VOL. 19, NO. 12, JUNE 2024 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

HEART DISEASE PREDICTION USING KNN ALGORITHM APPROACH

S. Ghouhar Taj and K. Kalaivani
Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies (VISTAS), PV Vaithiyalingam Rd, Velan Nagar, India
E-Mail: sgtaj786@gmail.com

ABSTRACT

Machine getting to know and records-based totally techniques for predicting and diagnosing coronary heart ailment may be an extraordinary medical gain, but it's far a chief undertaking to improve. In many countries there may be a shortage of cardiovascular professionals and a giant quantity of instances of misdiagnosis may be addressed by way of setting up correct and powerful early-stage cardiac forecasts with the aid of medical decision-making analysis of virtual patient information. This has a look at aimed picking out excessive-overall performance devices getting to know variants for such diagnostic purposes. Several gadget-getting-to-know algorithms have been used which might be in comparison and done with accuracy and accuracy in predicting heart disease. The scores of the importance of each detail are restrained to all algorithms used except MLP and KNN. All elements are calculated based totally on the cost points to find those who provide high-danger coronary heart disease prognosis. The look discovered that using Kaggle's 3-segment cardiac database based on pro-k (KNN), choice tree (DT), and random Forests (RF) RF technique algorithms done ninety-seven.2% accuracy and 97.2% sensitivity to make clear. Therefore, we've observed that an easily supervised machine gaining knowledge of a set of rules can be wielded to make coronary heart disease conjecture with the very best accuracy and the most satisfactory possible use.

Keywords: MLP, KNN, choice tree, random forest, cardiac database.

Manuscript Received 23 April 2024; Revised 15 July 2024; Published 5 September 2024

1. INTRODUCTION

Cardiovascular illnesses (CVD) are presently the number one purpose of dying internationally and the Arena Fitness Organization 2020 anticipated this to be around 17.9 million deaths every 12 months [1]. Early-degree detection of CVD is a critical manner of decreasing this toll. Of the numerous techniques for improving this disorder detection and prognosis is records mining. Those strategies that relate permit hidden knowledge to be extracted and to become aware of relationships among attributes in the dataset, and this is a promising approach for the CVD category [2-4]. The delivery of excessively great medical offerings that might be low-cost to sufferers is an important venture dealing with health companies. The shipping of appropriate service calls for both correct analysis of sufferers and identification of effective treatment, whilst fending off erroneous diagnoses [5]. Early-degree detection of CVD additionally minimizes price and reduces CVD mortality. Facts mining techniques can do the task successfully at a totally low value using a class set of rules, which performs a key role in scientific research [6]. here we investigated how such reasonably-priced and simple algorithms might be of enough utility to use clinically and factor the manner to impr Cardiovascular ailment (CVD) is presently the leading purpose of loss of life globally and the sector fitness employer 2020 estimates that this demise rate is estimated at 17.9 million annually. Early detection of CVD is an essential way to lessen this toll. In lots of approaches to improve the prognosis and diagnosis, statistics mining. Those associated strategies allow encrypted records to extract and perceive relationships between databases within the database and are a promising method for classifying CVD [2-4]. The transport of highly satisfactory low-cost medical offerings to sufferers is a prime assignment. Managing fitness companies. Correct service shipping calls for both proper patient diagnosis and powerful remedy identification, while keeping off misdiagnosis [5]. Early detection of CVD additionally reduces costs and reduces CVD mortality. Information mining strategies can perform paintings efficiently at very low prices with the use of a category set of rules, which performs a vital position in scientific research [6]. Here we've explored how cheap and simple algorithms may be beneficial enough for clinical use, and point the way for improved services.

2. RELATED WORK

Researchers have used a variety of mining statistics in conjunction with organizational policies, favors, and integration to develop a predictive coronary heart prediction model. Shiva Kazempour Dehkordi1 and Hedieh Sajedi suggested a speculative version hinged entirely on the medical doctor's system to use the mining statistics method [7]. They have changed a set of rules shout Skating to improve the precision of the gadget. Skating, like Boosting and Bagging, is an integrated method. On a different label, they examined four class algorithms: DT, Nave Bayes (NB), Enough Close Friends (KNN), and Skating. They demonstrate that a high-precision finer is impacted. This departmental algorithm was 73 percent accurate. However, when compared to other class algorithms and methodologies, this is a low-performance approach. Jan *et al.*, in 2018 wield an integrated mathematical method of using a limited value calculator collected within the UCI garage area (mainly

Cleveland and Hungary) where a mixture of 5 precise departmental algorithms and an RF, neural network. , NB, segregation by descent screening machine and assisted assisting machine (SVM) has changed into a lease [8].

They point to what appears to be the performance of a set of low-level policies to become a retrospective, as in the experiment, RF gave a maximum precision of 98.136%. In 2011, Soni .J *et al*. DT appears to be wielding a set of genetic codes to improve normal classroom performance, and this is comparable to various algorithms including NB and filtering techniques [9]. They found ninety-nine percent accuracy.2% within the proposed device. In 2017, Hend Mansoor et al. investigated the performance of LR and RF classification methods in compensating for exposure risks in CVD patients [10]. They demonstrated that the LR version outperformed the RF break-up method. The LR version was 89 percent accurate, whereas the RF version was 88 percent accurate. In 2013, Austin et al. compared the actual quality of typical deciduous trees to deciduous trees [6]. Normal LR ensures fulfillment = "hide"> fantastic = "tipsBox"> in determining the existence of HD. In 2018, Le *et al*. employed three strategies to improve 58 reference characters in data obtained by the device's UCI master [11].

They demonstrated that the line vector assist (SVM) with linear kernel offered the best performance, with 89.99 accuracy. Tarawneh and Embarak suggested a 12-item strategy and compared its performance against KNN, J48, GA, DT, synthetic neural community (ANN), SVM, and NB [12]. When compared to the special algorithms employed, the suggested cover method yielded 89.2 percent accuracy, which is just as effective as "hide "> excellent =" tipsBox "> basic. In 2013, Chitra and Seenivasagam proposed using a first-line cascaded neural network (CNN) to improve the accuracy of forecasting cardiac disease [13]. CNN employs a cascade structure in which a network is reinforced with neurons that are locked one after the other, unaltered after that, and which contains a hidden community with a network of neurons. The suggested method's final result is compared to SVM, which delivers 82 percent, and CNN 85 percent with accuracy and specification of 0.87 and 0.7775, respectively. They upgraded CNN as a class predictor of coronary heart disease with unsurpassed accuracy after examining these parameters, and the model with CNN exactifier is more accurate than SVM. To increase the accuracy of the divider, Latha and Jeeva employed a split-sharing approach with a Cleveland mathematical set and merged a majority vote with MP, RF, BN, and NB using a feature selection method [14]. Six sets of separators were assessed for results and outcomes.

They build what is considered one of the included types and look for performance to get a better version of the aggregate. They stated that the majority of people voting for MP, RF, BN, and NB wield the career choice technique provided 7339ff1fc90882f8f31ca1efdd2ac191 universal performance with eighty 5.48% accuracy and suggested this mixed process for predicting heart disease. However, it is now feasible to identify techniques that outperform their intended version in terms of accuracy. Mohan *et al*., 2019 introduced a speculative version of the cardiovascular system employing integrated device study strategies [15]. In this project, they utilised Rattle, a graphical user interface data mining device that employs R, to differentiate HD mainly based on data acquired from the Cleveland UCI archive. This has resulted in enhanced degree performance, with an HD hypothesis version with a hybrid RF with linear (HRFLM) model achieving 88.7 percent accuracy. When they compared the suggested version to the = "extraordinary =" tipsBox "> differentiation algorithm, they demonstrated that their employed model outperformed the other different classifiers. The use of machine learning and recording techniques is clear in the discussion above. -10. The study aims to identify those who differentiate people who can successfully have heart disease so that they can benefit from the clinic.

3. EXPERIMENTAL SETUP

In this look, a variety of statistics, solid and rapid facts about the heart, are investigated to construct our predicted model. In Kaggle, the website has altered and accumulated [16]. This database has 14 characteristics. Table 1 provides suggestions for all appendices. The website has information on 1025 persons who have been impacted, including 713 males and 312 different temporary women, of whom 499 (48.68 percent) are healthy and 526 (51.32 percent) have heart disease. Three hundred (57.03 percent) of heart disease patients are male, while 226 (42.97 percent) are female.

For preliminary analysis statistics we have to wield the 3. eight.three models as a tool for digging facts in behavioral research and the Python 3. eight.5 model analysis and identification (EDA) facts. Preliminary information processing is mandatory for any application of the application or method of digging facts, as the overall performance of the device that receives the device information depends on how well the records are organized and supported. The uncertain value was first used to manage deficit values, after which another filter, called the Inter quartile range (IQR), was changed to be wielded to determine foreign and translucent values in the pre-processing component. IQR is a method of measuring diversity when it comes to the record range. The outlier is an information element that makes up more than the predicted statistical range and can be considered for analytical purposes or not due to telerecord errors or non-major events [17]. Machine literacy or mining knowledge techniques [18] are essential for extracting such external material for better analytical or mathematical results. Based on the findings, the records were classified into three deciles: Q3, Q2, and Q1. Q1 and Q3 are information limits in this case. Count the i value of IQR by IQR = Q3 - Q1. Then lower border Bl and upper border Bu were obtained wield the following equations [19]: right here, the lower impact than Bl and more than Bu is considered outside. To grade an uneven collection of data, the synthetic minority oversampling method (SMOTE) was applied. therefore, some experimental records analysis

(EDA) was achieved (just like the box shape) to ensure that the information did now not include outside information, and the facts have been represented as an IQR and temperature map to determine the correlation between the functions, and the KDE plan for both sufferers and those who aren't sick in step with age. Figure-1 indicates the work flow of data analysis.

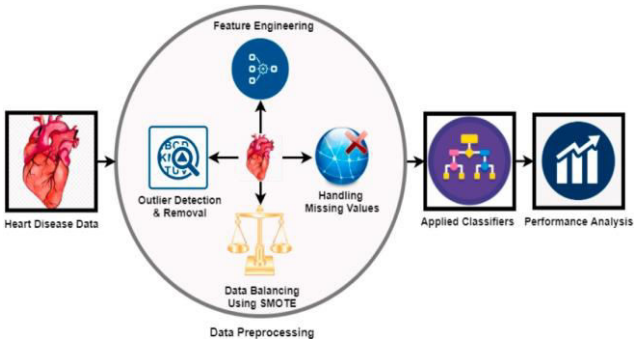

Figure-1. Work flow of data analysis.

$$B_l = 1.0 - 1.50*IQR \quad (1)$$

$$B_u = Q_3 + 1.50*IQR \quad (2)$$

3.1 Feature Extraction

Table-1 explains the details of the feature.

Table-1. Details of feature.

S. No	Attribute Name
1	Age
2	Gender
3	Blood presure
4	serum
5	fasting
6	Electro
7	Heart
8	Exercise
9	ST des
10	ST seg
11	fluoroscopy
12	thal
13	Target

3.2 Performance Evaluation Metrics

Six (06) department algorithms have been used inside the database to acquire the fine-performing algorithm against the accuracy and different mathematical variables with 10-fold verification. Multilayer perception (MP), nearest k acquaintances (KNN), random forest (RF), deciduous tree (DT), drop-down (LR), and AdaboostM1 were the algorithms employed (ABM1). These algorithms are compared based on the measures used to evaluate their overall performance. This newsletter provides a brief overview of this performance assessment.

A confusion matrix changed into acquired to obtain the sensitivity, specificity, and precision of the outcome of every set of rules. The formulation stated under is used to calculate all parameters [11, 13]:

$$\text{Sensitivity} = TP / TP + FN \quad (3)$$

$$\text{Specificity} = TN / TN + FP \quad (4)$$

$$\text{Specificity} = (TP + TN) / TP + FP + TN + FN \quad (5)$$

$$TPR = TP / TP + FN \quad (6)$$

$$FPR = FP / FP + TN \quad (7)$$

Right here, TP and TN indicate true and terrible finance, respectively, while FP and FN represent both false and bad; TPR represents true wonderful points and FPR is a massive false positive. Sensitivity is related to the role of the actual distortion defined by the divider accurately as facts and reflects the number of positive predictions using the divider using well-defined categories [20]. Clarification is the ability of the editor as he has to distinguish the negative effects [20]. The accuracy of the number of cases is well categorized according to section [11, 13, 20].

Kappa calculations, accuracy, do not forget, f-measure, Matthew coefficient (MCC), receiver performance (ROC), and profitable memory are some of the mathematical variables used to evaluate the success of various algorithms (percent). The classification agreement based on the diagnosis and prediction of positive characteristics is measured by Kappa records [21]. Accuracy is a viable test matrix, particularly when the proposed ML version must be evaluated based on predictable and realistic outcomes [20, 21]. Calculates the proportion of actual payments that surpass estimates. As a final result, it depends on the values of FP and TP. If it is important miles to obtain the extent to which the best predictors can be expected, remember another useful metric [20, 21], which represents part of the deliberately decorated results. The TP and FP values are used to assess memory recall. F-measure maintains a balance between precision and memory that distinguishes it from the rest of the class. The F-measure is the precise number between 0 and 1 that represents the highest mathematical values of accuracy and recall [20, 21].In the knowledge machine, MCC is used to test the integrity of binary and multi-stage categories. It describes the positive and negative features of truth and untruth and is frequently regarded as a stability matrix that may be employed even with subjects of varying proportions. MCC is, in fact, a composite of variables ranging from -1 to +1 [22]. These parameters are estimated using the following methods [22-24]:

$$Kappa, K = \frac{(P_r P_r(a) - P_r(e))}{1 - P_r(e)} \quad (8)$$

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$F - Measure = \frac{2*precision*recall}{precision+recall} \quad (11)$$

$$MCC = \frac{(TP*TN - FP*FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (12)$$

Supervised gadget gaining knowledge of algorithms unique varieties of supervised machine gaining knowledge of algorithms were used in this look. In supervised system mastering algorithms, a categorized statistics set is first used to get you up to speed with the primary algorithm. This applicable model is then uploaded to a categorized studies database to classify associated categories [25]. A brief review of these proposed systems studying formulation for diagnostics is supplied within the accompanying paragraph.

3.3 Nearest Neighbor (KNN)

KNN is one of the oldest and finest aviation algorithms [26, 27] or mathematical analysis approaches [28]. Fair refers to the number of close friends utilised, which may be defined quickly using the founder or computed if the needed figure is provided [24]. The same circumstances are therefore problematic in the same categories [29], and the new item is classified in each case using its similarity rate of 5bf1289bdb38b4a57d54c435c7e4aa1c [30]. If the unmarked sample is frequent, the sample space k adjacent to the anomalous pattern will be tested by a surrounding set of criteria. Predictions from numerous friends may be determined from tests mostly based on their grades, and surely one of the variables included to transform the difference into weight [28, 31]. A set of recommendations has several advantages since it evaluates and is extremely easy to utilise [28]. The detector is not very educated, but it is quite useful in identifying illnesses, particularly HD detection because it operates with a single event. The range n neighbors 2 and leaf size 40 were modified to a legitimate site parameter in this experiment.

3.3.1 Random forest

RF is a method of classifying records using DT-based studies [32]. It creates a great variety of woods and in addition produces a wooded area for logging, while it is shy of the pedagogy phase [33]. Each tree, a part of the tree area, predicts each class label during the test. If the beauty label is predicted for all trees, it means that the majority of voters are used to determine the final choice in all test statistics [34]. The category label that receives the most important type of vote is considered to be the maximum number of appropriate labels used in test facts. In all records within a series of records, this cycle is repeated. Random types of randomized controlled trials of this look like 123, which provided amazing performance of used data.

3.3.2 Decision tree (DT)

DT is one of the most well-known and widely used learning algorithms. DT creates option information using a mechanism that evaluates and connects the data department's output into a tree-like structure [25].DT tends to have a wide range of nodes, a very good category called root cutting or different sub-nodes. Differences in inputs or trends are displayed across all nodes = "hide"> internal = "tipsBox"> which include at least one child node. Relying on the final results of the observation, the separation strategies find the right location for the child, where the experimental and door-to-end approach ends before the leaf position is extended [34]. Leaves or terminal buses interact with the selection results. DT is considered pure for comprehension and study and is an important part of clinical diagnostic agreements [35]. The maximum depth of this algorithm release has shifted to 7 definitions and a divider with this high resolution produced 7339ff1fc90882f8f31ca1efdd2ac191 recordings used in this look.

3.3.3 Adaboost M1 (ABM1)

One of the most popular learning approaches is ABM1. It uses a curved developing technique to provide better segmentation outcomes in the form of a few notably distinct divisions and a strong one [36]. All apparent symptoms are given the same weight at the first level. The size of the magnitude error is used to calculate the differential coefficient, which varies with the common phase coefficient. As a consequence, the classifier's error rate is regarded as the separator's coefficient. As a result, the ABM1 algorithm can raise the load of random vaccination while decreasing the weight of the randomly created watch. The following repeat will stress the maximum weight in the incorrectly placed entertainment area. Finally, all high-level dividers are joined into a solid component using a line binding method [37] to offer the overall performance of a section. When the number of n estimators is set to 100, this separator achieves good universal effectiveness in this test.

3.3.4 Logistic regression (LR)

LR distinguishes supervised machine learning methods [38] and is an extension of the usual retrieval model used on a website, expressing the likelihood or probability of an event occurring [39]. When LR finds the possibility for a new target component of a certain class, the outcome is as close to 0 as feasible. As a result, the limit is specified, which specifies the partition of LR usage into two groups. For example, the number of opportunities stated above 0.5 is referred to as 'category A,' rather than 'category B.' The LR model may be expanded as a multi-item retraction to design phase flexibility with more than two variables [40].In this investigation, the most random event number is 1234, and 100 equals equally discovered on the website utilised.

3.3.5 Multilayer perceptron (MLP)

MLP is a well-developed neural-based classification technique that incorporates three or more

layers: an input layer, an output layer, and one or more hidden layers between the input and output layers [41]. The entire layer is made up of several 'neurons' that connect all of the layers. MLP is a multivariate multivariate indirect mappings calculator that produces a volume of training data to be read and constructed [2] from training data using retrospective learning methods [42]. The MLP phase structure includes enough input variables and network type specifications, data processing, and appropriate classification, network infrastructure design, success parameters specification, training algorithm specification (associated weight performance), and ultimately. [43] Model testing. In this investigation, the automated arrangement yielded the greatest results for this separator.

4. RESULTS AND DISCUSSIONS

4.1 Result of Exploratory Data Analysis (EDA)

Impact of data analysis (EDA) on this database, evaluation statistics evaluation is achieved to better apprehend the characteristics of the database. The consequences of this evaluation are defined in the following paragraph.

Figure-2 suggests the distribution of the elements of the dataset. Prices or pints without boxes and outer should ache. In the first segment, all of the outliers found are proven in this discern. Outliers were acquired through the usage of an interquartile range and extracted from the database. Facts do not display outsiders in this database after this filter is out. After casting off all outside items, a few databases are used for additional analysis.

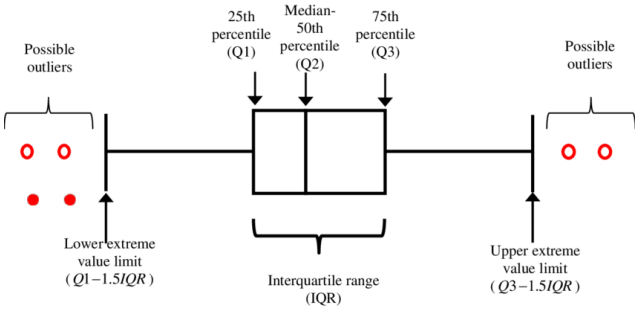

Figure-2. Box plot.

Figure-3 temperature map representing values related to the relationship between elements. The correlation between the elements and their associated values is included in all coloured cells, with the colour of the cell reflecting the capacity to communicate, while the average cost less than zero indicates a negative correlation and zero costs indicate no correlation.

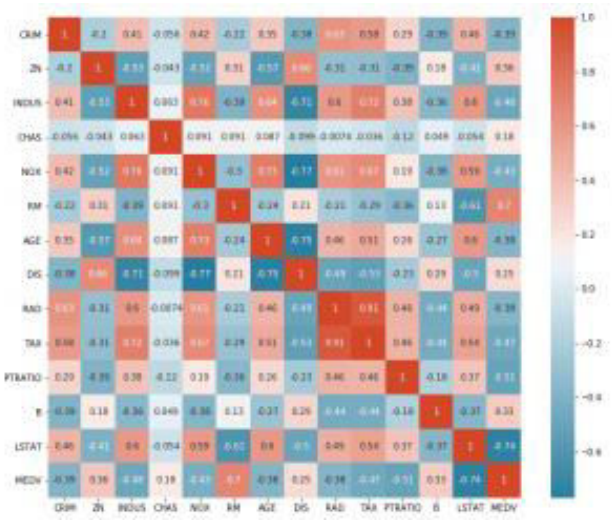

Figure-3. Using a heat map, find relationships between all of the dataset's attributes.

Figure-4 depicts the distribution of the most comprehensive database of healthy and non-sick patients. In terms of the information utilised it can be shown that patients during the 60 years are the afflicted group. As a result of the conspiracy, old age is a significant element of the heart, and the danger of illness increases with age.

4.2 Result of Machine Learning Analysis

For this study, a set of cardiovascular data was evaluated, extracted, and extracted utilising extracts from a variety of distinct classification algorithms, including MLP, KNN, DT, RF, LR, and ABM1. All of these are partition techniques.

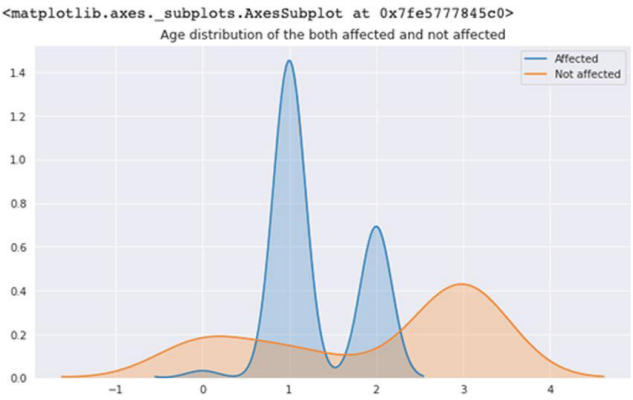

Figure-4. KDE plot for both affected and unaffected.

used for 10-fold verification techniques in the database. The cross-sectional overall performance warranty parameters are compared to determine the only set of rules for predicting the incidence of heart sickness. Parent 1 suggests the whole method.

Table-2 shows all parameters of the end-to-end performance result of the systematic planning phase used, in particular, sensitivity, readability, and accuracy. All of these suggest relevant results, with the KNN, RF, and DT parameters.

Table-2. Classification results of different classification algorithm.

Classifier Technique	Sensitivity	Specificity	Accuracy
LogisticRegression	0.800	0.950	89.610
AdaboostM1	0.900	0.970	95.210
Multilayer perceptron	0.940	0.960	97.510
K-nearest neighbor	0.960	0.970	97.610
Decision tree	0.970	0.970	97.650
Random forest	0.980	0.980	97.710

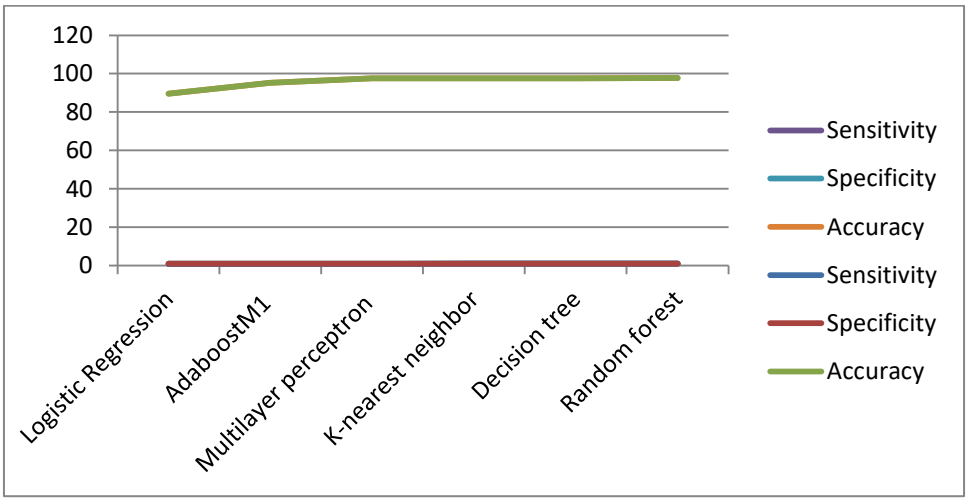

Figure-5. Precision-recall curve for different methods.

Provide high accuracy, sensitivity, and readability, observed through MLP displaying better overall staging than LR and ABM1. Figure-5 describes the Precision-Recall Curve for different methods.

Table-3 shows the kappa information and MCC values for the unique division methods. According to the results shown in the table, MLP is a better participant than ABM1 and LR. KNN, RF, and DT perform admirably and command an enormous price.

Table-3. Kappa and MCC values.

Classifier Technique	Kappa	MCC
Logistic Regression	0.720	0.760
.AdaboostM1	0.870	0.880
Multilayer perceptron	0.950	0.950
K-nearest neighbor	0.960	0.960
Decision tree	0.970	0.970
Random forest	0.980	0.980

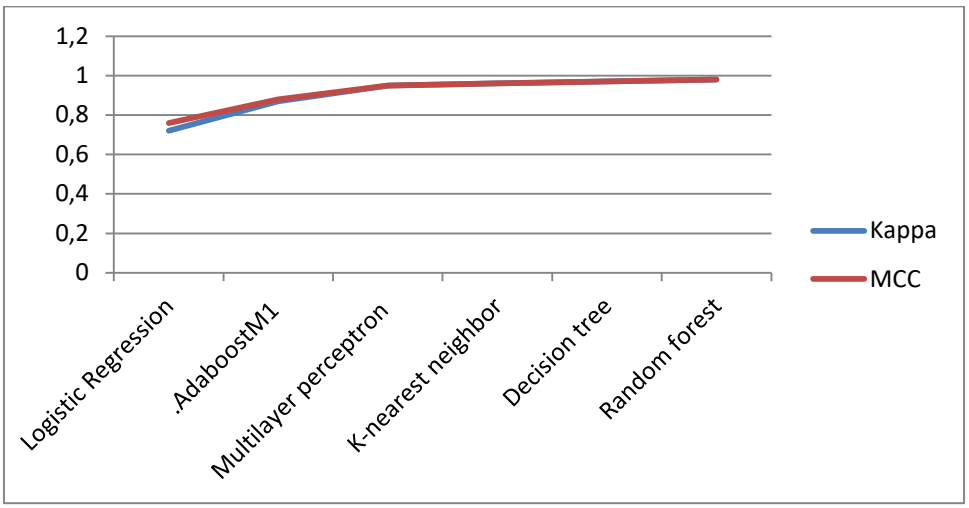

Figure-6. Graph for evaluation by Kappa and MCC.

Table-4 shows that LR and ABM1 perform worse than MLP in terms of accuracy, memory, and f steps. At the same time, KNN, RF, and DT function admirably, with 97.2 percent accuracy.

Table-4. Precision, recall, and f-measures.

Classifier Technique	Precision	Recall	F-Measure LR
AdaboostM1	0.900	0.950	0.950
Multilayer perceptron	0.980	0.980	0.980
K-nearest neighbor	0.980	0.980	0.980
Decision tree	0.960	0.960	0.960
Random forest	0.970	0.970	0.970

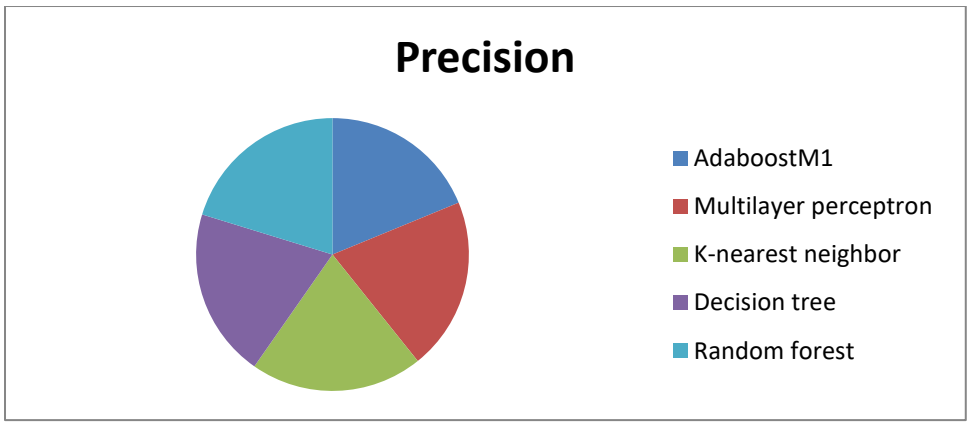

Figure-7. Pie chart for precision, recall, and f-measures.

Table-5 displays the area beneath the ROC and PRC for each subdivision algorithm. The region beneath the ROC indicates the general neighbourhood of true effective measurement and false fantastic degree, whereas the region beneath the PRC represents the general vicinity of accuracy and memory. Even though LR and MLP have a tight relationship, ABM1 implies higher effectiveness. KNN, RF, and DT, on the other hand, produced the best benefits.

Table-5. Value of area under ROC and PRC.

Classifier Technique	AUROC	AUPRC
Logistic Regression	0.910	0.910
AdaboostM1	0.990	0.990
Multilayer perceptron	0.920	0.910
K-nearest neighbor	0.950	0.950
Decision tree	0.960	0.960
Random forest	0.970	0.970

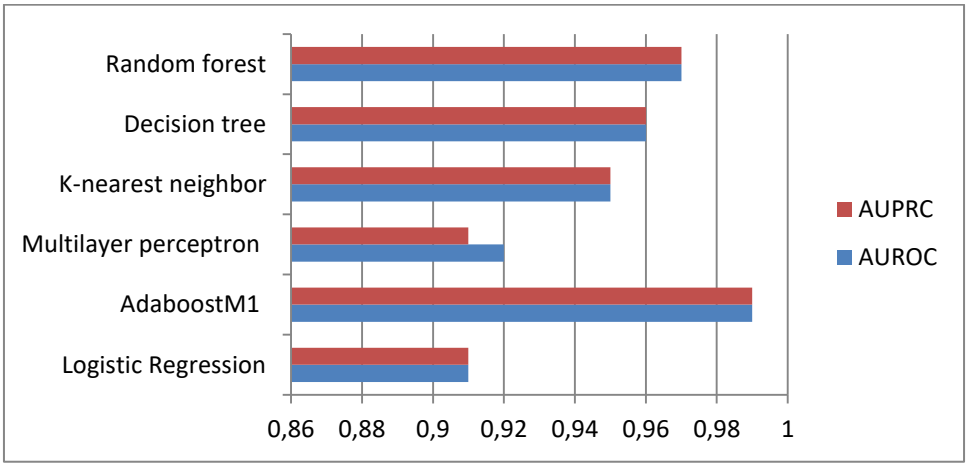

Figure-8. Bar graph for value of the area under ROC and PRC.

Table-6. Feature value and coefficient scores for the various algorithms used.

Features Name	LR	ABM1	DT	RF
Cp	0.883	0.04	0.242	0.135
Oldpeak	-0.542	0.10	0.092	0.121
Ca	-0.714	0.08	0.153	0.115
Thalach	0.0298	0.18	0.088	0.114
Thal	-0.849	0.04	0.075	0.112
Age	0.006	0.10	0.097	0.088
Chol	-0.004	0.32	0.078	0.074
Trestbps	-0.010	0.04	0.049	0.070
Exang	-0.944	0.02	0.064	0.055
Slope	0.733	0.02	0.000	0.049
Sex	-1.610	0.06	0.034	0.033
Restecg	0.399	0.00	0.023	0.019
Fbs	-0.285	0.00	0.000	0.009

Separator algorithms have been used other than MLP and KNN as those algorithms do not produce the cost of one of these feature or coefficient values. These factors the value and coefficient score are represented in the desk in keeping with the corresponding elements. The desk is proven in Figure-9 to better apprehend the extent of characteristics and importance according to the category algorithms.

Figure-9 is a diagram of Table-6. The determination suggests the level of the characteristic primarily based on the characteristic cost and the equal ratings of all the classification algorithms used except for

MLP and KNN. The number additionally regularly represents the most accountable elements of the heart ailment.

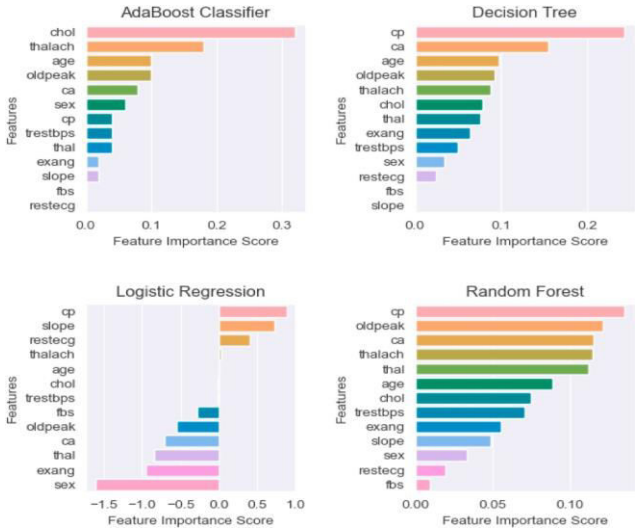

Figure-9. Bar graphs for feature importance.

Table-7 shows the five maximum crucial elements in terms of characteristic fee and dating fee. In step with the table, its miles located that chest ache (cp) is an important issue or element in coronary heart disease prognosis and diagnosis. Apart from age, the number of high coronary heart fees completed (thalach), ST pressure due to relaxation-related exercising (old peak), and the number of massive vessels (zero-3) coloured with fluoroscopy (ca) are also crucial elements affecting heart disorder.

Table-7. According to the above algorithms five features of heart disease.

Feature Ranking	LR	ABM1	DT	RF 1st
1	Slope	thalac	ca	oldpea
2	Restec	age	age	ca
3	Thalac	oldpea	oldpea	thalac
4	Age	ca	thalac	tha

In quick, we compiled a coronary coronary heart price database, processed it in advance as essential, and finished it to better understand the database. We then used six machine studying algorithms, ABM1, LR, MLP, KNN, DT, and RF, and evaluated their predictions primarily based mostly on the accuracy, sensitivity, precision, kappa records, accuracy, memory, F-diploma, and -MCC. ROC curve and Precision-keep in mind curve. We've got placed the effectiveness of all of the algorithms used, wherein KNN, DT, and RF have demonstrated notable typical performance with 100 hundred accuracy indicating that those are very effective in predicting heart ailment. We additionally predicted the function cost and coefficient values of all the algorithms used besides MLP and KNN as those algorithms no longer produce any characteristic effect of charge or coefficient values. The effect results of the function price are demonstrated in Table-6, wherein the one's factors are measured and represented by using the usage of photos in Figure-8 counting on the element fee issue. This analysis identified the most predictable components of cardiovascular diagnosis that propose potential advantages to nurses seeking to assume the incidence of coronary heart illness of their sufferers. however, it has to be cited that the quantity of facts on coronary coronary heart disorder furnished by way of this database became no longer massive sufficient to deal with all the issues and that additional records and evaluation are had to expose the producing of a sturdy predictive model but, inside the destiny, we are hoping to better recognize the limitations of this technique and that additional information analysis will permit extra accurate predictions of heart ailment and related situations the use of device gaining knowledge of techniques.

5. CONCLUSIONS

Heart illness is volatile to health, leading to probably lethal headaches which include coronary heart assaults because of its capability to expect sickness stages, the importance of records mining, and device getting-to-know techniques can be wielded to assume its occurrence. proper here, we wield a cardiovascular database to evaluate the usage of ML techniques for predicting heart disease and observed that the 3 different techniques KNN, RF, and DT labored very well with one hundred% precision. In addition, the price elements for each function are predicted by the least-bit algorithms used besides MLP and KNN. Those elements are rated primarily based absolutely on the result of the price of the characteristic. The look at aimed to locate the great ML techniques, among several properly-common and smooth-to-use algorithms, which found that, as a minimum on this database, they labored nicely. That is step one in the usage of ML strategies however suggests that it could be a top-notch addition to affected person care.

REFERENCES

[1] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Accessed 02 June 2021].

[2] R. D. Canlas. 2009. Data Mining in Healthcare: Current Applications and Issues, School of Information Systems & Management, Carnegie Mellon University, Australia.

[3] Christoph Helma, Eva Gottmann, Stefan Kramer. 2000. Knowledge discovery and data mining in toxicology, Stat. Methods Med. Res. 9(4): 329-358.

[4] I.-N. Lee, S.-C. Liao, M. Embrechts. 2000. Data mining techniques applied to medical information, Med. Inf. Internet Med. 25(2): 81-102.

[5] L. Parthiban, R. Subramanian. 2008. Intelligent heart disease prediction system using CANFIS and genetic algorithm, Int. J. Biol., Biomed. Med. Sci. 3(3).

[6] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee. 2013. Using methods from the data mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, J. Clin. Epidemiol. 66(4): 398-407.

[7] S. K. Dehkordi, H. Sajedi. 2018. Prediction of disease based on prescription using data mining methods, Health Technol. 9(1): 3744.

[8] M. Jan, A. A. Awan, M. S. Khalid, S. Nisar. 2018. Ensemble approach for developing a smart heart disease prediction system using classification algorithms, Res. Rep. Clin. Cardiol. 9: 33-45.

[9] J. Soni, U. Ansari, D. Sharma, S. Soni. 2011. Predictive data mining for medical diagnosis: an overview of heart disease prediction, Int. J. Comput. Appl. 17(8): 43-48.

[10] H. M. Islam, Y. Elgendy, R. Segal, A. A. Bavry, J. Bian. 2017. Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: a machine learning approach, J. Heart & Lung. 1-7.

[11] H. M. Le, T. D. Tran, L. A. N. G. Van Tran. 2018. Automatic heart disease prediction using feature selection and data mining technique, J. Comput. Sci. Cybern. 34(1): 33-48.

[12] M. Tarawneh, O. Embarak, February. 2019. Hybrid approach for heart disease prediction using data mining techniques, Acta Sci. Nutr. Health. 3(7): 147-151.

[13] R. Chitra, V. Seenivasagam. 2013. Heart disease prediction system using supervised learning classifier, Bonfring Int. J. Softw. Eng. Soft Comput. 3(1): 01-07.

[14] C. B. C. Latha, S. C. Jeeva. 2019. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Info. Med. Unlocked. 16: 100203.

[15] S. Mohan, C. Thirumalai, G. Srivastava. 2019. Effective heart disease prediction using hybrid machine learning techniques, IEEE Access 7: 81542-81554.

[16] https://www.kaggle.com/johnsmith88/heart-disease-dataset [Accessed 02 June 2021].

[17] M. R. Rahman, T. Islam, T. Zaman, M. Shahjaman, M. R. Karim, F. Huq, J. M. Quinn, R. D. Holsinger, E. Gov, M. A. Moni. 2019. Identification of molecular signatures and pathways to identify novel therapeutic targets in alzheimer's disease: insights from a systems biomedicine perspective, Genomics. 112(2): 1290-1299.

[18] Four Techniques for Outlier Detection, https://www.kdnuggets.com/2018/12 /four-techniques-outlier-detection.html.

[19] Md Satu, Syeda Atik, Mohammad Moni. 2019. A Novel Hybrid Machine Learning Model to Predict Diabetes Mellitus.

[20] S. Asaduzzaman, M. R. Ahmed, H. Rehana, S. Chakraborty, M. S. Islam, T. Bhuiyan. 2021. Machine learning to reveal an astute risk predictive framework for Gynecologic Cancer and its impact on women psychology: Bangladeshi perspective, BMC Bioinf. 22(1): 1-17.

[21] T. Akter, M. S. Satu, M. I. Khan, M. H. Ali, S. Uddin, P. Lio, J. M. Quinn, M. A. Moni. 2019. Machine learning-based models for early stage detection of autism spectrum disorders, IEEE Access. 7: 166509-166527.

[22] S. M. Vieira, U. Kaymak, J. M. C. Sousa. 2019. Cohen's kappa coefficient as a performance measure for feature selection, in International Conference on Fuzzy Systems, 2010 [online] Available at, http://yadda.icm.edu.pl/yadda/element/bwmeta1.element. Ieee-000005584447. (Accessed 21 August 2019).

[23] Z. Lei, Y. Sun, Y. A. Nanehkaran, S. Yang, M. S. Islam, H. Lei, D. Zhang. 2020. A novel data-driven robust framework based on machine learning and knowledge graph for disease classification, Future Generat. Comput. Syst. 102: 534-548.

[24] X. Luo, F. Lin, Y. Chen, S. Zhu, Z. Xu, Z. Huo, M. Yu, J. Peng. 2019. Coupling logistic model tree and random subspace to predict the landslide susceptibility areas with considering the uncertainty of environmental features, Sci. Rep. 9(1): 1-13.

[25] S. Uddin, A. Khan, M. E. Hossain, M. A. Moni. 2019. Comparing different supervised machine learning

algorithms for disease prediction, BMC Med. Inf. Decis. Making. 19(1): 1-16.

[26] T. Cover, P. Hart. 1967. Nearest neighbor pattern classification, IEEE Trans. Inf. Theor. 13(1): 21-27.

[27] B. V. Dasarathy. 1991. Nearest neighbor (NN) norms: NN pattern classification techniques, IEEE Comput. Soc. Tutorial. 10012834200.

[28] K. H. Raviya, B. Gajjar. 2013. Performance Evaluation of different data mining classification algorithms using WEKA, Indian J. Research. 2(1): 19-21.

[29] S. B. Kotsiantis, I. Zaharakis, P. Pintelas. 2007. Supervised machine learning: a review of classification techniques, Emerg. Artif. Intel. Appl. Comput. Eng. 160: 3-24.

[30] R. L. De Mantaras, E. Armengol. 1998. Machine learning from examples: inductive and Lazy methods, Data knowledge. Eng. 25(1-2): 99-123.

[31] S. Vijayarani, S. Sudha. 2013. Comparative analysis of classification function techniques for heart disease prediction, Int. J. Innov. Research. Compute. Commun. Eng. 1(3): 735-741.

[32] L. Breiman. 2001. Random forests, Mach. Learn. 45(1): 5-32.

[33] S. M. M. Hasan, M. A. Mamun, M. P. Uddin, M. A. Hossain. 2018. February. Comparative analysis of classification approaches for heart disease prediction, in 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE. pp. 1-4.

[34] J. R. Quinlan. 1986. Induction of decision trees, Mach. Learn. 81-106.

[35] J. A. Cruz, D. S. Wishart. 2006. Applications of machine learning in cancer prediction and prognosis, Canc. Inf. 2: 117693510600200030.

[36] K. Li, G. Zhou, J. Zhai, F. Li, M. Shao. 2019. Improved PSO_AdaBoost ensemble algorithm for imbalanced data and sensors. 19(6): 1476.

[37] C. Zhang, Y. Chen. 2017. Improved piecewise nonlinear combinatorial adaboost algorithm based on noise self-detection, Comput. Eng. 43: 163-168.

[38] D. W. Hosmer Jr., S. Lemeshow, R. X. Sturdivant. 2013. Applied Logistic Regression, vol. 398, John Wiley & Sons.

[39] S. Uddin, A. Khan, M. E. Hossain, M. A. Moni. 2019. Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inf. Decis. Making. 19(1): 1-16.

[40] S. Dreiseitl, L. Ohno-Machado. 2002. Logistic regression and artificial neural network classification models: a methodology review, J. Biomed. Inf. 35(5-6): 352-359.

[41] K. Kwon, D. Kim, H. Park. 2017. A parallel MR imaging method using multilayer perceptron, Med. Phys. 44(12): 6209-6224.

[42] S. Tajmiri, E. Azimi, M. R. Hosseini, Y. Azimi. 2020. Evolving multilayer perceptron and factorial design for modelling and optimization of dye decomposition by biosynthetized nano CdS-diatomite composite, Environ. Res. 182: 108997.

[43] Y. Azimi. 2019. Prediction of seismic wave intensity generated by bench blasting using intelligence committee machines, Int. J. Eng. 32(4): 617-627.

[44] G. Casalicchio, C. Molnar, B. Bischl. 2018. Visualizing the feature importance for black-box models, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Cham. pp. 655-670.

[45] V. A. Huynh-Thu, Y. Saeys, L. Wehenkel, P. Geurts. 2012. Statistical interpretation of machine learning-based feature importance scores for biomarker discovery, Bioinformatics. 28(13): 1766-1774.

[46] M. M. Ahamad, S. Aktar, M. Rashed-Al-Mahfuz, S. Uddin, P. Lio, H. Xu, M. A. Summers, J. M. Quinn, M. A. Moni. 2020. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients, Expert Syst. Appl. 160: 113661.

[47] https://datascience.stackexchange.com/questions/44700/how-do-i-get-the-fe ature-importace-for-a-mlpclassifier [Accessed on 01 June 2021].

[48] https://stats.stackexchange.com/questions/363662/can-you-derive-variable-im portance-from-a-nearest-neighbor-algorithm [Accessed on 01 June 2021].