama Institute of Science and Technology. His area of research is Block chain and cloud computing. He also presented his paper work in various reputed journals and Conferences.

Mrs. S. Hemalatha is working as an Assistant Professor at Sathyabama institute of Science and Technology, Chennai, India. She received her Bachelor's degree in B.E Computer Science and Engineering from Panimalar Engineering College and Master's in M.E Systems Engineering and operations research from College of Engineering, Anna University, Guindy Campus. She is pursuing her Ph.D at Sathyabama Institute of Science and Technology and her area of research is Machine learning. She has got teaching experience of 3 years and her areas of interest include data mining, machine learning and artificial intelligence.

Umme Fahad is pursuing her data science program Liverpool John Moores University, Liverpool, UK. She received her Bachelor's degree in B. Com from Sri Krishna devaraya University and completed her post executive program from IIITB. Her area of interest includes ML and DL models

How to cite this paper: J. Cruz Antony, E. Murali, D. Deepa, R. Vignesh, S. Hemalatha, Umme Fahad, "Data Transformation and Predictive Analytics of Cardiovascular Disease Using Machine and Ensemble Learning Techniques", International Journal of Intelligent Systems and Applications(IJISA), Vol.17, No.1, pp.88-97, 2025. DOI:10.5815/ijisa.2025.01.06