International Journal of Aquatic Science
ISSN: 2008-8019
Vol 12, Issue 03, 2021

Comparative Study On Classification Methods To Diagnosis The Diabetics

P.Dhivya[1], P.Parthasarathi[2] Dr.A.Bazila Banu[3]

[1, 2,3] *Bannari Amman Institute of Technology, Erode, India*

Email: [1]dhivyap@bitsathy.ac.in,[2]parthasarathip@bitsathy.ac.in, [3]bazilabanu@bitsathy.ac.in

Abstract. The information disclosure from clinical datasets is amazingly huge to make a successful clinical conclusion. The objective of the Machine Learning is to gain information from a dataset and adjust it into a fitting structure for additional utilization. Diabetic Mellitus remains as a broadly rising incessant infection, and this is an incredible test around the world. Today, it is basic in different age bunches extent as of youngsters to grown-ups. As the quantity of Diabetic Mellitus persistent have been multiplying each year explicitly in India. In the proposed work, comparative study on various classification algorithm such as Naïve Bayes, Random Forest, Decision tree, K Nearest Neighbor (KNN) , Support Vector Machine(SVM) on dataset to predict whether the given person affected with diabetic or not. In this work, a new ensemble method is identified to provide better accuracy such as 85.44 % compared with existing classification algorithm.

Keywords: Decision tree, K Nearest Neighbor (KNN), Logistic Regression, Random Forest, Support Vector Machine (SVM).

1. INTRODUCTION

Today, India has become the world's content capital with at most 50million people. The clinical experts feel that earlier identification and exact administration can help patients in getting an ordinary life. The sugar malady prompts a countless medical issue like Pressure, delay in mending of wounds, tiredness, obscured vision, and so forth. Additionally, according to the World Health Organization diabetes actuality sheet, 34lakhs passing are happened as a result of High Blood Sugar. The following symptoms are identified and test should be conducted. General symptoms of Diabetic are frequent pee, Loss of body weight, frequent hunger, Slow mending disease, Weight loss and frequent regurgitating**.** The diagnosis test conducted like Random BGL, Oral glucose tolerance test, Urine test and Fasting BGL [2].

Machine Learning has various applications in a few areas to be specific natural information investigation, media transmission industry, budgetary information examination, and so on. With the expanding research discoveries on the well being informatics area, different strategies are being found. The examination of house enormous amount of information is very mind boggling and needs outrageous information [3]. Today, E-human services execute Machine Learning techniques and furthermore media transmission strategies for well being related analysis. Just a couple of patients need a relentless well being check, in this way, need a specialist's support in a split second. The following steps to be carried out for processing the data [12].

Understanding about dataset:
The starter step is to grasp the necessities. Understanding is fundamental to have an obvious perception of the thought of the application and targets, regardless of whether it is to upgrade deals, anticipate financial exchange and so forth.

Cleaning and Technology selection: Here is the place commotion, just as insignificant information, is evacuated as of the Big Data set. Information cleaning is an amazingly mandatory advance since yield would be founded on quality. Information cleaning incorporates evacuating excess records, entering consistently right qualities for vacant or invalid records, expelling additional information fields, normalizing information group, and furthermore refreshing information in a convenient mode, etc[13]. Through change or dimensionality decrease approaches, information is transmuted into an appropriate structure preparing it for Machine Learning step. Appropriate strategy ought to be chosen for looking through the example as of the information. The model and parameter ought to be Suitable for the technique.

Knowledge extraction and Evaluation
Machine Learning is the real quest for designs as of the existent information utilizing the chose techniques. It is a post-preparing part that translates mined examples alongside connections. In case the example evaluated isn't useful, at that point the procedure perhaps will begin again from any of the previous advances and iterative procedure. The information discovered is blended and meant to the client in an agreeable and direct to appreciate group. The greater part of the occasions, perception systems are being used to make data justifiable by the clients just as by the translators [4].Machine learning to be found an important role in the field of prediction in business or medical data. The machine learning algorithm is classified into supervised, unsupervised and semi supervised. The Classification under supervised learning is the well-known Machine Learning task. A large amount of medical datasets are usually there in classification. Classification algorithm can allot the substance in a group to the targeted categories. A role of a model which delineates the attributes as a function of input attributes. In machine learning, classification is used to classify, predict and diagnosis the disease. The classification exactly predicts the targeted class for every case in the data. To learn a collection of data that is required to be clustered for on their features and combining them according to the similarities. Clustering is not utilizing the single attribute for prediction. All those input attributes are equally treated. The attribute values are required to be formalized prior to clustering [5].

2. LITERATURE SURVEY

In the literature survey various classification algorithm were discussed.

2.1 Neural Network Classifier
Neural framework involves units (neurons), composed in layers, which convert a data vector into some yield. Each unit takes data, applies a (normally nonlinear) ability to it and subsequently gives the respect the accompanying layer. Generally the frameworks are described to be feed-forward: a unit deals with its respect all the units on the accompanying layer, yet there is no analysis to the past layer [6].

2.2Principal Component Analysis
Huge datasets are progressively far reaching in numerous controls. So as to interpret such datasets, techniques are required to definitely decrease their dimensionality in an interpretable manner, with the end goal that a large portion of the data in the information is protected. Nu-

merous strategies have been created for this reason, however PCA is one of the most seasoned and most generally utilized. It's thought is basic—decreases the dimensionality of a dataset, while protecting as much 'fluctuation' (for example factual data) as could be expected under the circumstances. The components [as a whole] form an orthogonal foundation for the space of the data [11].

2.3 Random Forest

Random Forest is used to create the forest by applying different techniques randomly. There are Forest creations and prediction steps will be followed [7].

- Initially select the features from whole randomly
- Identify the node from the features using split point
- Split the selected node and then construct the tree

The above steps will be carried until it reaches 1.

2.4Logistic Regression

Logistic Regression predicts the probability of dependent variable such as yes or no (0 or 1).It is mainly used to fit the model for categorical information. It is classified into binary, multinomial and ordinal logistic regression. If the categorical output has only yes or no, then it is binary logistic regression. If there is more categorical information in random order, then it is multinomial logistic regression. In case of more categorical information in order, then it is ordinal logistic regression [4].

2.5K Nearest Neighbour

In the K Nearest neighbour, prediction is made based on the similar records then locates it. Based on the summarization of neighbours, the prediction is made. The Euclidean Distance is identified to find the nearest neighbour. Weight is assigned to each neighbour [3].

2.6 Support Vector Machine

It remains as a regulated learning process,i.e. informational index is prepared in such a way that it might offer pre-decided yield In AI, uphold vector machines (SVMs, moreover uphold vector networks[1]. A SVM model is a depiction of the models as centres in space and various groupings are isolated by an undeniable opening that is as wide as could be normal considering the present situation [8].

2.7 Naive Bayes Classifier

Bayes is an essential technique for building classifiers models where the class names are drawn from some restricted set. There is authentically not a singular count for getting such classifiers, anyway a gathering of computations subject to a normal principle and acknowledge that the assessment of a particular segment is liberated from the assessment of some other component of given class variable [10][14].

2.8 Decision trees

Decisions and their possible outcomes, including chance event results, resource costs, and utility. It is one way to deal with show a computation that just contains prohibitive control clarifications. A Decision tree is a flowchart-like structure in which every inside centre point addresses a "test" on a property (for instance whether or not a coin flip comes up heads or tails), each branch addresses the consequence of the test, and each leaf centre addresses a class mark. The ways from root to leaf address request manages everything. [9].

3. COMPARATIVE STUDY ON CLASSIFICATION ALGORITHM

In the proposed work, the followings things to be measured for pima Indian diabtes dataset was taken. The dataset contains 768 entries with the following feature [5].
The following output is obtained for various algorithms in fig1.

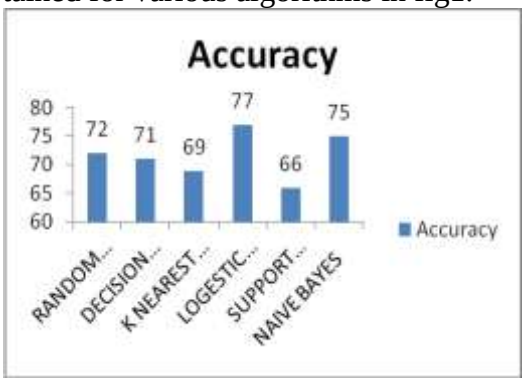

Fig.1. Accuracy prediction of various algorithms

----------RANDOM FOREST----------

	precision	recall	f1-score	support
0	0.80	0.78	0.79	208
1	0.57	0.60	0.58	100
micro avg	0.72	0.72	0.72	308
macro avg	0.68	0.69	0.69	308
weighted avg	0.73	0.72	0.72	308

[[162 46]
[40 60]]

----------SUPPORT VECTOR MACHINE----------

	precision	recall	f1-score	support
0	0.66	1.00	0.79	202
1	0.00	0.00	0.00	106
micro avg	0.66	0.66	0.66	308
macro avg	0.33	0.50	0.40	308
weighted avg	0.43	0.66	0.52	308

[[202 0]
[106 0]]

----------NAIVE BAYES----------

	precision	recall	f1-score	support
0	0.82	0.79	0.80	202
1	0.62	0.66	0.64	106
micro avg	0.75	0.75	0.75	308
macro avg	0.72	0.73	0.72	308
weighted avg	0.75	0.75	0.75	308

[[160 42]
[36 70]]

----------DECISION TREE----------

	precision	recall	f1-score	support
0	0.81	0.74	0.77	202
1	0.57	0.67	0.62	106
micro avg	0.71	0.71	0.71	308
macro avg	0.69	0.70	0.69	308
weighted avg	0.73	0.71	0.72	308

[[149 53]
[35 71]]

----------K NEAREST NEIGHBOURS----------

	precision	recall	f1-score	support
0	0.76	0.77	0.76	202
1	0.55	0.54	0.54	106
micro avg	0.69	0.69	0.69	308
macro avg	0.65	0.65	0.65	308
weighted avg	0.69	0.69	0.69	308

[[155 47]
[49 57]]

----------LOGESTIC REGRESSION----------

	precision	recall	f1-score	support
0	0.81	0.86	0.83	202
1	0.69	0.61	0.65	106
micro avg	0.77	0.77	0.77	308
macro avg	0.75	0.73	0.74	308
weighted avg	0.77	0.77	0.77	308

[[173 29]
[41 65]]

Fig. 2. Output of the classification algorithm

From the fig2, the logistic regression gives better performance with the accuracy of 77% and also naïve bayes algorithm performs nearly logistic regression. The working process classification algorithm of Machine Learning explained in fig 3.But the output is not up to the expected level. In the proposed work a new ensemble method is developed to increase the accuracy. The algorithm analysis is given in the figure 4.

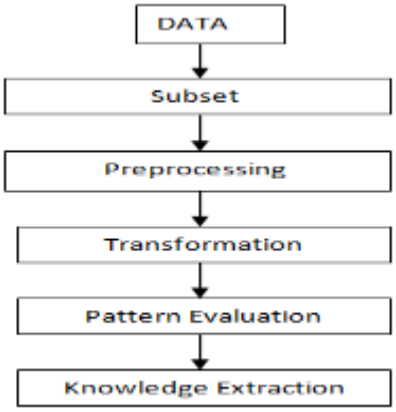

Fig. 3. Working process of Machine Learning

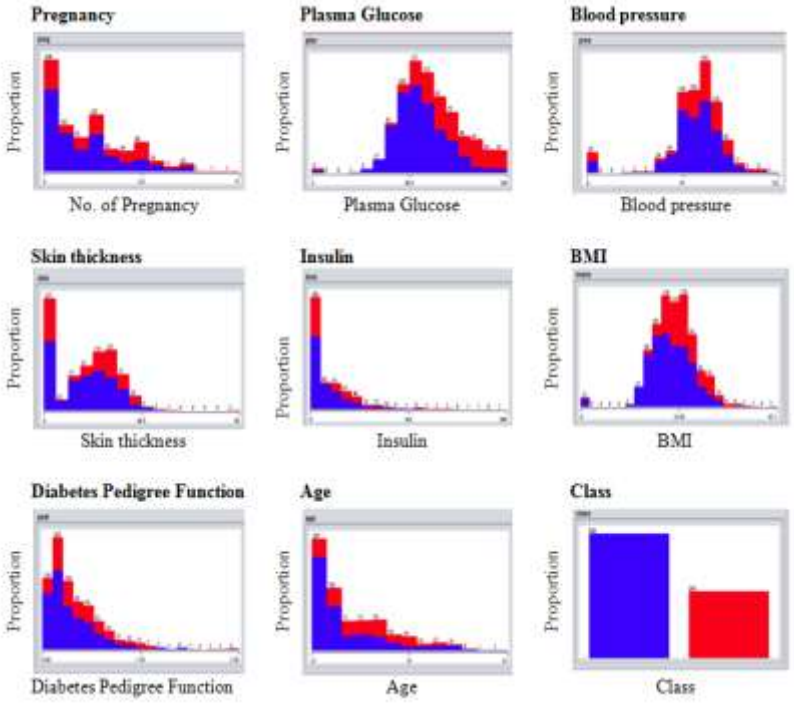

Fig. 4. Algorithm Analysis

4. PROPOSED WORK

In the proposed work, the new ensemble method was developed to obtain better accuracy. The Pareto Distribution method has been applied to get better accuracy. The Pareto distribution is investigate distributions associated Population sizes, the occurrence of natural resources, the size of companies, personal incomes, stock price fluctuations, and error cluster-

ing in communication circuits. In the modified new ensemble method gives the following accuracy.

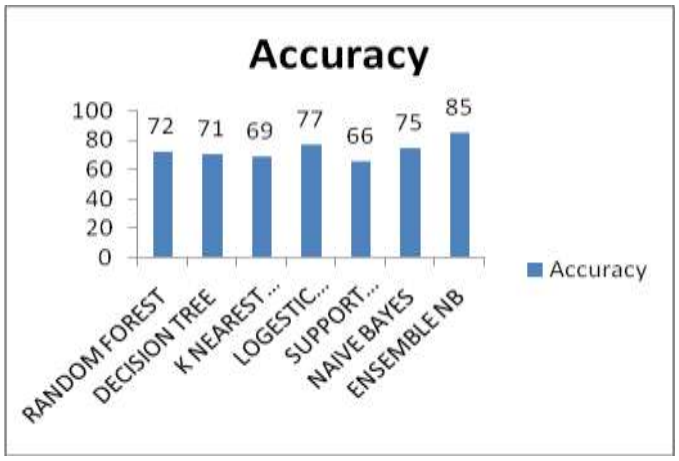

Fig. 5. Accuracy prediction of Ensemble NB

Thus the above ensemble NB gives better accuracy such as 85.45 which is greater than the accuracy obtained in normal logistic regression and Naïve bayes algorithm given in fig 5.

5. CONCLUSION

Machine Learning is the process of taking out the useful information and finds the pattern from the large database. Based on the obtained pattern, it will predict the output for future input. There is various classification algorithms designed to get better accuracy. In the comparative study various algorithm are discussed and its accuracy was obtained. By the comparison, Logistic regression and Naïve bayes performed well equally. By applying the Pareto Distribution in the Naïve Bayes produced the better accuracy with 85%. With this new ensemble method, the model performance increased with improved prediction accuracy.

6. REFERENCES

[1]. RuchaShinde, Sandhya Arjun, PriyankaPatil&JaishreeWaghmare 2015, An intelligent heart disease prediction system using K-means clustering and Naïve Bayes algorithm, International Journal of Computer Science and Information Technologies, vol. 6, no. 1, pp. 637-639.

[2]. Santhanam, T. &Padmavathi, M. S. 2015, Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis, Procedia Computer Science, vol. 47, pp. 76-83.

[3]. SiriwanSuebnukarn, PawornwanRittipakorn, BudsaraThongyoi, KwanwongBoonpitak, MansuangWongsapai&PanuPakdeesan 2013, Usability assessment of an electronic health record in a comprehensive dental clinic, SpringerPlus, vol. 2, no. 1, pp. 220.

[4]. SuselGóngora Alonso, Isabel de la Torre Díez, Joel JPC Rodrigues, SofianeHamrioui& Miguel López-Coronado 2017, 'A systematic review of techniques and sources of Big Data in the healthcare sector', Journal of Medical Systems, vol. 41, no. 11, pp. 183.

[5]. Yasodha, P &Kannan, M 2011, 'Analysis of a population of diabetic patients databases in WekaTool', International Journal of Scientific & Engineering Research, vol. 2, no. 5.

[6]. Chen M., Y. Hao, K. Hwang, L. Wang & L. Wang 2017, 'Disease prediction by machine learning over Big Data from healthcare communities', IEEE Access, vol. 5, pp. 8869–8879

[7]. Vuksan V., Xu, Z.Z., Jovanovski, E. et al. 2018, 'Efficacy and safety of American ginseng (Panaxquinquefolius L.) extract on glycemic control and cardiovascular risk factors in individuals with type 2 diabetes: a double-blind, randomized, cross-over clinical trial', European Journal of Nutrition, pp. 1-9.

[8]. Begum, S. A., R. Afroz, Q. Khanam, A. Khanom& T. S. Choudhury 2014, 'Diabetic Mellitus and gestational Diabetic Mellitus', Journal of Paediatric Surgeons of Bangladesh, vol. 5, no. 1, pp. 30-35.

[9]. Danielle AJM Schoenaker,, Yvonne Vergouwe, Sabita S. Soedamah-Muthu, Leonie K. Callaway & Gita D. Mishra 2018, 'Preconception risk of gestational diabetes: Development of a prediction model in nulliparous Australian women', Diabetes Research and Clinical Practice, vol. 146, pp. 48-57.

[10]. Enrico Capobianco 2017, 'Systems and precision medicine approaches to diabetes heterogeneity: a Big Data perspective', Clinical and Translational Medicine, vol. 6, no. 1, pp. 23.

[11]. Halili, F. & A. Rustemi 2016, 'Predictive modeling: Machine Learning regression technique applied in prototype', Internatonal Journal of Computer Science and Mobile Computing, vol. 5, no. 8, pp. 207-215.

[12]. H. Zhang, X. Feng, H. Liu, P. Guo, S. Krishnamoorthy and C. Zhang, "Cloud-Based Class Attendance Record System," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 2019, pp. 283-287, doi: 10.1109/ICCC47050.2019.9064482.

[13]. Yasoda, K., Ponmagal, R.S., Bhuvaneshwari, K.S. K Venkatachalam, " Automatic detection and classification of EEG artifacts using fuzzy kernel SVM and wavelet ICA (WICA)" Soft Computing Journal (2020).

[14]. S. Ramamoorthy, G. Ravikumar, B. Saravana Balaji, S. Balakrishnan, and K. Venkatachalam, "MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-8, 2020.