Cardiovascular Diseases Forecasting using Machine Learning Models

Heba R. Abdelhady [1], Mahmoud M. Ismail[*2]
[1,2] Decision Support Department, Faculty of Computers and Informatics Zagazig University, Zagazig, 44519, Egypt
Emails: HRAbdelhady@fci.zu.edu.eg; mmsabe@zu.edu.eg
*Correspondence: mmsabe@zu.edu.eg

Abstract

Providing medical treatment is a vital part of human existence. Diseases of the heart and blood arteries are often referred to as cardiovascular disease. Predicting cardiovascular illness early on allowed doctors to make adjustments for individuals at high risk, lowering their mortality rate. Machine learning techniques are necessary for making appropriate judgments in the forecasting of cardiac problems because of the vast amounts of medical data available in the healthcare business. Mixed machine-learning approaches are the subject of recent research on unifying these methods. The study proposed machine learning models to predict the heart disease. In order to determine whether or not a person has heart disease, this project presents a model for forecasting. To achieve this, we compare the accuracy of using rules to that of using the Support Vector Machine (SVM), Random forest (RF), and Decision Tree (DT) separately on the dataset.

Keywords: Machine Learning; Forecasting; Cardiovascular; Support Vector Machine; Decision Tree; Random Forest.

1. Introduction

Lifestyle and genetic factors are driving up the cost of healthcare today. Data accumulation occurs naturally over time. This means that the results of any health surveys are being squandered. However, with the advent of data analytics, this is no longer the case. Healthcare facilities and non-governmental organizations (NGOs) are using data to their advantage by deriving actionable insights. Cardiovascular disease is now the leading cause of death in developed countries. The manner this illness manifests in its victims makes speedy treatment difficult. Thus, making accurate diagnoses at the appropriate times is the most challenging aspect of medical practice[1]–[3].

The hospital's terrible reputation stems from its miscommunication and incorrect diagnoses. The Indian government has cast doubt on the widespread belief that the cost of this disease's cure is out of reach. The normal ranges for blood pressure, cholesterol, and heart rate vary from person to person. Medical studies show that a blood pressure of 120 over 80, cholesterol of 200, and pulse rate of 72 are all considered healthy[4]–[6].

The World Health Organisation estimates that cardiovascular disease is responsible for more than 12 million deaths annually. It's a terrible sickness that causes even worse disasters in India. The technique used to investigate the unhealthy situation is intricate[7], [8].

Accurate and exact measurements are required. Due to a dearth of specialists in certain areas, patients are being put in jeopardy. Cardiologists (the doctors who often treat patients with heart problems) are the ones who typically make the diagnosis. Integrating these methods into the healthcare information infrastructure has enormous potential benefits[9]–[12]. In order to determine which machine learning approach is most suited for integrating the survey

Doi: https://doi.org/10.54216/IJAACI.010204
Received: January 18, 2022 Accepted: May 28, 2022

data collected in June, it is necessary to compare many approaches. In this research, we propose a number of machine learning methods for estimating the probabilities of developing cardiovascular illnesses given a set of available variables. The employed medical datasets are drawn from studies that have piqued international interest[13]–[15].

Machine learning (ML), a subfield of artificial intelligence (AI), is finding growing use in cardiology. It's the basis for how computers may analyze information and make judgments or categorizations, with or without human intervention. Conceptually, ML is built around models that take in data (such as photos or text) and use a mix of mathematical optimization and statistical investigation to make predictions. Numerous ML techniques have found use in practical settings. A popular ML technique called SVM, for instance, can identify non-linear patterns for applications like face recognition, handwriting comprehension, and credit card fraud detection. The detection and handling of spam emails have been used by known boosting techniques used for forecasting and categorization. Random forest (RF) is another technique that may help in making judgments by taking the mean of a number of nodes[16], [17].

2. Machine Learning Methods

This section presented the three ML model such as SVM, RF, and DT. Figure 1 shows the steps of the proposed model.

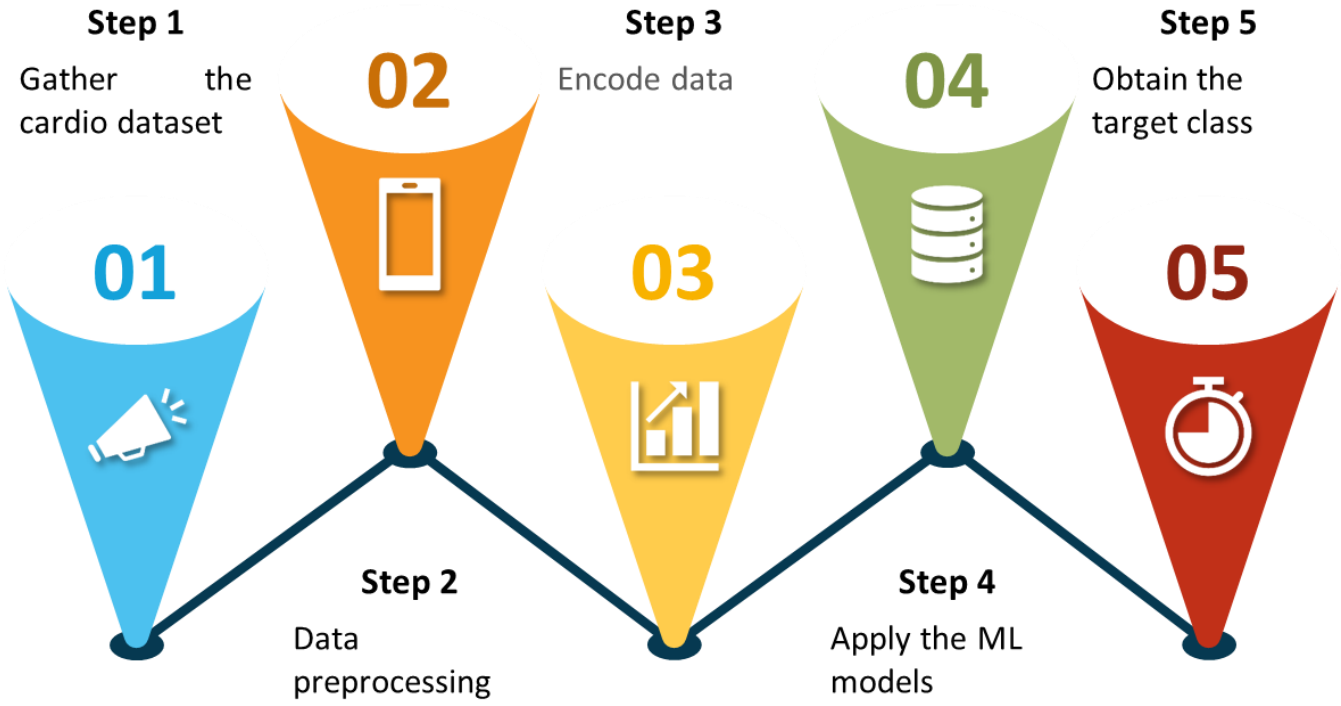

Figure 1: Steps of the suggested model.

2.1 Support Vector Machine (SVM)

The support vector (SV) was introduced by Hearst et al., who used statistical learning theory to categorize it as a nonlinear search technique. The SVM, a subclass of SV developed later to lower machine learning's anticipated error and prevent excessive over-fitting, was presented. SVM, a supervised learning machine that employs the statistical learning theory and the structural risk reduction rule, is widely used in disease modeling. SVM's training

Doi: https://doi.org/10.54216/IJAACI.010204
Received: January 18, 2022 Accepted: May 28, 2022

approach constructs models that allocate novel non-probabilistic binary linear classifiers that, by use of inverse reasoning, minimize the empirical classification mistake and maximize the geometrical margin. SVM is employed to make a future prediction of a quantity by training on historical data. Support vector regression (SVR) is an extension of the SVM that was developed as a regression tool during the last two decades[18], [19].

2.2 Decision Tree (DT)

The ML technique of DT is a useful tool in predictive modelling, especially for simulating disease. DT relies on a decision tree with branches that lead to leaf nodes that contain the desired values. Leaves in a classification tree (CT) indicate the names of classes, whereas branch nodes reflect conjunctions of characteristic labels. The final factors in a decision tree (DT) are discrete sets of values. A regression tree (RT) is a DT where the variable of interest takes on values that are continuous and when many trees are used in an ensemble. There are several similarities and distinctions between regression and classification trees. DTs are often used in ensemble forms to simulate and forecast disease since they are considered quick algorithms[20], [21]. Although the popular DT in ML, the classification and regression tree (CART), has been effectively used to disease modelling, further research is needed to determine whether or not it can be utilised for disease prediction. Another well-liked DT approach for predicting disease is the random forests (RF) technique. Multiple tree predictors are available inside RF. Each branch generates a distinct collection of independent values and response values for the predictor. In addition, the optimal set of classes is determined by a collection of these trees[22], [23].

3. Results and Discussion

This section presented the application and results of the ML applied. In this study, there are three models to apply to the dataset. We collected the dataset from Kaggle. There are 12 variables and 7000 rows in the obtained dataset. Table 1 shows the sample of the dataset. The table 1 shows the 12 variables including age, gender, height, weight, and so on. Then apply some descriptive statistics on the dataset as shown in Table 2.

Table 1: The sample of the dataset

	id	age	gender	Height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	19240	2	168	76.0	120	80	1	1	1	0	1	0
6999	9999	2260	1	158	126.0	140	90	2	2	0	0	1	1

Doi: https://doi.org/10.54216/IJAACI.010204
Received: January 18, 2022 Accepted: May 28, 2022

6	5	1											
69997	99996	19066	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	22431	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	20540	1	170	72.0	120	80	2	1	0	0	1	0

Table 2: The descriptive statistics of dataset.

	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**Max**
id	70000.0	49972.419900	28851.302323	0.0	25006.75	50001.5	74889.25	99999.0
age	70000.0	19468.865814	2467.251667	10798.0	17664.00	19703.0	21327.00	23713.0
gender	70000.0	1.349571	0.476838	1.0	1.00	1.0	2.00	2.0
height	70000.0	164.359229	8.210126	55.0	159.00	165.0	170.00	250.0
weight	70000.0	74.205690	14.395757	10.0	65.00	72.0	82.00	200.0
ap_hi	70000.0	128.817286	154.011419	-150.0	120.00	120.0	140.00	16020.0
ap_lo	70000.0	96.630414	188.472530	-70.0	80.00	80.0	90.00	11000.0
cholesterol	70000.0	1.366871	0.680250	1.0	1.00	1.0	2.00	3.0
gluc	70000.0	1.226457	0.572270	1.0	1.00	1.0	1.00	3.0
smoke	70000.0	0.088129	0.283484	0.0	0.00	0.0	0.00	1.0
alco	70000.0	0.053771	0.225568	0.0	0.00	0.0	0.00	1.0
active	70000.0	0.803729	0.397179	0.0	1.00	1.0	1.00	1.0
cardio	70000.0	0.499700	0.500003	0.0	0.00	0.0	1.00	1.0
year	70000.0	53.338686	6.765294	30.0	48.00	54.0	58.00	65.0

Then apply the three ML models. Figure 2 shows the count of the cardio according to the year. The target class has two classes. The first class no disease (0), and other class has a disease (1)

Doi: https://doi.org/10.54216/IJAACI.010204
Received: January 18, 2022 Accepted: May 28, 2022

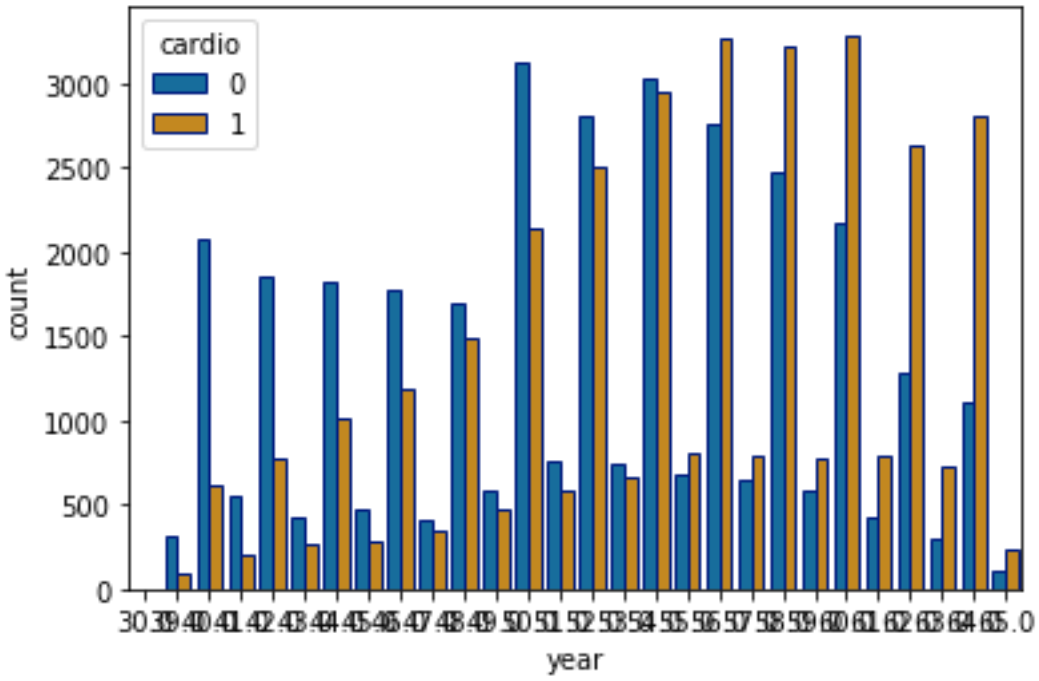

Figure 2: The target class distribution.

After Appling the ML models, we achieved the precision, recall, and f1 score as shown in Table 3. The RF and DT have the largest value followed by SVM.

Table 3: The values of evaluation ML models.

Models	Class	precision	recall	f1-score
RF	0	0.0	0.0	0.0
	1	0.0	0.0	0.0
SVM	0	0.0	0.0	0.0
	1	0.81	1.0	0.89
DT	0	1.0	1.0	1.0
	1	1.0	1.0	1.0

Figure 3. shows the accuracy value of three ML models. The RF has a 100%, SVM has 88.8 accuracy, and DT has 100% accuracy.

Doi: https://doi.org/10.54216/IJAACI.010204
Received: January 18, 2022 Accepted: May 28, 2022

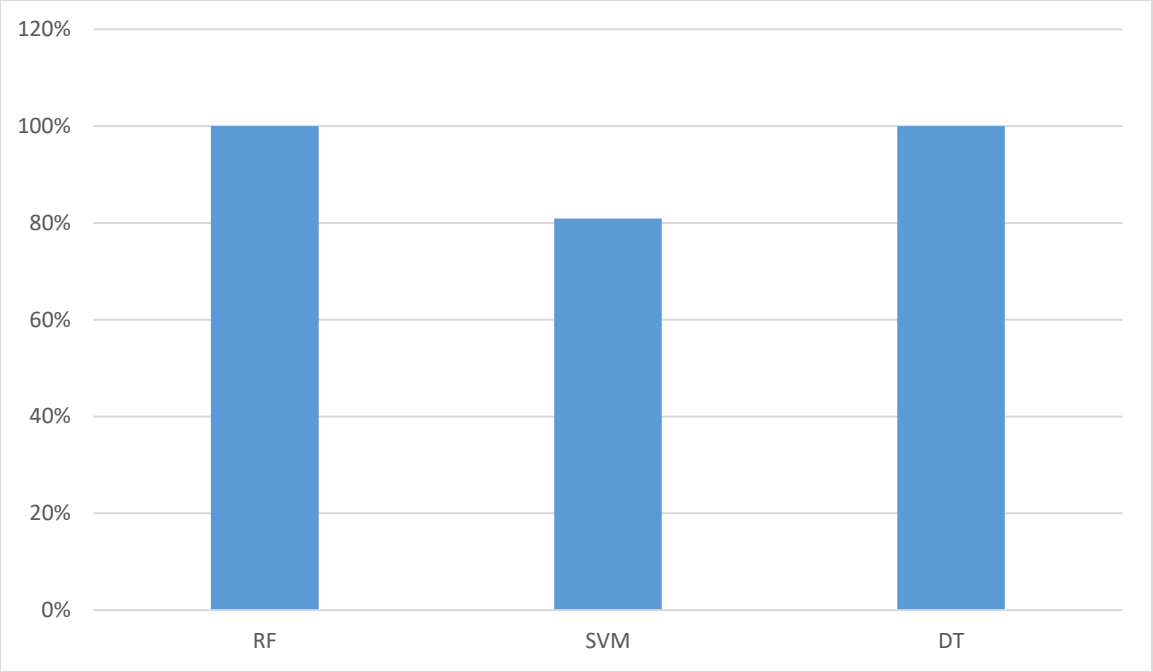

Figure 3: The comparison of three models by accuracy score.

4. Conclusion

Overall, the findings from ML models were encouraging, but there are still a number of obstacles to be cleared before they can be used in clinical use. Successful applications of SVM and boosting techniques are widespread in cardiovascular medicine. However, good clinical context translation requires the selection of acceptable methods for suitable study topics, compared with human experts, testing cohorts, and a summary of all available assessment matrices. Most critically, further research is required to directly compare ML algorithms to more traditional risk models. This study proposed three machine learning models to predict the cardio disease. The ML models are SVM, RF, and DT. The RF and DT have largest accuracy followed by the SVM.

References

[1] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," *Heart*, vol. 104, no. 14, pp. 1156–1164, 2018.

[2] W. Sun, P. Zhang, Z. Wang, and D. Li, "Prediction of cardiovascular diseases based on machine learning," *ASP Trans. Internet Things*, vol. 1, no. 1, pp. 30–35, 2021.

[3] C. Krittanawong *et al.*, "Machine learning prediction in cardiovascular diseases: a meta-analysis," *Sci. Rep.*, vol. 10, no. 1, p. 16057, 2020.

[4] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, IEEE, 2018, pp. 1–7.

[5] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges," *Eur. Heart J.*, vol. 38, no. 23, pp. 1805–1814, 2017.

Doi: https://doi.org/10.54216/IJAACI.010204
Received: January 18, 2022 Accepted: May 28, 2022

[6] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, vol. 12, no. 4, p. e0174944, 2017.

[7] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–15, 2019.

[8] S. Mezzatesta, C. Torino, P. De Meo, G. Fiumara, and A. Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," *Comput. Methods Programs Biomed.*, vol. 177, pp. 9–15, 2019.

[9] P.-Y. Tseng *et al.*, "Prediction of the development of acute kidney injury following cardiac surgery by machine learning," *Crit. care*, vol. 24, no. 1, pp. 1–13, 2020.

[10] G. Kissas, Y. Yang, E. Hwuang, W. R. Witschey, J. A. Detre, and P. Perdikaris, "Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4D flow MRI data using physics-informed neural networks," *Comput. Methods Appl. Mech. Eng.*, vol. 358, p. 112623, 2020.

[11] M. Motwani *et al.*, "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis," *Eur. Heart J.*, vol. 38, no. 7, pp. 500–507, 2017.

[12] E. K. Oikonomou *et al.*, "A novel machine learning-derived radiotranscriptomic signature of perivascular fat improves cardiac risk prediction using coronary CT angiography," *Eur. Heart J.*, vol. 40, no. 43, pp. 3529–3543, 2019.

[13] J. I. Z. Chen and P. Hengjinda, "Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study," *J. Artif. Intell.*, vol. 3, no. 01, pp. 17–33, 2021.

[14] P. Ghosh *et al.*, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[15] B. Ambale-Venkatesh *et al.*, "Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis," *Circ. Res.*, vol. 121, no. 9, pp. 1092–1101, 2017.

[16] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[17] G. Carleo *et al.*, "Machine learning and the physical sciences," *Rev. Mod. Phys.*, vol. 91, no. 4, p. 45002, 2019.

[18] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 66, no. 3, pp. 247–259, 2011.

[19] R. G. Brereton and G. R. Lloyd, "Support vector machines for classification and regression," *Analyst*, vol. 135, no. 2, pp. 230–267, 2010.

[20] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[21] Y. Ben-Haim and E. Tom-Tov, "A Streaming Parallel Decision Tree Algorithm.," *J. Mach. Learn. Res.*, vol. 11, no. 2, 2010.

[22] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert Syst. Appl.*, vol. 134, pp. 93–101, 2019.

[23] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 6, p. e5518, 2018.

Doi: https://doi.org/10.54216/IJAACI.010204
Received: January 18, 2022 Accepted: May 28, 2022