International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Description about the Genesis and Role of Machine Learning Techniques in the Prediction of Heart Disease

Abhishek Saxena[1*], Harish Kumar Taluja[2], Neeta Verma[3]

Submitted: 05/01/2024 **Revised:** 15/02/2024 **Accepted:** 22/02/2024

Abstract*:* There is a vast impact and scope for Artificial Intelligence (AI) which is spreading in every sphere. However, major alterations were recently made by it in the medical domain, particularly in the cardiovascular community. By making a framework for its advancement, a noteworthy contribution to utilizing the large multidimensional health sector datasets was made in AI. Therefore, it can be incorporated into the medical domain as of the start of basic laboratory research to clinical application and all the way up to the delivery of healthcare. Due to the difficulty in early diagnosis of Heart Disease (HD)by medical experts, it has emerged over the past ten years as the leading cause of death worldwide. Machine learning, an area of AI, serves as a beacon of hope for medical practitioners since early cardiac disease treatment cannot be applied without accurate prediction. To highlight the AI advancements in healthcare, this study describes the genesis and functions of several Machine Learning (ML) techniques on the dataset of HD extracted as of the UCI machine repository for Cardiovascular Disease prediction. A high level of accuracy for it was exhibited by results acquired from the many methodologies that use it, which also suggests that utilizing ML for it may be a superior option.

Keywords: *Artificial Intelligence, Machine Learning, Support Vector Machine, Naïve Bayes Classifier, Random Forest, Decision Tree*

1. Introduction

Many vital organs are available in the human body with every part mentioning its role and function in the human body's smooth working. Actually, there is an interconnection among those organs, and are mainly responsible for offering aid to the human being's life. The significance of the heart cannot be ignored among these diverse organs. The vital organ that has the capacity to supply blood to various body parts is the Heart, which is also responsible to carry oxygen to the brain. Since the heart reflects the entire growth of a person's life, it is the most valuable organ that is needed to be fit and healthy.

However, HDs have emerged to be the most deadly disease during the past decade. It is found to be the sole reason for the death of many people around the world. This disease still remains a challenge for medical specialists, since the early diagnosis is unpredicted.

The early treatment of HD is a curse without its appropriate prediction. Therefore, to aid healthcare professionals, ML technology's significant role has been identified as a ray of hope by several researchers.

The ML technique's role in the dataset taken as of the UCI machine repository of HD was elaborated on in this paper. 303 values with 76 attributes are contained in this database. However, all the published experiments generally referred to utilize a subset of 14 of them. It is also vital to note that this is the one and only dataset available for HD-related research work and hence utilized for this study also. The dataset is primarily preprocessed and then cleaned to exclude the missing attributes, noise, etc. It is then utilized for training the algorithm. Grounded on various parameters, the results are being analyzed for testing various ML techniques.

2. Machine Learning

In 1969, the term "Machine Learning" was initially introduced by Arthur Samuel. As the term indicates that computers were made to learn from the data by this algorithm along with elevate themselves without being programmed explicitly.

The term 'Learning', which is similar to intelligence, is a broad term that is critical to define it accurately.

ML *is the subset of* AI in simple words. The machines are provided the capacity to learn autonomously by the ML along with elevate them as of the experience without being programmed *explicitly to do so.*

The discipline of AI that is actually for constructing computer programs that can be elevated autonomously via experience is ML. To solve the issues, it would make predictions by the machine's ability to think.

Owing to this, ML [12] *has paved the way for getting absolute analysis regarding its occurrence with vast and rapid technological advancement in the AI field. It also exhibits complete insight regarding its early detection. ML can be mainly utilized to forecast* HD *while it has a large scope in several medical field applications.*

*[1] Research Scholar, Department of Computer Science & Engineering, School of Engineering & Technology, Noida International University, Gr Noida, Uttar Pradesh, India.
Email Id: abhishek09cse@outlook.com
[2] Professor, Department of Computer Science & Engineering, School of Engineering & Technology, Noida International University, Gr Noida, Uttar Pradesh, India. Email Id: harishtaluja@gmail.com
[3] Professor, Department of Computer Science Engineering, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India
Email Id: neeta140@gmail.com*

This remaining paper is partitioned into the ensuing sections: The literature review being implemented by the various researchers in this domain is studied in *Section 3*, the proposed architecture is described in *Section 4*, *Section5* defines the methodology being adopted by specifying the role of various classifiers of ML techniques being used in this work; the obtained outcomes between the various classifiers are described in*Section 6*, *Section 7*defines the references.

3. Literature Review

Cardiovascular diseases are the major reasons behind the trouble in the working of the heart. The early models utilized to forecast it had however aided in making decisions regarding the modifications. A very significant role was played by the heart in the life of living organisms. More accuracy and perfection were required in the diagnosis and prediction of HDsas little error can cause either a fatigue problem or even death of a person [12]. Also, it is estimated that there is an extreme amount of deaths associated with a heart as per several world reports. It is also found that its percentage is growing exponentially at a rapid rate per year.

To spread its preliminary awareness and detection, many researchers found that there is an urgent need for a prediction system [16] to deal with such issues. Complete support and a platform for the prediction of any kind of disease were provided by ML, which is a subset of AI taking data as of natural ways.

A vital role was played by ML techniques in a deeper understanding of medical data in medical healthcare along with emerges as a better option for delivering accurate outcomes as ML models have been utilized for HD detection in the past [13],[14].

A significant feature to retrieve valuable information as of the corporate datasets of large databases is data classification, which is one of the most prominent tasks in ML. Thus, [1] analyzed that some ML techniques that are utilized in the diagnosis of HD are (i) Artificial Neural Networks,(ii) K- Nearest Neighbors (KNN), (iii) Naive Bayes (NB), along with(iv)Logistic Regression (LR).

Since it requires patient data and prior records, it is very difficult to predict in advance if a person will have HD. Therefore, identifying the parameters is a challenging task, which requires the use of extra instruments when making clinical choices regarding HD.

Another researcher [11] utilized the Kaggle dataset with 303 instances for training the model and 14 variables for the class labels, defined as the cause of HD in order to elevate the system's effectiveness. Although they applied a dataset to LR, they came to the conclusion that the proposed approach provided superior accuracy than NB and Random Forest (RF).

Similarly, for the prediction and categorization of patients with HD, the methods of LR, RF, and KNN regression were also employed by another researcher [2]. By utilizing medical history, the dataset in this study comprises patient records for chest discomfort, blood sugar levels, blood pressure, and other conditions. It was discovered that by using these records, as opposed to the NB classifier previously utilized, high accuracy is attained. Since the model uses more training data, it can make more accurate predictions and determine if a certain person has HD or not.

Similar to this, more study has been elaborated by many researchers using these ML techniques [3], [15], [4], [5], [6], [7], [8] [9]. On the UCI dataset [10], various algorithms were tested by these researchers along with discovered good accuracy and outcomes.

4. Proposed Architecture

Studying the impact and the role of ML algorithms, like (i) LR,(ii) KNN,(iii) SupportVector Machine (SVM), (iv) RF, (v) Decision Tree (DT), along with (vi) NB is the chief objective of this work. However, HD prediction is a complicated task along with requires a high level of accuracy. Therefore, the following algorithm has been proposed for this purpose.

ALGORITHM

Step 1: *Choose the dataset of heart disease.*

Step 2: *Perform the preprocessing of data.*

Step 3*: Prepare the data by removing noise and missing values.*

Step 4*: Make proper feature selection in it.*

Step 5*: Split the dataset into'2'portions: 80% for training & 20% for testing.*

Step 6*: Apply the selected algorithm.*

Step *7: Train the model.*

Step 8*: Apply cross-validation in the training data.*

Step 9*: Perform the testing by using 20% of the data.*

Step 10*: Obtain the predicted result.*

5. Methodology

5.1 Dataset Collection:

The dataset in this research is chosen via [11] that had been originally obtained as of the UCI repository [10] containing the Cleveland database. 303 values with 76 attributes are comprised in this database; however, all publically submitted experimentations suggest utilizing a fourteen subset among them. Specifically, this is the unique database that is available for being utilized by several ML researchers up-till now. Hence, the same dataset is utilized in this work for predicting HD irrelevant of the disease type.

Table 1: Attributes & Description of dataset

Features	Type	Description	Values
Age	Real	Patient’s age	in years

Sex	Categorical	Sex	Male = 1,
			Female = 0

CP	Categorical	Sort of Chest Pain	Value one: typical angina Value two: atypical angina Value three: non-anginal pain Value four: asymptomatic
Trestbps	Real	RelaxingBP	Gauged in mm Hg.
Chol	Real	Serum Cholesterol	Gauged in mm/dL

FBS	Real	Fasting blood sugar > 120 mg/dl)	(one = true;
			zero = false)

RestECG	Categorical	Relaxing electrocardiographic outcomes	Value zero: Normal Value one: possessing aberrationin ST-T wave Value two: Estes' criteriashowapossible or certain left ventricular hypertrophy
Thalach		Maximal Heart Rate attained	60-200 bph

Exang	Categorical	Exercise engendered Angina	(yes = 1;
			no = 0)

Oldpeak	Real	exercise-inducedST depression associated with rest	
Slope	Categorical	heavy exercise ST’sslope	Value one: upsloping Value two: flat Value three: down-sloping
Ca	Real	number of foremost vessels colored by fluoroscopy	Values betwixt(zero-three)

Thal	Categorical	Thalassemia (fault type)	normal =three, fixed fault = six, reversible fault =seven
Num	Real	HD diagnosis	Value 1: Yes Value 0: No

5.2 Training Algorithms:

The proposed research work trains and classifies the HDaccuracy by utilizing a variety of ML algorithms, including LR, KNN, SVM, DT, NB Classifier, and Deep Neural Network:

(i) ***Logistic Regression***: An algorithm utilized for classification in ML is LR. It is grounded on the probability concept and is employed to assign the data into a discrete class. The target or dependent variable can only have one of two possible types, either 1 or 0, in the binary or binomial version of LR, which is the simplest type. The LR hypothesis allows a model to specify the link between several predictor variables and binomial target variables with a tendency to constrain the cost function ranging betwixt 0 and 1. It also transforms the output using the sigmoid logic function. The linear function present in it is principally utilized as an input to another function, such as g in the ensuing relation −

$h\theta(x) = g(\theta Tx)$ *where* $0\leq h\theta\leq 1h\theta(x) = g(\theta Tx)$ *where* $0\leq h\theta\leq 1$

Here, the logistic or sigmoid function is defined as g, which can be offered as −

$g(z) = 11+e-z$ *where* $z = \theta Tx$

(ii) ***K-Nearest Neighbor***:-The prediction of the new categorization sample was enabled by a nonparametric lazy learning algorithm named KNN. It could be applied to forecast issues involving regression and classification. It is also used by many diverse groups. However, it is mostly employed in the classification of industrial issues because it does well across all the parameters being looked at. As it is simple to understand and requires less computing time, it is typically chosen for evaluating the effectiveness of a technique.

(iii) ***Support Vector Machine***:-SVM is utilized to categorize linear and non-linear data. By maximizing the marginal distance for both classes along with diminishing categorization errors, each data item is primarily mapped into an n-dimensional feature space. Here, the quantity of features is signified as n. It next identifies the Hyperplane (HP) that partitions the data items into '2' classes. The distance betwixt the decision HP and its nearby instance, a member of that class, is the marginal distance for a class. Formally, every data point is plotted at first as a point in n-dimensional space with each feature's value signified by the specific coordinate's value. Next, by locating the HP that isolates both classes by the greatest margin, the classification is executed.

(iv) ***Naïve Baye's Classifier***: A group of probabilistic classifiers grounded on the NB theorem is the NB classifier. Here, identifying the posterior probabilities is the main interest, that is, a label's probability provided with few observed features, P($L \mid features$). It can be articulated in quantitative form with the assistance of Bayes theorem as -

P(L|features)=P(L)P(features|L)P(features)

P(L|features)=P(L)P(features|L)$P(features)$

Where, the class's posterior probability is signified as($L \mid features$), the class's preceding probability is notated as (L), the provided class's predictor probability is specified as ($features \mid L$) and ($features$) is the predictor's preceding probability. To make predictions, the NB classifier relies heavily on its assumption of robust independence between features [17]. It is therefore ideal for utilization in the field of medical science for identifying disorders since it works well most of the time and is quite simple to construct.

(v) ***Decision Tree***:- DT learning is a methodology for approximating discrete-valued target functions. To give classification, it sorts the instances from root to leaf of the tree. To assess the quality of the split, it then assesses the homogeneity of the target variable within these subgroups. With internal nodes, branches, and a terminal node, it is a powerful and accurate algorithm for predictive modeling. With a class label intended for each leaf node, the internal node in this performs a "test" on the features; also, the branches contain the test results.

5.3 Evaluating Methods*:*

By concentrating on a few criteria of metrics, like accuracy, F-score, ROC, AUC, precision, and recall, the proposed model's evaluation is executed.

(i) Accuracy: is measured to identify correct values for classification. It is one amongst the utmost vital performance metrics. It is delineated as the addition of entire true values divided by total values, as shown :

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

(ii) Recall: is utilized to define the methodology's capacity to predict positive values. The minor portion of adequate examples over the entire quantity of applicable examples that are recovered is termed Recall. It is gauged as:

$$\text{Recall or Sensitivity} = \frac{TP}{(TP + FN)} \quad (2)$$

(iii)Precision: is utilized to analyze and predict how frequently a model's positive value is correct. It is delineated as the total True Positives (TP)divided by the overall number of predicted positive values. It is given as:

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (3)$$

(iv) F1 Score: the measurement of the mean value of precision together with recall is referred to as the F-1 score. It is signified as:

$$\text{F1 Score} = \frac{2 \text{ x } TP}{2 \text{ x } TP + FN + FP} \quad (4)$$

(v) Receiver Operating Curve: It exhibits the classification system's effectiveness at all classification levels. It is also termed the receiver operating characteristic curve. TP and false positive are the '2' parameters displayed utilizing this curve. The area under the curve (AUC), which is utilized as a ROC curve description, is a gauge of the classifier's capability to diversify betwixt the classes. The model is more effective at distinguishing between positive and negative groups if the AUC is higher.

6. The Results

6.1 Data Preprocessing:

Table 2: The Summary Statistics is

	Age	Sex	cp	-	caa	thall	output
Count	303.000	303.000	303.000	-	303.000	303.000	303.000
Mean	54.366337	0.683168	0.966997	-	0.729373	2.313531	0.544554
Std	9.082101	0.466011	1.032052	-	1.022606	0.612277	0.498835
Min	29.000000	0.000000	0.0000000	-	0.0000000	0.000000	0.000000
25%	29.000000	0.000000	0.000000	-	0.0000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	-	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	-	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	-	4.000000	3.000000	1.000000

6.2 Result for LR:

Table 3: Classification Report for LR

	Precision %	Recall %	F1-score %	Support %
0	0.94	0.76	0.84	38
1	0.85	0.96	0.90	53
Accuracy			0.88	91
Macro Average (M.A.)	0.89	0.86	0.87	91
Weighted Average (W.A.)	0.89	0.88	0.88	91

Accuracy score is: 87.9%

6.3 Result for KNN

Table 4: Classification Report for K-NN

	Precision %	Recall %	F1-score %	Support %
0	0.87	0.68	0.76	38
1	0.80	0.92	0.86	53
Accuracy			0.82	91
M.A.	0.83	0.80	0.81	91
W.A.	0.83	0.82	0.82	91

Accuracy Score is : 82.41%

6.4 Result for Support Vector Machine:

Table 5: Classification Report for SVM

	Precision %	Recall %	F1-score %	Support %
0	0.90	0.71	0.79	38
1	0.82	0.94	0.88	53
Accuracy			0.85	91
M.A.	0.86	0.83	0.84	91
W.A.	0.85	0.85	0.84	91

Accuracy Score: 84.61%

6.5 Result of Naive Bayes Classifier:

Table 6: Classification Report for NB Classifier

	Precision %	Recall %	f1-score %	Support %
0	0.88	0.74	0.80	38
1	0.83	0.92	0.88	53
Accuracy			0.85	91
M.A.	0.86	0.83	0.84	91
W.A.	0.85	0.83	0.84	91

Accuracy Score: 84.61%

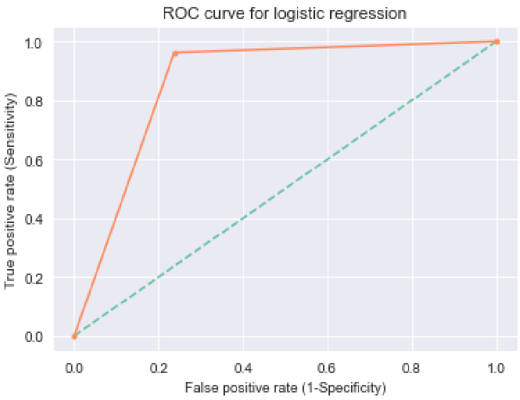
6.6 Result of Decision Tree

Table 7: Classification Report for DT

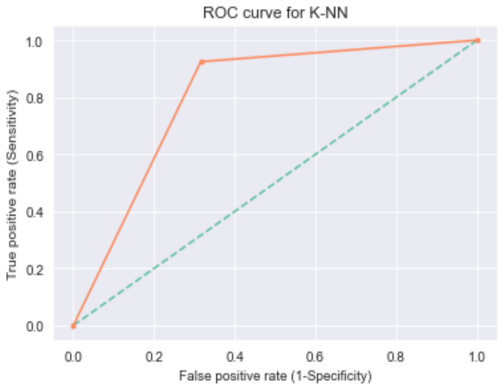
	Precision %	Recall %	f1-score %	Support %
0	0.70	0.79	0.74	38
1	0.83	0.75	0.79	53
Accuracy			0.77	91
M.A.	0.77	0.77	0.77	91
W.A.	0.85	0.83	0.84	91

Accuracy Score: 76.92%

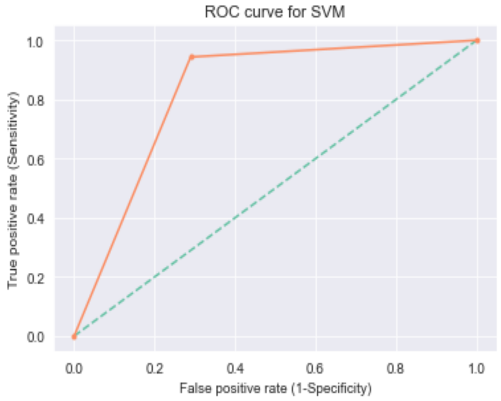
6.7 Results of the ROC Curve:



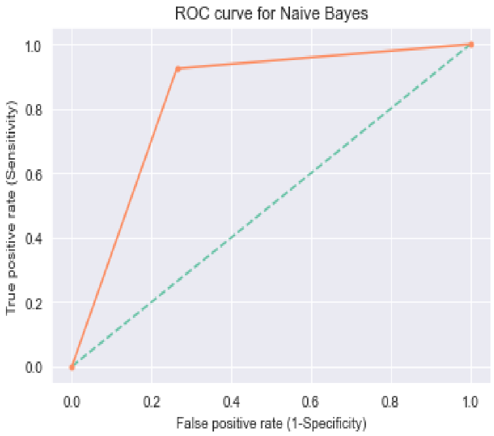
(i) LR



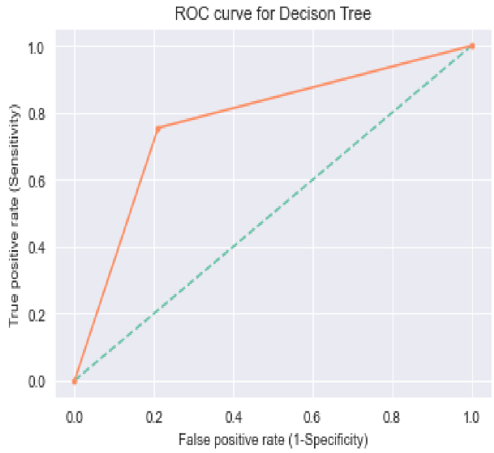
(ii) K-NN



(iii) SVM



(iv) NB



(v) DT

6.8 Comparison Chart:

Various classifiers give the accuracy for the above dataset used as follows:-

Table 8: Accuracy Comparison Chart

S.No	Algorithm	Accuracy
1	LR	87.9%
2	KNN	82.41%
3	SVM	84.61%
4	NB	84.61%
5	DT	76.92%

It has been witnessed grounded on the above accuracy charts that a good accuracy was given by the various classifiers of ML techniques on the above dataset.

Declarations:

Conflict of interest: The authors declare that they have no conflict of interest.

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

Consent of publication: Not applicable.

Availability of data and materials: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Competing interests: The authors declare that they have no competing interests

Funding: This work has no funding resource.

Author's contributions: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [1]**Abhishek Saxena,** [2]**Harish Kumar Taluja,** [3]**Neeta Verma.** The first draft of the manuscript was written by [1]**Abhishek Saxena** and all authors commented on previous versions of the manuscript.

All authors read and approved the final manuscript.

Acknowledgement: We thank the anonymous referees for their useful suggestions.

References

[1]. Yadav SS, Jadhav SM, Nagrale S, Patil N (2020) Application of machine learning for the detection of heart disease, 2nd International Conference on Innovative Mechanisms for Industry Applications, (ICIMIA), pp. 165–172, Bangalore, India, March 2020

[2]. Harshit Jindal et al (2021) Heart disease prediction using machine learning algorithms. IOP Conference Series: Material Science and Engineering. DOI 10.1088/1757-899X/1022/1/012072

[3]. Aljanabi M, Qutqut H, Hijjawi M (2018) Machine learning classification techniques for heart disease prediction: a review. International Journal of Engineering and Technology. 2018;7:5373–5379.

[4]. Shah D, Patel S, Bharti SK (2020) Heart disease prediction using machine learning techniques. SN Computer Science 1(6):1-6. https://doi.org/10.1007/s42979-020-00365-y

[5]. Santhi P, Ajay R, Harshini D, Jamuna Sri SS (2021) A Survey on Heart Attack Prediction Using Machine Learning. Turkish Journal of Computer and Mathematics Education. 12(2):2303-2308.

[6]. Rubini PE, Subasini CA, Katharine AV, Kumaresan V, Kumar SG, Nithya TM (2021) A Cardiovascular Disease Prediction using Machine Learning Algorithms. Annals of the Romanian Society for Cell Biology 25(2):904-912.

[7]. Norman A, Harding J, Zhukova D (2021) Machine learning in the health industry: predicting congestive heart failure and impactors. SMU Data Science Review 5(1).

[8]. Sharma C, Shambhu S, Das P, Jain S, Sakshid. Features Contributing Towards Heart Disease Prediction Using Machine Learning, ACI'21: Workshop on Advances in Computational Intelligence at ISIC, February 25-27, 2021, Delhi, India.

[9]. Boukhatem C, Youseff HY, Nassif AB (2022) Heart Disease Prediction using Machine Learning, Advances in Science and Engineering Technology International Conferences (ASET) 1-6. DOI:10.1109/ASET53988.2022.9734880

[10]. UCI Machine Learning Repository. Available from: https://archive.ics. uci.edu/ml/index.php. Accessed November 01, 2018.

[11]. https://www.kaggle.com/heart-disease-uci?select=heart.csv

[12]. Awan SE, Sohel F, Sanfilippo FM, Bennamoun M, Dwivedi G (2018) Machine learning in heart failure. Current Opinion in Cardiology 33(2):190-195. doi:10.1097/hco.0000000000000491

[13]. Dinesh KG, Arumugaraj K, Santhosh KD, Mareeswari V (2018) Prediction of cardiovascular disease using machine learning algorithms. IEEE. DOI: 10.1109/ICCTCT.2018.8550857

[14]. Singh A, Kumar R (2020) Heart disease prediction using machine learning algorithms. International Conference on Electrical and Electronics Engineering (ICE3). DOI: 10.1109/ICE348803.2020.9122958.

[15]. Sanz González, R, Luque Juárez, J, M.ª, Martino, L, Liz Rivas, L, Delgado Morán, J, J, & Payá Santos, C, A. (2024) Artificial Intelligence Applications for Criminology and Police Sciences. International Journal of Humanities and Social Science. Vol. 14, No. 2, pp. 139-148. https://doi.org/10.15640/jehd.v14n2a14

[16]. Almustafa KM. Prediction of heart disease and classifiers' sensitivity analysis, BMC Bioinformatics. 2020;21.

[17]. Lafta R, Li Y, Tseng VS (2015) An Intelligent Recommender System based on Short Term Risk Prediction for Heart Disease patients, IEEE/WIC/ ACM International Conference on Web Intelligence and Intelligent Agent Technology. Singapore: IEEE.

[18]. Repaka AN, Ravikanti SD, Franklin RG (2019) Design And Implementing Heart Disease Prediction Using Naives Bayesian, 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India. 292-297. doi:10.1109/ICOEI.2019.8862604.