Cluster Computing
https://doi.org/10.1007/s10586-017-1532-x

# Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster

**N. Yuvaraj**[1] · **K. R. SriPreethaa**[1]

Received: 7 December 2017 / Accepted: 13 December 2017


**Abstract**
Health care systems are merely designed to meet the needs of increasing population globally. People around the globe are affected with different types of deadliest diseases. Among the different types of commonly existing diseases, diabetes is a major cause of blindness, kidney failure, heart attacks, etc. Health care monitoring systems for different diseases and symptoms are available all around the world. The rapid development in the fields of Information and Communication Technologies made remarkable improvements in health care systems. Various Machine Learning algorithms are proposed which automates the working model of health care systems and enhances the accuracy of disease prediction. Hadoop cluster based distributed computing framework supports in efficient processing and storing of extremely large datasets in cloud environment. This work proposes the novel implementation of machine learning algorithms in hadoop based clusters for diabetes prediction. The results show that the machine learning algorithms can able to produce highly accurate diabetes predictive healthcare systems. Pima Indians Diabetes Database from National Institute of Diabetes and Digestive Diseases is used to evaluate the working of algorithm.

**Keywords** Healthcare systems · Machine learning algorithms · Hadoop clusters · Predictive analysis · Cluster computing

## 1 Introduction

Health is currently on the international agenda. Health care systems are designed to meet the health care needs of people. As population continues to grow rapidly and age hence there will be increasing demand for acute exacerbation of chronic illness and may routine [1]. Health problems are the one that nevertheless requires a prompt action. Health services aims at contributing to the improved health or diagnosis, treatment and rehabilitation of sick people. The several perspectives of health services are (i) as actions to organize the inputs necessary for the provision of effective interventions (ii) as inclusive of promotion, prevention, cure, rehabilitation and palliation efforts, and (iii) as oriented towards either individuals or populations [2].

Sound information plays a critical role in the modern health care and efficiency of health systems [3]. The use of health information lies at the root of evidence-based policy and management in health care [4]. Information and communication technologies are being utilized to improve health care systems in developing countries through the standardization of health information, computer-aided diagnosis and treatment monitoring and informing population groups on health and treatment. An automated hospital information system helps to improve quality of care because of their far-reaching capabilities.

Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar. Hyperglycemia, or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels. In 2014, 8.5% of adults aged 18 years and older had diabetes [5]. In 2015, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths. In 2015, United States estimated the number and percentage of diagnosed and undiagnosed diabetes among

✉ N. Yuvaraj
drnyuvaraj@gmail.com

K. R. SriPreethaa
sripreethaakr@gmail.com

[1] Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India

Published online: 27 December 2017

adults aged ≥ 18 years where the number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014. The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014 [5]. Diabetes prevalence has been rising more rapidly in middle and low income countries. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. In 2015, an estimated 1.6 million deaths were directly caused by diabetes [5]. Another 2.2 million deaths were attributable to high blood glucose in 2012. Almost half of all deaths attributable to high blood glucose occur before the age of 70 years. World Health Organization projects that diabetes will be the seventh leading cause of death in 2030. Accurate and efficient diabetes predictive system is needed to overcome all the deadly statistics mentioned [6].

A Hadoop cluster is a special type of cluster that is specifically designed for storing and analyzing huge amounts of unstructured data [7]. A Hadoop cluster is a computational cluster that distributes the workload among multiple cluster nodes that works to process the data in parallel. The primary benefit of using hadoop clusters is the one, ideally suited for analyzing big data [8]. Since the data collected from a huge set of patients to make a predictive analysis is obviously going to be a big data. A set of machine learning algorithms are available to make the predictive analysis system [9]. Among the different algorithms decision tree, naïve Bayes and random forest algorithms are identified as most efficient, scalable and powerful in handling the huge volume of data.

The organization of this paper is as follows: The next Sect. 2 discusses the work related to healthcare systems. In Sect. 3, the proposed model is presented followed by the evaluation measures in Sect. 4. Finally, the paper is concluded with future work in Sect. 5.

## 2 Related works

Machine learning is a well-established research area of computer science and it is playing a main role in the development of the classification and the predictive analysis systems [10]. Minimizing the total number of features reduces the total time taken for feature selection [11]. Feature fusion and multiple dictionary models can be implemented for higher accuracy [12]. When Euclidean distance between the classifications response vectors are used as the similarity measure it enhance the accuracy of classification [13]. Data mining algorithms performs better in development of intelligent computing systems for making better decisions [9,14]. Some of the familiar machine learning algorithms were perceptrons, decision tree learners like ID3 and CART, Naïve Bayes classifier and random forest algorithms [9]. Predictive data mining methods are supervised and it focuses to induce models or theories from class-labeled data [15]. The induced models can be used for prediction and classification [16]. Maximum likelihood model is proposed to overcome the difficulties in complex performance measure that occur in the predictive models [17]. Local mean based classification can be implemented to overcome the complexities that occur in large sparse databases [18].

The National Diabetes Statistics Report is a periodic publication of the Centers for Disease Control and Prevention (CDC) that provides updated statistics about diabetes in the United States for a scientific audience [5]. It includes information on prevalence and incidence of diabetes, prediabetes, risk factors for complications, acute and long-term complications, deaths, and costs [6]. An estimated 33.9% of U.S. adults aged 18 years or older (84.1 million people) had prediabetes in 2015, based on their fasting glucose or A1C level. Nearly half (48.3%) of adults aged 65 years or older had prediabetes. Diabetes was the seventh leading cause of death in the United States in 2015 [5]. Secured wireless body area network supports a lot in developing predictive analysis system in health care [19]. Big data is put forward for the first time in 2009, and then have found application in the field of multiple business and development, especially the mature usage on medical field [8]. In big data environment, users put forward higher request to the storage service on the availability, reliability and durability of data [20]. Hierarchical model based hadoop clusters is efficient in handling huge volume of data [10]. The platform is based on Hadoop big data platform architecture, relying on HDFS, Map Reduce and MongoDB, etc. distributed framework which are deployed into more cheap hardware equipment, for the application with high throughput data access mechanism [21,22]. Map Reduce based distributed computing framework works efficiently with machine learning algorithms in training the distributed data blocks [23]. The single cluster and analytical hierarchy process is used to compute data along with Hadoop to provide fault tolerance over failures, less processing time and communication errors [24].

With the advent of social media, Internet of Things and other data sources handling the huge volume of data remains a challenging task. From the literature it is clear that an intelligent cloud based cluster model is effective and efficient for managing the huge volume of data [25]. Neural network and genetic algorithms perform better for mining the hidden patterns [26]. Due to the availability of huge volume of data from different data sources distributed computing based cluster along with cloud environment supports in effective handling of data [27].

From the literature it is clear that the disease detection and surveillance systems provide epidemiologic intelligence that allows all the persons in the healthcare to deploy preventive measures and help clinic and hospital administrators make automatic and intelligent decisions.

## 3 Methodology

The dataset used, type of feature selection technique used, machine learning algorithm implemented and the ways of integrating R into hadoop is discussed in the following sections.

### 3.1 Dataset

Creating a predictive analysis system ensures a comprehensive data-set and corpus to learn from and a test data-set to ensure that the machine learning algorithms accuracy meets anticipated standards. Accuracy of the classification algorithms is highly sensitive and plays a main role in the health care systems. To develop a Diabetes prediction model in health care system the data collected by National Institute of Diabetes from 75,664 patients is used. The dataset consists of thirteen attributes that defines the various physical composition of the patients. Among the thirteen different attributes some of the attributes like Plasma glucose concentration, Diastolic blood pressure, Triceps skin fold thickness, 2-hour serum insulin, Body mass index, Diabetes pedigree function and Age in years plays a main role in prediction. Among the set of available data 70% of the data is used as a training data and the remaining 30% of the data is used for testing purpose.

### 3.2 Feature selection

The feature vectors generated from the available dataset is diversified and high due to the availability of large number of attributes about a patient. All the available attributes about the patient will not contribute for the disease prediction. Including the unwanted features for prediction is a time consuming process and its affects the accuracy of the system. Different statistical techniques are used in the literature for feature selection but have been found to be suboptimal. Information gain used as a feature selection method in the literature proves that it improves the accuracy of feature selection.

### 3.3 Information gain

Information gain (IG) is a feature selection method by using this noise is reduced and irrelevant features influence a classifier. IG measures information in bits about class prediction, if information is available in the presence of a feature and corresponding class distribution. Information gain measure chooses a test attribute at each node in a tree. An attribute with highest information gain is a current node's test attribute, which lowers the information to classify samples in partitions. Entropy is used to measures the disorder amount or system uncertainty. Higher entropy corresponds to a sample having mixed labels of collection in a classification setting. Lower entropy corresponds to pure partitions. Entropy of a sample D is defined in information theory as follows:

$$\mathrm{H(D)} = -\sum_{i=1}^{k} \mathrm{P(ci \mid D)} \log 2\ \mathrm{P(ci|D)} \tag{1}$$

where P(ci | D) is data point probability in D being labeled with class c i, and k is number of classes.

P(ci | D) is estimated directly from data as follows:

$$p\,(ci \mid D) = \frac{|\{\mathrm{xj} \in \mathrm{D} \mid \mathrm{xj\ has\ label\ yj} = \mathrm{ci}\}|}{|\mathrm{D}|} \tag{2}$$

Also a decision/split's weighted entropy is defined as follows:

$$\mathrm{H(DL, DR)} = \frac{|DL|\,H(D\,L)}{|D\}} + \frac{|\mathrm{D\,R}|}{|D\}} H(D\,R) \tag{3}$$

where D is partitioned into DL and DR due to a split decision. Finally, information gain for a split is defined as:

$$\mathrm{Gain\ (D,\ D_L, D_R)} = \mathrm{H\,(D)} - \mathrm{H(D_L, D_R)} \tag{4}$$

Gain is expected reduction in entropy due to knowing an attribute's value.

### 3.4 Machine learning algorithms

A machine learning algorithm plays a main role in developing Intelligent automation systems. There is a bundle of machine learning algorithms that are available which works on different logical entities for developing the intelligent predictive systems. The following section discusses the most commonly used machine learning algorithms.

#### 3.4.1 Neural network

Neural networks are models of biological neural networks working in a similar fashion in terms of passing the information, processing the available data and making decisions. Neural network is a set of connected input and output unit in which each connection has a weight associated with it. Neural Network are not competitor to conventional computing methods rather it should be seen as complementary as the most successful machine learning algorithms which operate in conjunction with existing, traditional techniques. Back propagation is a type of neural network learning algorithm. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural Network are capable of learn from examples and these system adapts itself during a training period, based on the existing examples of similar problems without a desired solution to each problem. After sufficient

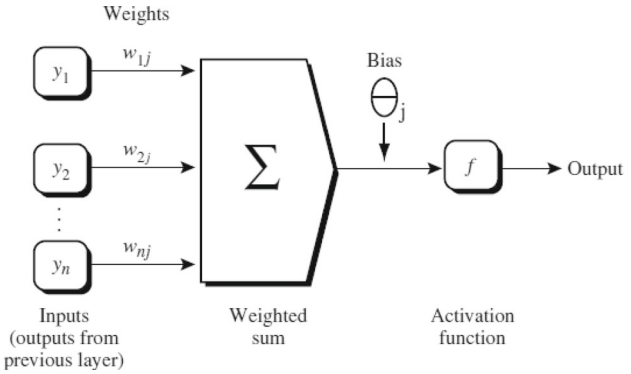


**Fig. 1** Neural network system

training the neural computer is able to relate the problem data to the solutions, inputs to outputs, and it is then able to offer a viable solution to a brand new problem (Fig. 1).

Given a unit j in a hidden or output layer, the net input, Ij, to unit j is given as follows

$$I_j = \sum_i w_{ij} O_i + \theta_j \tag{5}$$

where, $w_{ij}$ is the weight of the connection from unit i in the previous layer to unit j, $O_i$ is the output of unit i from the previous layer. $\theta$ j is the bias of the unit.

Given the net input Ij to unit j, then O j, the output of unit j, is computed as

$$O_j = \frac{1}{1 + e^{-1} j} \tag{6}$$

where, $O_i$ is the output from the unit j, Ij is the input to the unit j.

### 3.4.2 Support vector machine

Support vector machine is a type of machine learning algorithm that attempts to find a hyper-plane separating the different classes of the training instances with the maximum error margin. In this method an attempt is made to form different hyper plane for making the classification and the best separating hyper plane is calculated by measuring the distance function. Support vector machine is like trying to create a fence between two different classes and letting the few tuples in the data set to be at one side of the fence and the remaining at the other end. Support vector machine are called as large margin classifiers since they create a large margin between data points and the decision boundary. The most important training instances are the ones that are making up the boundary. These models are much complex than distance based classifiers but this will take number of parameters as input while making the classification which results in improving the overall accuracy of the classifiers.

A separating hyper plane in support vector machine can be written as

$$\mathrm{W} \cdot \mathrm{X} + \mathrm{b} = 0 \tag{7}$$

where, W is a weight vector, namely, W = {w1, w2, . . ., wn}, n is the number of attributes, b is a scalar referred to as a bias, X denotes the value of the attributes.

### 3.4.3 Decision tree

A decision tree is a machine learning algorithm that follows a flowchart-like tree structure, where each internal node or a non leaf node denotes a test on an attribute. Each branch in the tree is used to represent the outcome of the test. Finally the leaf node of the tree holds a class label which denotes the predicted final value. There are different set of decision tree algorithms in existence in which ID3 is one the most efficient decision tree algorithm that is used in most of the applications. The learning and classification steps of decision tree induction are simple and fast. The main goal of this algorithm is to construct an optimal decision tree based on a specified target function. Since decision tree algorithm is non-parametric in nature it can efficiently deal with large, complicated datasets without imposing a complicated parametric structure.

The available dataset is divided in to training and validation datasets to handle with the available large amount of data. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model. While building up a decision tree one of the attribute is selected to act as a root node for tree construction. There are various attribute selection measures are available. In this work the information gain is used because of its high efficiency in classifying the attributes based on the information gain value. The accuracy of the tree constructed based on the results obtained by giving the available dataset as an input.

### 3.4.4 Naive Bayes

Naive Bayes (NB) is a classifier that builds probability based model which works based on Bayes Theorem. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. The conditional independence of NB classifier makes the data to train faster. It assumes all the vectors in the feature vectors as independent and applies the Bayes rule in the sentence.

The basic formulation and working of Bayes theorem is explained as follows. Let $\boldsymbol{X}$ be a data tuple holding a set of values available in the data set. In Bayesian terms, $\boldsymbol{X}$ is considered "evidence." X is described by measurements made on a set of $n$ attributes. Let Hbe some hypothesis, such as that the data tuple $\boldsymbol{X}$ belongs to a specified class $C$. Tuples of different classes will have different hypothesis. For classification problems, there is a need to understand two different probabilities namely posterior probability and prior probability. P(H | $\boldsymbol{X}$) and P(X | H) is the posterior probability. P(H) is called as prior probability. The posterior probability depends on more information when compared with the prior probability. The Bayes theorem and its probability terminologies is represented in Eq. (5).

$$P(H|\boldsymbol{X}) = \frac{P(H|\boldsymbol{X})P(H)}{P(\boldsymbol{X})} \tag{8}$$

where, P(H | X) is the posterior probability of H conditioned on X, P(X | H) is the posterior probability of X conditioned on H, P(H) is the prior probability of the hypothesis H, P(X) is the prior probability of the evidence X.

Bayes theorem describes the probability of an event, based on conditions explained as posterior and prior probability that might be related to the event. Michal Haindl [28] has discussed that despite all the complicated mathematics, implementing a Bayes classifier is all about counting the number of words, documents and categories. Once the number of positive and negative words in a sentence is evaluated, then it can be combined to calculate the probability for each of the possible classes. The document is then classified according to the highest calculated probability.

#### 3.4.5 Random forest

Random forests or random decision forests are ensemble learning method for classification, regression that operates by constructing a multi node of decision trees at training time and the mode of the classes or mean prediction of the individual trees. Habit of over fitting the training set is corrected by Random format. Random forest algorithm is a supervised classification algorithm. In general, the more trees in the forest the more robustness the forest looks like. In the same way the random forest classifier, the higher the number of trees in the forest gives the higher accuracy results. The pseudocode for random forest algorithm can split into two stages. One is Random forest creation pseudocode and the other is Pseudocode to perform prediction from the created random forest classifier. Random Forest pseudocode is as follows:

1. Randomly select "k" features from total "m" features. Where k << m
2. Among the "k" features, calculate the node "d" using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1–3 steps until "l" number of nodes has been reached.
5. Build forest by repeating steps 1–4 for "n" number times to create "n" number of trees.

The beginning of random forest algorithm starts with randomly selecting "k" features out of total "m" features. Initially all the features and observations are taken in random. In the next stage the randomly selected "k" features to find the root node by using the best split approach. The next stage, the daughter nodes are calculated using the same best split approach. The first three stages is used to form the tree with a root node and having the target as the leaf node. Finally, the stages 1–4 is used to create "n" randomly created trees. This randomly created trees forms the random forest.

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1,\ldots,x_n$ with responses $Y = y_1,\ldots,y_n$, bagging repeatedly ($B$ times) selects a random sample with replacement of the training set and fits trees to these samples. For $b = 1,\ldots,B$:

- Sample, with replacement, $n$ training examples from $X$, $Y$; call these $X_b$, $Y_b$.
- Train a classification or regression tree $f_b$ on $X_b$, $Y_b$.

After training, predictions for unseen samples $x'$ can be made by averaging the predictions from all the individual regression trees on $x'$:

$$f^{\wedge} = \frac{1}{B}\sum_{b=1}^{B}, \ldots, \; fb(x') \tag{9}$$

### 3.5 Integrating R in to Hadoop

R is one of the most powerful machine learning platforms because of its breadth of the techniques such as data analysis, visualization, sampling, supervised learning and model evaluation techniques. R can access as the state-of-the-art algorithm and runs on any workstation platform. R is a computer language, interpreter and platform with a power in packages and functions. There are hundreds of machine learning packages and thousands of techniques available in R. The exploration and prototyping in the interactive environment can be performed using R. The environment is also very good for exploring new problem and great to use systematic process. R scripts can run from the command line called shell scripts from targets in a Make file to support larger dataset.

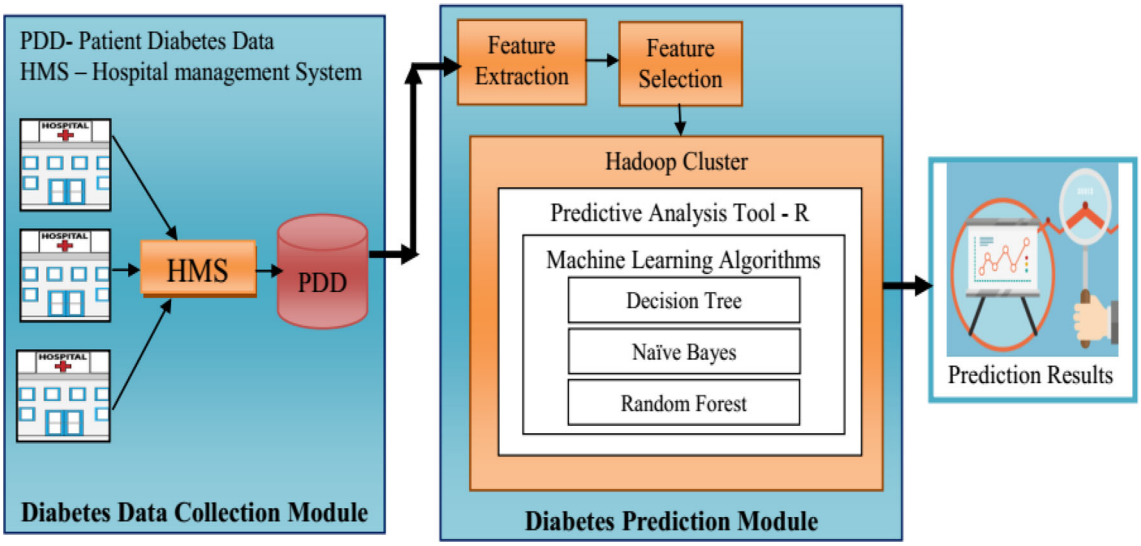


**Fig. 2** Predictive analysis system

**Table 1** Contingency table for evaluation

| | Actual label (expectation) | |
|---|---|---|
| Predicted label (observation) | a (true positive) correct result | b (false positive) unexpected result |
| | c (false negative) missing result | d (true negative) correct result |

Hadoop is a free, open source software platform for writing and running applications that process a large amount of data for predictive analytics. It enables the distributed parallel processing of large heterogeneous datasets generated from different sources. Hadoop uses the two main components such as MapReduce and Hadoop Distributed File Systems for performing predictive analytics. It can easily aggregate the huge mass of diverse data for investigating with complex operations such as clustering billions of documents. For experimental purpose a hadoop cluster with four nodes is framed and it is associated with a predictive analysis software R.

R can be integrated with Hadoop to scale data analytics to big data analytics. It can be performed by means of RHadoop, RHipe, ORCH, Hadoop Streaming (R package) and Hadoop Streaming (Hadoop Streaming Utility). Among the different methods mentioned RHadoop is great open source solution provided by Revolution Analytics. It has collection of five R packages such as rhdfs, rhbase, plyrmr, rmr2 and ravro to manage and analyze data with hadoop for recent tested releases.

Based on the architecture of the predictive system has two modules namely Diabetes data collection module and diabetes prediction module. Figure 2 shows the entire architecture diagram of the predictive analysis system.

### 3.6 Evaluation measure

The qualities of generated results by different machine learning algorithms were measured in terms of classification accuracy, precision value, recall value and f-score value. For classification of input dataset, contingency table of four different measures namely true positive, true negative, false positive and false negative were used. The system is trained to calculate these four measures by comparing the actual data and testing data based. on which accuracy measure is calculated. Contingency table is illustrated in the following Table 1.

Various evaluation parameters used for measuring the accuracy of machine learning algorithms are defined as follows:

*Precision* It is defined as the percentage of selected entities that are correctly classified in class C out of all the available entities in the dataset D. It is calculated using Eq. (7).

$$\mathrm{Pr}\,ecision = \frac{a}{a+b} \tag{10}$$

*Recall* It is defined as the percentage of correct entities that are selected in class C from all the available entities in the dataset actually belonging to class C. It is calculated using Eq. (8).

**Table 2** Evaluation parameters of machine learning algorithm

| Parameters | Decision tree | Naïve Bayes | Random forest |
|---|---|---|---|
| Precision | 87 | 91 | 94 |
| Recall | 77 | 82 | 88 |
| F-measure | 82 | 86 | 91 |
| Accuracy | 88 | 91 | 94 |

$$\text{Recall} = \frac{a}{a+c} \tag{11}$$

*F-measure* A measure that combines precision and recall is the harmonic mean of precision and recall. It is determined using below Eq. (9).

$$F - measure = 2.\frac{\text{Pr}\,ecision.Recall}{\text{Pr}\,ecision + Recall} \tag{12}$$

*Accuracy* It is defined as the Proportion of total number of predictions that are correctly classified in class C. It is determined using below Eq. (10).

$$Accuracy = \frac{a+d}{a+b+c+d} \tag{13}$$

## 4 Results and discussion

Investigations were carried out by selecting best features from the diabetes data using Information Gain. To handle the large dataset efficiently the R tool is embedded in to hadoop clusters. The extracted features were classified using the different machine learning algorithms namely decision tree, naïve bayes and random forest algorithm to make diabetes prediction. Values of all the evaluation parameters of different machine learning algorithms are represented in Table 2.

Figures 3, 4, 5 and 6 shows the comparison of the machine learning algorithms based on the classification accuracy, precision, recall and F-measure values.

From the Fig. 6 it is clear that precision measure of naïve bayes algorithm is 4% high than decision tree algorithm. Random forest algorithm produces 3% high in precision value than naïve bayes algorithm and 7% high in precision value than decision tree algorithm. Regarding the recall value, naïve bayes algorithm produces 5% higher result than decision tree algorithm. Random forest algorithm produces 6% high in recall value than naïve bayes algorithm and 11% high in recall value than decision tree algorithm.

As the results of F-measure value is concerned naïve Bayes algorithm produces 4% higher result than decision tree algorithm. Random forest algorithm produces 5% high

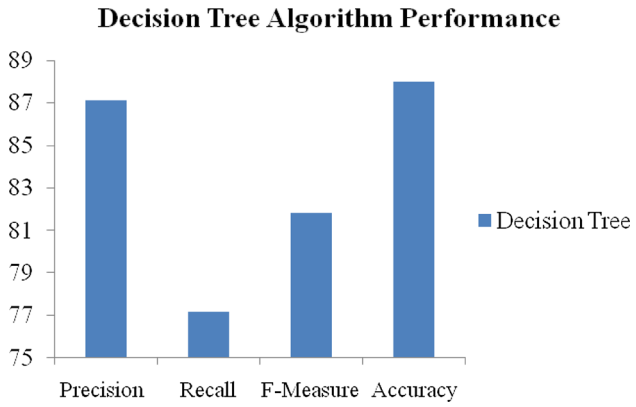


**Fig. 3** Performance measures of decision tree algorithm

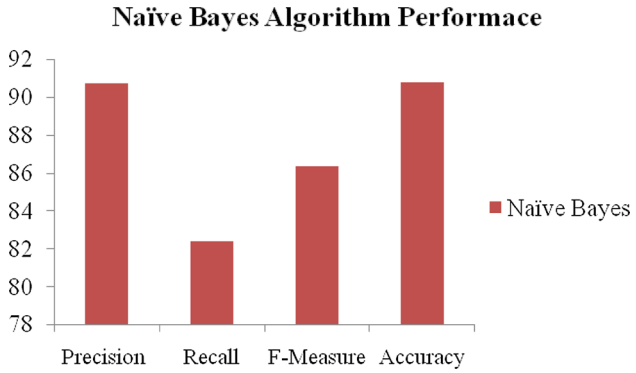


**Fig. 4** Performance measures of decision tree algorithm

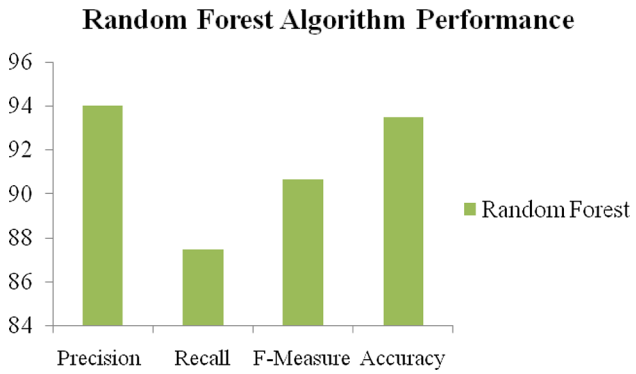


**Fig. 5** Performance measures of random forest algorithm

in F-measure value than naïve bayes algorithm and 9% high in F-measure value than decision tree algorithm.

Finally comparing the overall accuracy of the different machine learning algorithms naïve bayes algorithm produces 3% higher accuracy than decision tree algorithm. Random forest algorithm produces 3% high in accuracy than naïve bayes algorithm and 6% high in accuracy than decision tree algorithm.

From the result statistics obtained it is very clear that hadoop cluster based random forest algorithm performs much better in terms of all the different performance mea-

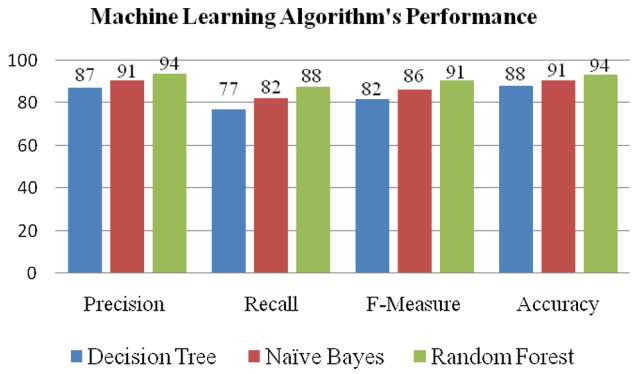


**Fig. 6** Performance measures of the machine learning algorithms

sures when compared to the other two machine learning algorithms.

## 5 Conclusion and future work

A healthcare system plays a main role nowadays in monitoring the health related aspects of the humans around the globe. A keen attention towards the diabetes and its precautionary measures is needed as at most urgent to prevent the various deadly side effects as mentioned. Random forest algorithm produces highest accuracy than decision tree and naïve bayes algorithm in 4 node hadoop cluster environment. As a future work Meta heuristic algorithms can be implemented as a part of machine learning algorithms by having more number of nodes in hadoop clusters.

## References


1. Song, T.M.: Efficient utilization of big data on health and welfare. Health Welf. Policy Forum. **193**, 68–76 (2012)
2. Kumar, S., Chakravarty, A.: ABC-VED analysis of expendable medical stores at a tertiary care hospital. Med. J. Armed Forces India **71**(1), 24–27 (2015)
3. Yuvaraj, N., Sabari, A.: An extensive survey on information retrieval and information recommendation algorithms implemented in user personalization. Aust. J. Basic Appl. Sci. **9**(31), 571–575 (2016)
4. Gebicki, M., Mooney, E., Chen, S.-J.G., Mazur, L.M.: Evaluation of hospital medication inventory policies. Health Care Manage. Sci. **17**, 215–229 (2014)
5. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf
6. https://www.cdc.gov/features/diabetes-statistic-report/index.html
7. Qi, Y., Jie, L.: Research of cloud storage security technology based on HDFS. Comput. Eng. Des. **8**, 2700–2705 (2013)
8. Huang, B., Xu, S., Pu, W.: Design and implementation of MapReduce based data mining, platform. Comput. Eng. Des. **2**, 495–501 (2013)
9. Yuvaraj, N., Sabari, A.: Twitter sentiment classification using binary shuffled frog algorithm. Intell. Autom. Soft Comput. **1**(1), 1–9 (2016)
10. Huang, W., Wang, H., Zhang, Y., Zhang, S.: A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop. Clust. Comput. (2017). https://doi.org/10.1007/s10586-017-1205-9
11. Bakshi, S., Sa, P.K., Wang, H., Barpanda, S.S., Majhi, B.: Fast periocular authentication in handheld devices with reduced phase intensive local pattern. Multimed. Tools Appl. (2017). https://doi.org/10.1007/s11042-017-4965-6
12. Chen, Q., Zhang, G., Yang, X., Li, S., Li, Y., Wang, H.H.: Single image shadow detection and removal based on feature fusion and multiple dictionary learning. Multimed. Tools Appl. (2017). https://doi.org/10.1007/s11042-017-5299-0
13. Wang, H., Wang, J.: An effective image representation method using kernel classification. In: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 853–858 (2014)
14. Zhang, J., Williams, S.O., Wang, H.: Intelligent computing system based on pattern recognition and data mining algorithms. Sustain. Comput. (2017). https://doi.org/10.1016/j.suscom.2017.10.010
15. Yuvaraj, N., Sabari, A.: Performance analysis of supervised machine learning algorithms for opinion mining in e-commerce websites. Middle-East J. Sci. Res. **1**(1), 341–345 (2016)
16. Chapelle, O., Sindhwani, V., Keerthi, S.S.: Optimization techniques for semi- supervised support vector machines. J. Mach. Learn. Res. **9**, 203–233 (2013)
17. Zhang, N., Chandrasekar, P.: Sparse learning of maximum likelihood model for optimization of complex loss function. Neural Comput. Appl. **28**(5), 1057–1067 (2017)
18. Zhang, S., Wang, H., Huang, W.: Two-stage plant species recognition by local mean clustering and Weighted sparse representation classification. Clust. Comput. **20**(2), 1517–1525 (2017)
19. Smys, S., Kumar, A.D.: Secured WBANs for pervasive m-healthcare social networks. In: 10th International Conference IEEE on Intelligent Systems and Control (ISCO), January 2016, pp. 1–4. (2016)
20. Huang, S., Wang, B., Wang, G.: A survey on MapReduce optimization technologies. J. Front. Comput. Sci. Technol. **10**, 885–905 (2013)
21. Gao, S., Li, L., Li, W., Janowicz, K., Zhang, Y.: Constructing gazetteers from volunteered big geo-data based on Hadoop. Comput. Environ. Urban Syst. **61**, 172–186 (2017)
22. Li, J., Cui, J., Wang, D., et al.: Survey of MapReduce parallel programming model. Acta Electronica Sinica **11**, 2635–2642 (2011)
23. Chen, J., Chen, H., Wan, X., Zheng, G.: MR-ELM: a MapReduce-based framework for large-scale ELM training in big data era. Neural Comput. Appl. **27**(1), 101–110 (2016)
24. Huang, W., et al.: A novel cluster computing technique based on signal clustering and analytic hierarchy model using hadoop. Clust. Comput. (2007). https://doi.org/10.1007/s10586-017-1205-9
25. Cai, Z., Deng, L., Li, D., Yao, X., Cox, D., Wang, H.: A FCM cluster: cloud networking model for intelligent transportation in the city of Macau. Clust. Comput. (2017). https://doi.org/10.1007/s10586-017-1216-6
26. Pattern mining model based on improved neural network and modified genetic algorithm for cloud mobile networks
27. Wang, Y., Li, J., Wang, H.H.: Cluster and cloud computing framework for scientific metrology in flow control. Clust. Comput. (2017). https://doi.org/10.1007/s10586-017-1199-3
28. Haindl, M., Somol, P., Ververidis, D., Kotropoulos, C.: Feature selection based on mutual correlation. In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (eds.) Progress in Pattern Recognition, Image Analysis and Applications. CIARP 2006. Lecture Notes in Computer Science, vol. 4225, pp. 569–577. Springer, Berlin, Heidelberg (2006)

**N. Yuvaraj** Associate Professor in Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India. He has completed his Masters of Engineering in Sofware Engineering. He has completed his Doctorate in the area of Data Science. He has two years of Industrial Experience and eight years of teaching experience. He has published over 21 technical papers in various International Journals and conferences. His areas of interest include data science, sentimental analysis, data mining, data analytics and information retrieval, distributed computing frameworks.

**K. R. SriPreethaa** Assistant Professor (Sr.G) in Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore. She has 9 years of teaching experience. Her area of research includes Data analytics, sentiment analysis, Information retrieval and SEO on which she has published over 17 technical papers in conferences and journals.