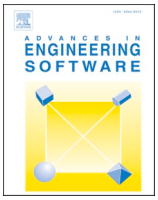
Advances in Engineering Software 173 (2022) 103285

Contents lists available at ScienceDirect

Advances in Engineering Software

journal homepage: www.elsevier.com/locate/advengsoft



Research paper

An assessment of machine learning algorithms for healthcare analysis based on improved MapReduce

J. Sukanya [a,*], K. Rajiv Gandhi [b], V. Palanisamy [c]

[a] Alagappa University, Karaikudi, Tamil Nadu, India
[b] Government Arts and Science College, Alagappa University, Paramakudi, Tamil Nadu, India
[c] Department of Computer Applications, Alagappa University, Karaikudi, Tamil Nadu, India

ARTICLE INFO

Keywords:
Heart sickness
K-Means
MapReduce
Navie-Bayes
Machine learning
PSNB: IMR

ABSTRACT

Peoples who're specially affected with heart sickness and it's far one of the man-kill illnesses in the international level. Most of researchers to awareness at the prediction, clustering, rule generation, decision tree and machine learning algorithm for figuring out and predicting the danger of the sufferers primarily based on the medical information. The overall performances of the crucial functions are based on the machine-learning concept. By studying the algorithm, the researcher can pick out the time and reminiscence wanted for the execution. As such, there are many different types of machine learning algorithms are categorized into three important classifications namely unsupervised learning, supervised learning and reinforcement learning. Unsupervised learning consists of all varieties of clustering algorithms at the same time as supervised learning algorithm consists of all of the category strategies. But the author is considered the two algorithms are Supervised and Unsupervised learning algorithm to examine the overall performance. This research paper includes the six elements to evaluate the overall performance of K-Means, Navie-Bayes and enhanced PSNB-IMR Algorithm with various parameters.

1. Introduction

The Volume of facts are expanded daily for nearly in all vicinity of software because of digitalization of the world. In particular, clinical associated facts are commonly utilized in prediction, clustering, rule generation, selection tree and device studying set of rules. However, there are greater a hit device studying algorithms are applied to discover and manner to therapy sicknesses of coronary heart associated fitness problems to the human being. The researcher enforces the device studying set of rules is called, PSNM: IMR Parallel Semi Naïve Bayes Algorithm with Improved Map Reducer .A heart specialist can acquire info on every out patient approximately one hundred instances greater regularly than every day on a normal foundation with periodic health center appointments, every so often giving the health practitioner an early be aware approximately worries that would forestall a coronary heart attack. The facts amassed via way of means of those clinical units is maximum regularly voluminous and exponentially growing; this includes rigorous and nuanced analyzes each to enhance medical selections and to direct take a look at into higher procedures, for this reason improving results.

Data evaluation and deep studying strategies are used to construct progressive equipment to assist physicians and people within side the healthcare enterprise make early-level selections on coronary heart attacks. Big Data is advanced via way of means of a growing multitude of means, together with Web clicks, net transactions, user-generated facts and social networking, and actually, genomics facts, in addition to intentionally generated content material via way of means of sensor networks or company transactions. The maximum crucial trends at gift encompass the use of human genetic facts in drug studies, the alternate of medical trial results, in our perception this paper is one a few of the critical investigations withinside the territory of locating the top-ok power esteems in considering massive facts. The use of the MapReduce method in this problem is probably a momentous association that guarantees advanced effectiveness. Contrasted with beyond works, this exam moreover has several distinctive components of development that have a few understandings in difficult to understand facts or whole facts. During this sense, endeavors recognition on lacking facts in addition to on tremendous, lacking Big Data.

The researcher contributed their studies in this research paper are lessen the facts extent measurement the usage of IMR dimension reduction algorithm to reduce the data volume of dimension and select the suitable and secure parameters to predict the disease of heart related

* Corresponding author.
E-mail address: jjsukanya1254@gmail.com (J. Sukanya).

https://doi.org/10.1016/j.advengsoft.2022.103285
Received 16 July 2022; Received in revised form 16 August 2022; Accepted 31 August 2022
Available online 21 September 2022

issues using PSNB prediction algorithm.

Different techniques had been proposed for figuring out the danger evaluation of coronary heart sufferers. In this article, we have got analyzed the performances of three machine learning algorithms like parallel K-Means clustering algorithm, MapReduce based Naive Bayes algorithm and PSNB: IMR algorithm for predicting the data of coronary heart sufferers. The records contain the following 14 characteristics like Age, Sex, Cp, trestbpd, serumCho, fbs, restecg, thalach, exang, oldpeak, peakSlope, numVessels, Thal and Disease. The records are obtained from UCI machine learning repository. The overall performance evaluation of any algorithm relies upon on different factors which include variety of iterations, variety of machines, time, area, accuracy, error rates, scalability, sensitivity, specificity and F-1 rating. We were taken into consideration the above cited elements for studying the performances of our proposed algorithms PSNB: IMR algorithms with K-Means and Navie- Bayes and.

2. Related study

There are numerous researchers do their studies within side the subject of fitness care enterprise and category [1] have proposed clustering primarily based totally category strategies for predict the sufferers of coronary heart illnesses. This paper discussed the following steps. The very beginning step is clustering that allows name of the cluster and varied form of records. The idea confers the gathered records for the estimate assessment. The dataset is calculated for resulting the centroid factor and calculate Euclidean distance from the centroid factor, and the specific object is recognized in which denotes the resemblance among the records. They are prepared to use the reduction strategies for reducing the complication of the big datasets.

The proposed cluster and reduction methods are carried out in Mat lab and expanded the correctness from the prevailing method [2] were mounted with K-Means compressed sensing concept in mixture with SVD approach. They have used a big electrocardiogram dataset which includes 668486 beats. The authors are wants to clarified proposed strategy with a conjunction of PCA as a dimensionality reduction method. The projected algorithm additionally decreased thirteen percentage clustering power intakes as compared to the present clustering algorithm. They additionally advised that the projected algorithm has many sensible packages which include wi-fi ECG structures, Holter tracking, digital fitness, and cell fitness.

The creator [3] has furnished one of kind methods to enforce the idea of machine learning and to get the quality outcomes. Machine mastering is the idea to research the large quantity of records. In beyond centuries, the most quantity of time became wasted via way of means of our programmers extended the records on their research and getting the final result. The research says more time taken for very large dataset. While the final result came, they get mistakes in attaining the favored output. The MapReduce primarily based totally gadget mastering algorithm is used to research the records in fraction of seconds.

Authors [4] have as compared the algorithms with one of a kind overall performance measure with the assist of gadget mastering. From their observations, it's been diagnosed that the quality outcomes are proven via way of means of K-Nearest Neighbor, Random Forest, Naïve Bayes and Artificial Neural Network. The proposed fashions are examined with coronary heart sickness dataset and the techniques are led to higher accuracy of 99.65%.

Also, the authors [5] are offering a MapReduce sickness tracking application for actual-time examine of the hyperlink among weather records and Dengue fever transmission. Endorse a neighbor classifier okay-nearest for disparity records discount the usage of MapReduce, making use of the technique to a extensive DNA dataset of ninety million base pairs. In some other text, Lin et al. [6] upload MapReduce to de Bruijn graphs such that metagenomic gene category is finished extra efficaciously and reliably. Research has additionally been achieved to increase the MapReduce gadget and its assisting Apache Hadoop software program to be used in huge records; Weng et al. [5] indicates upgrades to metadata renovation within side the Hadoop Distributed File System for expanded scalability, displaying MapReduce persevering with significance in present day packages.

3. Proposed system methodoligies

Cloud computing is the maximum crucial device that is used to accumulate and retrieve the affected person's records at the call for basis. In this approach, the uncooked records of coronary heart sufferers had been accumulated from cleveland. Records [7], hungarian.records, lengthy-beach-va.records and switzerland.records. Next, pre-processing became final stage for improve the accuracy of proposed algorithms. The dataset is ten thousand bytes in length, with 20% of the records utilized for training and the final 80% for testing [8]. In this approach, clustering primarily based totally category became implemented to are expecting the danger evaluation of coronary heart sufferers [9]. The records partitioning is the maximum crucial operations within side the disbursed environment. Dataset became partitioned into extra variety of sub parts. The scoring method is one of the quality techniques to decide the dominant values many of the given dataset. Weights had been assigned for the ones subparts [10]. The parallel clustering became finished at Hadoop. The output of parallel clustering became feed into classification algorithm. The changed conjunctive attributes had been extracted from the outcomes.

The most appropriate solution will be to measure the score for each particular dimension, depending on the intrinsic characteristics of MapReduce. The MapReduce mapper part measures the ranking, since it only accommodates a set number of dimensions. There are also three internal sets which form the IMR algorithm and help us acquire the dominant top-k values for each object m(Case) in a dataset. The key purpose of the IMR algorithm is to aid by taking advantage of the MapReduce system to make score estimation of each user's features as simple as possible.

The scoring system is a crucial aspect which can greatly influence efficiency of the algorithms. The score is a metric in this sense which indicates how strong an object must be to be a top-k dominant rating. The enhanced PSNB algorithm is considered to develop accuracy by taking advantage of both Bayesian network algorithm and Apache Spark parallelism, and to reduce execution time. The first stage of the PSNB algorithm is generated by dependent and autonomous Direct Acyclic Graph (DAG) based on the Bayesian control network. Since the program includes several influential parameters, an important DAG is developed. Likewise, all other DAGs are created on the basis of dependent, independent parameters. In the next point, the DAGs for parallel execution are allocated to the Spark RDD objects. In the next step, in the Transformation: flat map function of the PSNB algorithm, the mean ($\mu_{\varnothing}$), variance ($\sigma^2_{\varnothing}$) and conditional likelihood ($\wp$) are performed. In the corresponding step of the PSNB algorithm, the class possibility is calculated by the Transformation: Hadoop map function by using previous outputs of $\mu_{\varnothing\sigma^2_{\varnothing}\ and\ \wp}$. Ultimately, prediction output is stored in the PSNB algorithm using actions and save feature for the users. For each priority user, all flat map and map jobs are done on correct Hadoop cluster nodes depending on the IMR algorithm.

4. Performance assessment of machine learning algorithm

4.1. PSNB-IMR algorithm

The probabilistic approach enhanced Parallel Semi-Naive Bayes (PSNB) algorithm is implemented to categorize the coronary heart sufferers primarily based totally at the medical signs of sufferers which had been recorded via way of means of BAN (Body Area Network). The IMR algorithm became carried out to finish the MapReduce system [11]. The mapper additives check all of the attributes and assigns rank to all

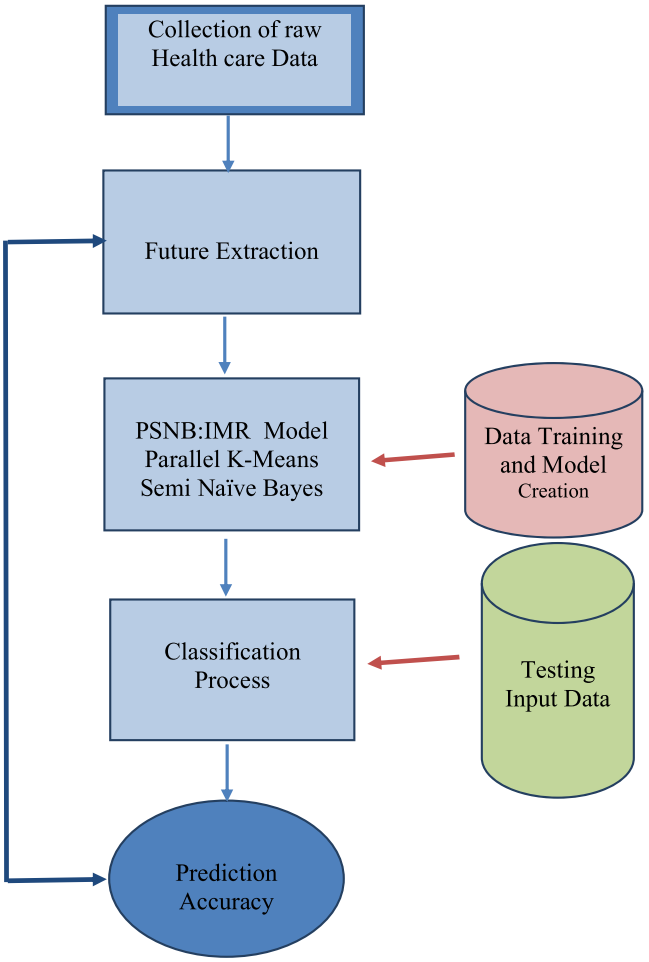


Fig. 1. Overall process of proposed architecture.

attributes. The identity of predictive accuracy of the proposed algorithm PSNB: IMR, assessments had been carried out and as compared with the ultra-modern algorithms. The records units for coronary heart sickness are taken as remarks. Fig. 1 illustrates the proposed frame work for the step-by-step procedure to predict the model and analyze. A disseminated framework that is adaptable is introduced for the expectation of information which brought about expanded execution with least time intricacy. Thinking about the unstructured information, the proposed PSNB: IMR is planned as adaptable learning framework for administered learning. The framework that is prepared well actually want to anticipate the yield which is positively not kn using the Bayesian classification algorithm perspective, the probability $p(C|D)$ of a understood sample data D consider class C can be specified as follows:

$$p(C|D) = \frac{p(C)p(D|C)}{p(D)} \tag{1}$$

Let C is the class label and it is maximized probability, it considers as max-a-posteriori classification (MAP). Let $p(C)$ is general for all classes and also let $p(D)$ is equivalent to all classes. The max-a-posteriori classification may reduction to max-likelihood classification.

$$\widehat{C} = argmax\ p\left(C|_c^D\right) = argmax\left(_c^D|C\right) \tag{2}$$

Let d_1, d_2, ..., d_n are the attributes of D, based on the Naïve Bayes assumption are given its class label C, we can compute the possibility as given below.

$$p(D|C) = p\left(d_1, d_{2,\cdots,}d_n|C\right) = p(d_i\ |C)_{i=1} \tag{3}$$

By placing the log probability,

$$\widehat{C} = argmax\ log\ p(D|C) = \text{arg}max_c \sum_{i=1}^{n} logp(d_i|C) \tag{4}$$

The new projected algorithm of enhanced Parallel Semi-Naive Bayes classification as given below

1 Given test data T_i and training data R
2 $s \leftarrow$ Generate m feature subsets randomly from original feature space
3 for all training data R_k,$k = 1, 2, \cdots, n$ do
4 // The for loop are mentioned for parallel processing of the given dataset
5 for all feature subsets $S_j \in S$, $j = 1, 2, \cdots, m$ do
  a $t_j \leftarrow pointofthesubspaceS_jfromsampledataT_i$
  b $r_j \leftarrow pointofthesubspaceS_jfromtrainingdataR_k$
  c $ifdist(t_j,r_j) \leq threshold(r_j)$// Determin the matching result $i.total_j[C]$ $total_j[C] + 1$// C is class label of a matched point r_j
  d end ifend for
6 end for [11]
7 return Final counting result array $total_j$[11]

4.2. Parallel K-Means algorithm

In this proposed method, MapReduce based parallel K-Means clustering algorithm was implemented for the prediction of risk analysis of heart patients. In this work, Hadoop was used to execute the Map Reduce process which was present in an AWS. For mapping, the data was converted into key and value pairs. The shuffling was done at the key and value pairs. The dataset was divided into subparts which were given to the different set of machines. The proposed method also tested with the dataset which have 14 attributes and 303 samples. The preprocessing step was implemented for the given attributes. After that, the dataset was feed into the feature selection task for identifying the most important attributes. In the feature selection task, chi-squared based test and correlation matrix with heat map was applied. Next step, only 4 features namely cp, restecg, thalach and slope were selected as important features. The parallel K-Means clustering algorithm was applied to the dataset with selected attributes. Two different clusters were formed and represented in different colours with their centroids.

4.3. Semi Naïve Bayes algorithm

In general, classification algorithm classifies the data into predetermined class labels. The proposed method also classifies the risk analysis of heart patients. The proposed method involves three segments. At first, significant attributes were selected by Relief feature selection method. Then MapReduce implementation was done by Spark. At last, Classification of heart patients was done by using semi naive bayes classification algorithm.

The dataset consists of 13 attributes. The Relief feature selection was used to identify important attributes that causes of heart diseases. The 10 nearest hits and nearest misses were identified for finding the rank of all the attributes. The ranks were assigned to all the attributes. Among 13 attributes, only top 5 attributes were selected for the classification process. They were cp, exang, chol, thal and slope as they have highest score among all the attributes. The MapReduce concept was implemented in 10 machines. For mapping process, each machine was assigned 76 instances as input with top 5 attributes. The proposed method was executed at 10 different machines with different configurations. The semi naive bayes algorithm was implemented in 10 machines for parallel execution. The classification results were obtained from the 10 machines. The underlying test when the proposed calculation is executed with pre-handling of information which incorporates methods like standardization, missing worth taking care of, information change, extraction of elements and information preparing. The fundamental idea of information investigation is to extricate valuable data

from information by recognizing every one of the potential relations among the information.

5. Result analysis and discussions

5.1. Parameters for comparative analysis

5.1.1. Time and space complexity

A question may be solved fashionable a difference of habit. One endures learn in what way or manner to determine the act of several algorithms and select high-quality individual for a likely question. The authors are mainly captured into the record of finances the following limit one exist temporal length of event or entity's existence complicatedness and second one happen room complicatedness all along the assessment of the reasoning. The time complicatedness of an invention measures how long an treasure takes to run as a function of the distance of the recommendation. Similarly, an treasure's space complicatedness measures in what way or manner much scope or thought it takes to run as a function of the length of the recommendation. Many determinant influence opportunity and space complicatedness, containing fittings, computer software for basic operation, processors, and so on. However, nobody of these traits happens taken into or charges while determine the invention. We exist focusing an invention's killing occasion.

5.1.2. Sensitivity and specificity

The sensitivity of the test cases is predicting true positives in each accessible category. The specificity of the test cases is measuring a model's ability to predict true negatives in each accessible category. Any category model can benefit from these measures. Below the Eqs. (5) and (6) are the formulae for computing these measures.

$$Sensitivity = \frac{True\ Positive}{True\ Positives + False\ Negatives} \tag{5}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \tag{6}$$

These determinants happen used to analyze the efficiency of categorization system. True Positive (TP), action of person being treated for medical problem that the projected system expresses an outcome in advance happen positive that exist existent certain. False Negative (FN), the numbers of person being treated for medical problem that the projected system express an outcome in advance exist negative that exist actually beneficial. False Positive (FP), process of patients that the projected plan expresses an outcome in advance were certain that exist actually negative. True Negative (TN), program of person being treated for medical problem that the projected system expresses an outcome in advance exist negative that were existent negative.

5.1.3. Accuracy

Predicting a class label from instances in a problem domain is called classification predictive modelling. The most frequent classification accuracy statistics are used to assess a classification prediction model's performance. The beginning fashionable improving categorization precision or correctness search out create a forecast for each sample fashionable a test dataset utilizing a categorization model. The express an outcome in advance labels are therefore distinguished to the popular labels for the test set examples. The balance between parts of whole of instance fashionable the test set that happen properly express an outcome in advance, detached by all declaration made in advance created on the test set, happen used to decide precision or correctness.

The difference between correctly classified instances and total number of instances are referred as Accuracy. This measurement is considered as the most important parameter to analyze the performances of the proposed system.

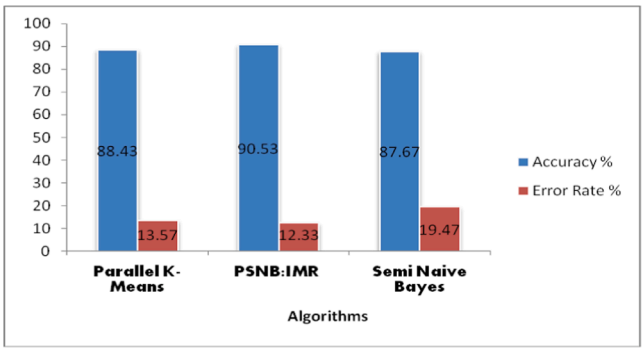


Fig. 2. Obtained accuracies and error rate.

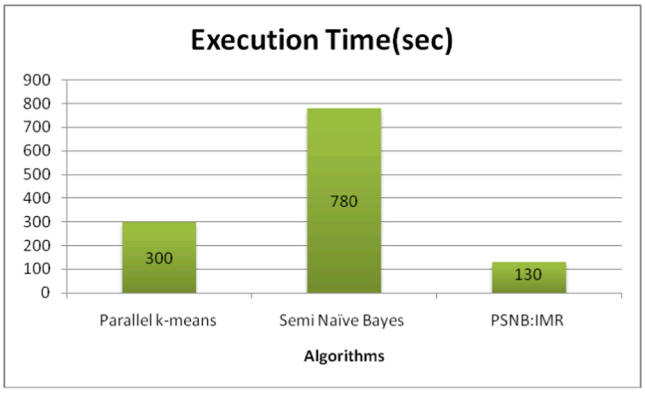


Fig. 3. Execution time.

5.1.4. Error rates

Conversely, the wrong rate maybe thought-out as the total number of wrong predictions fashioned in contact the test set detached by all declaration made in advance created ahead of the test set. The precision or correctness and error rate exist complements of each one, message that we can continually compute or estimate amount individual from the added.

5.1.5. Scalability

The number of processors required by an algorithm may be determined by the size of the problem. If a model is scalable, it can reduce the number of processors specified by an algorithm to fit an available at the cost of increased time while keeping the same efficiency. If the increasing of data, does not affect the performances of the system thus leads to the term scalability.

5.1.6. F1-score

If we calculate the difference between specificity and sensitivity and there is an unequal class distribution, F1 Score could be a preferable statistic to employ (large number of Actual Negatives). We have been also considering F1 score as the one of most important factor for analyzing the performances of proposed system.

5.2. Result analysis

The parallel K-Means method clustering algorithm affords accuracy as 88.43%, Semi Naive Bayes category algorithm affords accuracy as 87.67% as an imply accuracy from the 25 machines and the PSNB: IMR algorithm affords 90.53% as a median accuracy from the ten nodes. Similarly, The parallel K Means-method clustering algorithm affords mistakes charge as 13.57%, Semi Naive Bayes category algorithm affords Error rate as 19.4731and the PSNB: IMR algorithm affords 12.33% as an Error rate. From the observations, it's been diagnosed that the

Table 1
Summary of performance analysis of all the proposed systems.

Algorithms	Accuracy	Error rates	Execution time	No of Systems connected	Scalability	Extendibility
Parallel K-Means Clustering	88.43	13.57	300 s	10	✓	✓
Semi Naïve Bayes	87.67	19.47	780 s	10	✓	✓
PSNB: IMR algorithm	90.53	12.33	130 s	10	✓	✓

PSNB: IMR algorithm affords excessive accuracy and occasional mistakes charge as it's far as compared with different proposed techniques. Fig. 2. Shows that the accuracies and error rates of all of the proposed techniques.

The parallel K Means method clustering algorithm, Semi Naive Bayes category algorithm and the PSNB: IMR algorithm takes 300 s, 780 s and 130 s, respectively, to attain the favored outcomes. From the observations, it's been proved that the PSNB: IMR algorithm takes much less quantity of time while it's far as compared with different proposed techniques.

Fig. 3 suggests that the execution instances of all proposed techniques. Algorithm Execution Time parallel K-Means method clustering algorithm 300 s Semi Naïve Bayes category algorithm 780 s PSNB: IMR algorithm a 130 s From the observations, it has been proved that PSNB: IMR algorithm and Parallel K-Means algorithm works properly in extra variety of structures with much less quantity of time while it's far as compared with different proposed techniques. Table 1. Shows that overall assessment of proposed algorithms.

6. Conclusion and future scope

All the assessment techniques for machine learning are scalable and extendable with MapReduce system. This paper introduces three assessment machine learning algorithms are Parallel K-Means Clustering, Semi Navie Bayes and PSNB: IMR then compares existing algorithms Parallel K-Means Clustering and Semi Navie Bayes with the proposed algorithm of PSNB: IMR by analyzing accuracy, error rate, execution time, no. of systems used, scalability and extensibility. Besides, it summarizes the application range of different algorithms. As in keeping with evaluation, Parallel K-Means algorithm and proposed PNSB: IMR are appropriate for prediction of heart disease. As in keeping with the scalability, extendibility, time and area complexity, the proposed PNSB: IMR algorithm is the quality one for predicting the heart diseases. The proposed PSNB: IMR algorithm is a faster way of processing massive datasets while at the same time wisely controlling the computer resources and retaining time efficiency. The proposed PSNB: IMR algorithm offers excellent efficiency, and is two or three times faster than the single computer process in most situations. For evolving big data, the proposed PSNB: IMR can be well used to develop classification systems and provide Top K dominant database processing with a nearly real-time solution. The suggested model can be easily used explicitly for the collection and interpretation of emergency patient data in various medical applications. Such ideas and changes may be discussed later either to further enhance the efficiency or to include stronger, needs-based classification schemes. In future, we can add more number of nodes, input as a streaming data in cloud applications, and implement various index parameters to our model to show the accuracy of results.

Author statement

Not applicable.

Declaration of Competing Interest

The authors declare that we have no conflict of interest.

References

[1] Singh R, Rajesh E. Prediction of heart diseases by clustering and classification techniques. Int J Comput Sci Eng 2019;7(5):861–6.
[2] Bolouchestani M, Krishnan S. Advanced K-Means clustering algorithm for large ECG datasets based o a collaboration of compressed sensing theory and k-svd approach. Signal, image and video processing, 10. Springer; 2016. Vol.
[3] Pooja, Sharma A, Sharma A. Machine learning: "a review of techniques of machine learning. J Appl Sci Comput 2018;5(7):1076–5131. VolumeJuly 2018.ISSN NO.
[4] Khourdifi Y, Bahaji M. Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. Int J Intell Eng Syst 2019;12(1):242–52. 2019.
[5] Weng CH, Huang TCK, Han RP. Disease prediction with different types of neural network classifiers. Telemat Inform 2016;33(2):277–92.
[6] Lin B, Guo W, Xiong N, Chen G, Vasilakos AV, Zhang H. A pretreatment workflow scheduling approach for big data applications in multicloud environments. IEEE Trans Netw Serv Manag 2016;13(3):581–94.
[7] Mirmozaffari M, Alinezhad A, Gilanpour A. Data mining apriori algorithm for heart disease prediction. Int J Comput Commun Instrum Eng (IJCCIE) 2017;4(1):20–3.
[8] Zaheilas N, Kalogeraki V. Real-time scheduling of skewed MapReduce jobs in heterogeneous environments. In: Proceedings of the 11th international conference on autonomic computing (ICAC 14). USENIX Association; 2014. p. 189–200.
[9] Deza E, Deza MM. Encyclopedia of distances. Springer; 2009. p. 94. page.
[10] Reuther CB, Arcand W, Bestor D, Bergeron B, Hubbell M, Jones M, Michaleas P, Prout A, Rosa A, Kepner J. Scalable system scheduling for HPC and big data. J Parallel Distrib Comput 2018;111:76–92.
[11] Choi SW, Ho Lee C. A FPGAbased parallel semi-Naive Bayes classifier implementation. IEICE Electron Express 2013;10(19):1–7.