Diabetes Mellitus Prediction using Ensemble Machine Learning Techniques

Jyoti, Peri Arjun

Abstract: ***The healthcare industry is inflicted with the plethora of patient data which is being supplemented each day manifold. Researchers have been continually using this data to help the healthcare industry improve upon the way major diseases could be handled. They are even working upon the way the patients could be informed timely of the symptoms that could avoid the major hazards related to them. Diabetes is one such disease that is growing at an alarming rate today. In fact, it can inflict numerous severe damages; blurred vision, myopia, burning extremities, kidney and heart failure. It occurs when sugar levels reach a certain threshold, or the human body cannot contain enough insulin to regulate the threshold. Therefore, patients affected by Diabetes must be informed so that proper treatments can be taken to control Diabetes. For this reason, early prediction and classification of Diabetes are significant. This work makes use of Machine Learning algorithms to improve the accuracy of prediction of the Diabetes. A dataset obtained as an output of K-Mean Clustering Algorithm was fed to an ensemble model with principal component analysis and K-means clustering. Our ensemble method produced only eight incorrectly classified instances, which was lowest compared to other methods. The experiments also showed that ensemble classifier models performed better than the base classifiers alone. Its result was compared with the same Dataset being applied on specific methods like random forest, Support Vector Machine, Decision Tree, Multilayer perceptron, and Naïve Bayes classification methods. All methods were run using 10k fold cross-validation.***

Keywords: Diabetes, Machine learning, Ensemble, Dataset.

I. INTRODUCTION

Prediabetes is a disorder in which the amount of blood glucose is sufficiently high to be classified as Diabetes, but not as extreme as usual. The pancreas insulin-producing cells of the body have a type of type-1 Diabetes, which eventually kills over 90 percent of them. For all people with Diabetes, just approximately 5 or 10 percent have Type 1. [1]. The chance of having type 2 diabetes is increased for those with African, Hispanic, American Indian, and Mexican, or Latino American ancestry who reside in the USA. Diabetes may cause permanent tissue harm and dysfunction, particularly eyes, kidneys, ears, blood vessels, and nerves. Diabetes can be classified into Type 1 diabetes (T1D) and Type 2 diabetes (T2D). Type 1 diabetes patients are usually younger, often under 30. Typical health signs include elevated hunger, repeated urination, and excessive blood pressure. A form of Diabetes cannot be treated successfully with oral medications alone, so patients require insulin therapy. Type 2 diabetes is prevalent in middle-aged and elderly persons, frequently combined with obesity, hypertension, dyslipidaemia, arteriosclerosis, and other diseases. Recently, various algorithms have been used to forecast Diabetes, including conventional machine learning [9], such as support vector machine (SVM), decision tree (DT), technical regression, etc. Researchers [4] separated Diabetes patients from average persons by utilizing PCA and neuro-fuzzy inference. [15] used quantum particle swarm optimization (QPSO) algorithm and weighted least squares supporting vector machine(WLS-SVM) to predict type 2 diabetes [7] proposed a diabetes prediction system, called LDA-MWSVM. The writers used Linear Discriminant Analysis (LDA) to minimize measurements and isolate the functions. [6] Built prediction models based on logistic regression for different Type 2 diabetes prediction onsets to deal with high-dimensional datasets. [10] Concentrated on glucose and used support vector regression (SVR) to model diabetes as a multivariate regression issue. However, more and more experiments used fixed approaches to increase precision [9]. [11] Suggested a modern ensemble strategy, called rotation wood, incorporating 30 methods of machine learning. Han et al. (2015) suggested a machine learning approach that modified predictive principles for SVM. We study the ensemble algorithm with Principle component analysis, and K means clustering for predicting Diabetes and Dataset classification.

II. LITERATURE REVIEW

Ensemble methods are statistical and computational learning procedures. They are in sync with the human social learning of trying different opinions before making any final decision. Sets of learning machines are used to combine choices and provide more robust and accurate predictions on controlled and unattended learning problems [3, 4]. There is no single, theoretically sound explanation for classifier ensemble methods [4]. The Machine Learning approach suggested by [5], changing the SVM rules for prediction. Comparison analysis between Naïve Bayes, Decision Tree and K-NN algorithms has been performed. The decision tree was simple, and the methods together are statistical and computational learning procedures that reminisce about acts of human social learning before making a final decision to seek different opinions. There have recently been various algorithms for predicting Diabetes, including the standard learning method [9], for example, the vector support system (SVM), decision machine (SVM), Logistic regression, Decision Tree (DT), etc.[4], using the critical component analysis (PCA) and fuzzy neuro inference, have separated Diabetes from healthy people.

Revised Manuscript Received on June 22, 2020.

***** Correspondence Author

Dr. Jyoti Verma, Assistant Professor, Department of Computer Science, J.C. Bose University of Science and Technology, Haryana, India.

Peri Arjun, Student, Department of Computer Science, J.C. Bose University of Science and Technology, Haryana, India.

[5] have suggested a diabetes prediction method, called LDA-MWSVM, to prevent quantum particulate swarm optimizing algorithms and weighted least squares supporting the vector system (WLS-SVM) for the prediction of type 2 diabetes.

The authors have used this method to reduce dimensions and extract features with Linear Discriminant Analysis (LDA). [6] have developed logistically regressive prediction models for various types of type 2 diabetes predictions to deal with the high-dimensional datasets. Besides, more and more research has used ensemble approaches to improve accurately [7]. [8] focused on glucose and used vector regression (SVR) as a multivariate problem regression to predict Diabetes. [9] suggested a new ensemble approach incorporating 30 methods of auto-learning, namely, rotational forest. [10] suggested that the Machine Learning approach, changing the SVM rules for easy-to-interpret, fell short on two accounts, i.e., the target needed to have discrete values. Secondly, it suffers if there are many complex interactions between relevant attributes. The merits of the Naïve Bayes algorithm being an intuitive technique was that there was no need to set values of parameters before proceeding. The probabilities returned by this algorithm can be easily applied to further experiments or analyses. Its learning rate is fast, as it starts classifying with few datasets. It is also computationally fast when making decisions. The only drawback contrasting with the advantages is its assumption of class conditional independence[16] proposed intelligible support vector machines for diagnosis of diabetes mellitus, etc. The classifiers should be used for diabetes prediction, and they are recommended to improve them through the production of hybrid models [17]. In short, deep learning can extract useful data from EHRs by studying features related to diabetes outcomes in comparison with traditional machine learning models, and therefore helps to target people who are likely to develop the disease in order for them to change their way of life [18] A significant number of filter-based filtering techniques, such as SVM, gain ratio, data benefit and Decision tree were found in the literature. The key issues with a filter-based collection of features were (i) the bulk of the features do not accept consistency, (ii) the limitation of a particular filter-based system in the chosen function subset and (iii) poor prediction accuracy during classification. To solve these challenges, an Ensemble method with Principle component analysis and K-means clustering is used, which increases classification prediction accuracy.

III. MATERIALS AND METHODS

A. Data

Data acquisition from different sources is always raw data that may contain mistakes, outliers, or missing values. This data has to be preprocessed. There are various means and strategies to deal with different problem areas. These could be utilizing data cleaning, decertification, and data transformation to allow the use of those sets in the data mining process [11]. The data preparation alone is estimated to account for 60% of the entire data mining operation's expansion. The Dataset used for this analysis is 'Diabetes Dataset for Pima Indians.' This Dataset contains 768 instances, and each instance contains eight input (X1 to X8) attributes and an output (Y) attribute.

B. Classification

The Dataset contained some incomplete details, one of the key drawbacks. In WEKA 3.8 (Waikato Setting for Knowledge), two well-liked and useful functions were used to manage the problem. The 'ReplaceMissingValue' function was used to substitute missing data from the Dataset at first. For each nominal and numeric attribute, this feature combines all missed details with modes [12]. Another feature named Randomize was used that does not significantly affect the overall output by replacing the missed area.

C. Design

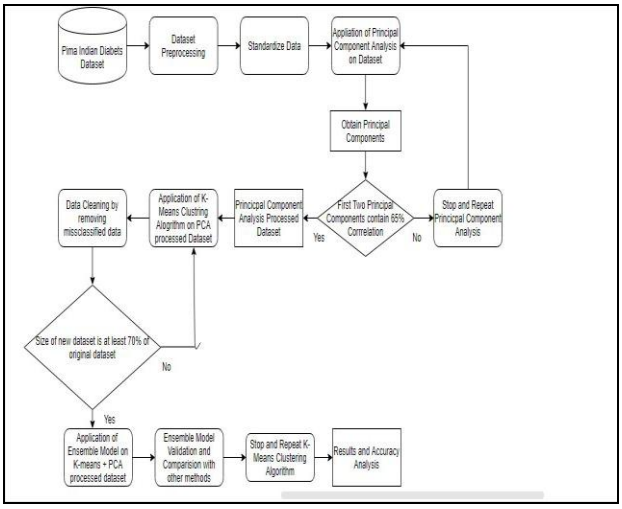


Fig 1. Flowchart for Prediction of Diabetes using Ensemble Machine Learning Algorithm

This above model in Fig 1 shows the proposed flowchart where the primary process of dataset standardization and classification is done. The principal component analysis (PCA), while preserving trends and patterns, simplifies the dynamics of extensive dimensional results. It is done by turning details into function summaries in fewer dimensions. High dimensional results are very popular in biological science and occur when many characteristics are calculated per sample, including the expression of several genes. That form of data poses a variety of problems that PCA mitigates: device expenses and an improvement in the error rate due to numerous check adjustments as each function is checked for accuracy.

IV. FEATURE SELECTION

The impetus behind the entire approach to learning was recently applied to other areas of computer learning, such as the collection of apps. The aim is then to produce more reliable performance than a single function selection approach by integrating the outputs of different feature selection models. However, are not only many versions usable, as is the case for classification ensembles, but also the different subsets of features obtained. Function collection sets can be categorized according to a number of parameters regarding one or more of the above, but the easiest distinction applies to the form of selectors used.

The ensemble is known as homogenous if the basis selectors are all of the same nature; otherwise, the ensemble is heterogeneous.

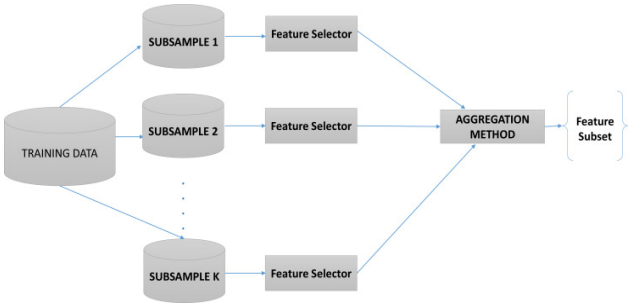


Fig 2. Ensemble feature selection method.

A. Principle Component analysis

The central study of components is primarily a modification of the co-ordinate. On an X-axis and a Y-axis, the initial details were shown. PCA attempts to rotate these two axes with two-dimensional data such that the current X-axis lies in the highest data volatility path. PCA includes perpendicular axes, because the option of X' decides Y' in two dimensions. This new collection of axes, X' and Y', helps you to read the transformed data. The first axis is in the direction of most of the variation, for more than two dimensions; the second is in the direction of the next variation. The independent variable, with PCA, is a unit vector that points towards a different co-ordinate axis. The axis that displays more variance is the axis with the highest value.

V. MEASUREMENT

A. K-means Clustering with PCA

We examine the link between those two commonly used methods in this paper. We show that the critical components in the K-means clustering system are simply an ongoing solution for the cluster participation predictor, i.e., PCA dimensional reduction automates data clustering according to the 'Means Objective rule'[20]. The PCA-based data reduction is, thus, importantly justified. To characterize the results, K-means method uses K simulations, the cluster centers. The sum of squared errors is reduced.

$$J_{\mathrm{K}} = \sum_{k=1}^{K} \sum_{i \in C_k} (\mathbf{x}_i - \mathbf{m}_k)^2 \qquad \text{.....(1)}$$

where $(P\ x_1, \cdots, x_n) = X$ is the data matrix and $m_k = i \in C_k\ xi/n_k$ is the centroid of cluster C_k and n_k is the number of points in C_k. Standard iterative solution to K-means suffers from a well-known problem- as iteration proceeds, the solutions are trapped in the local minima due to the greedy nature of the update algorithm.[18] K-means algorithm automatically handles a mixture of categorical and numerical attributes. Furthermore, the algorithm automatically normalizes numerical attributes when doing distance computations.

In these two types of data reduction, the PCA plays a crucial role in approximating the signal feature vectors of the cluster centroids. In summary, it ensures the automatic identification of the subspace of the cluster by the PCA dimension reduction that the K-means grouping is particularly useful.

B. Comparison

Table 1: Comparison of various algorithms and their comparison on F-measure.

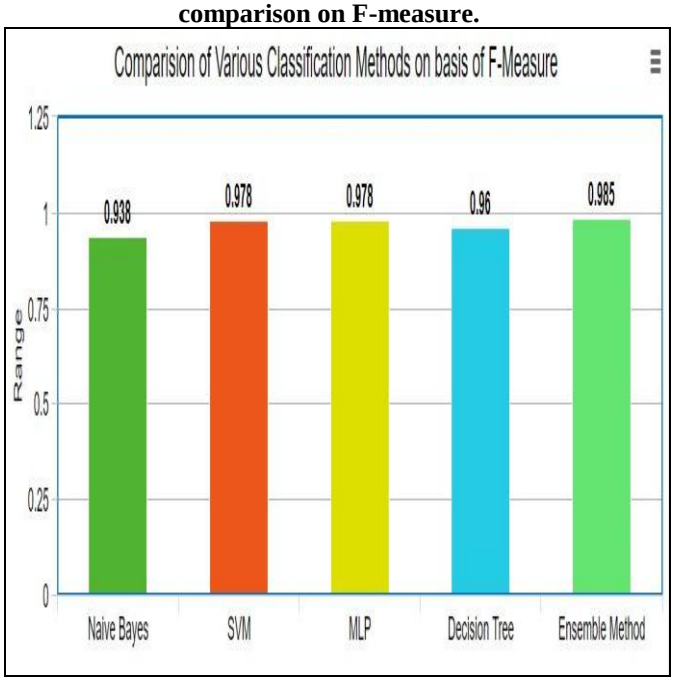


Table 3: Comparison of various algorithms and their comparison based on recall.

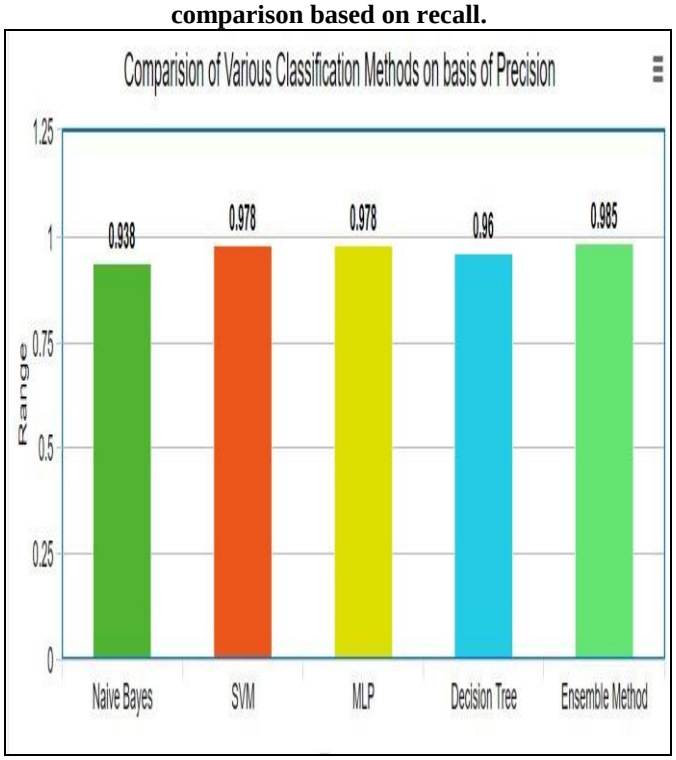


Table 2: Comparison of various algorithms and their comparison based on precision.

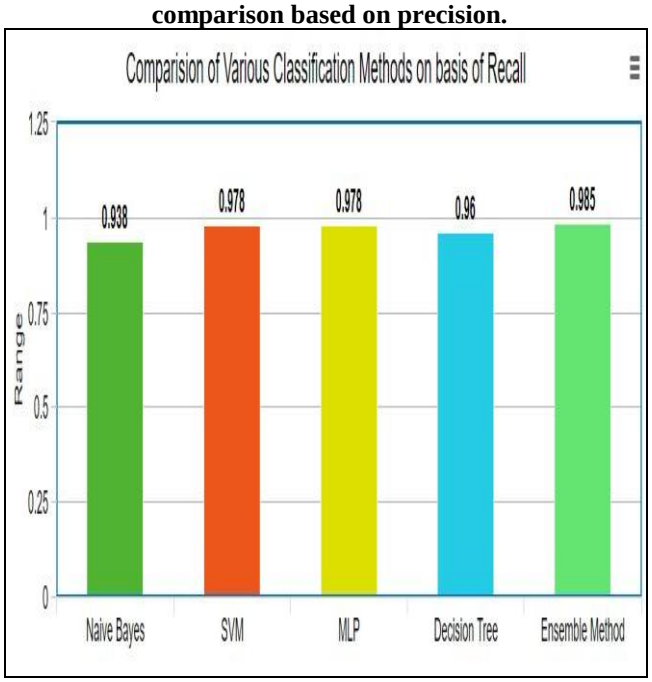
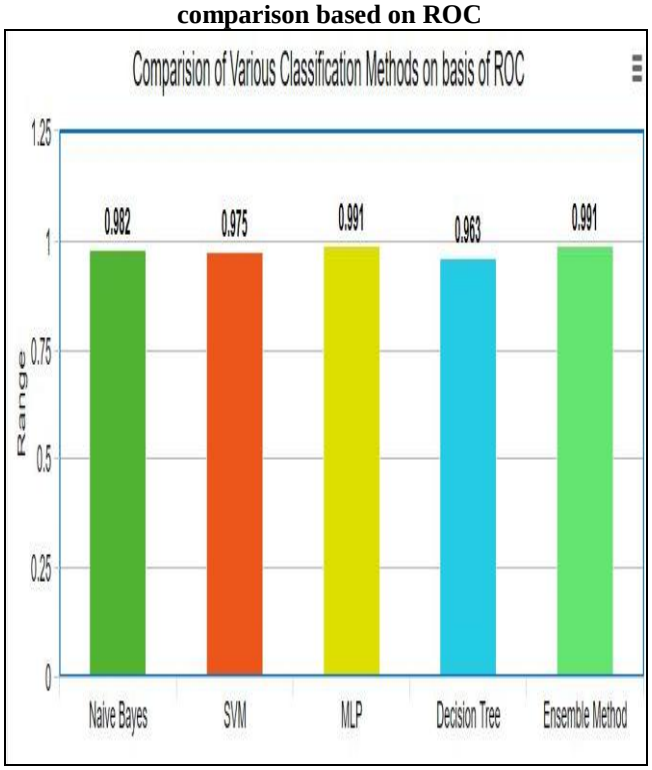
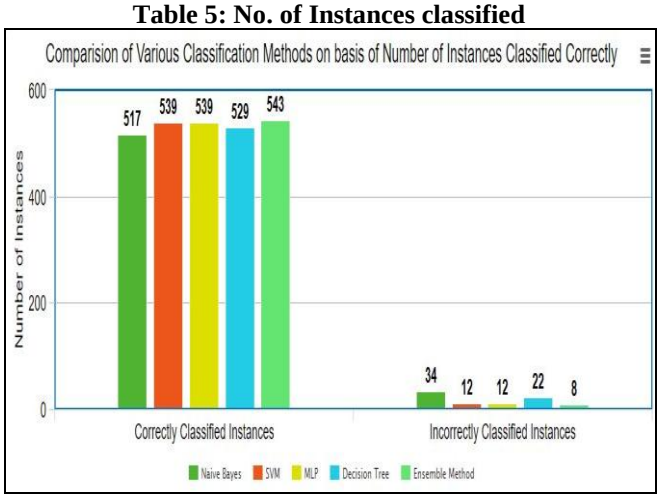


Table 4: Comparison of various algorithms and their comparison based on ROC



It experimented in that SVM algorithm, and Proposed ensemble method produced similar ROC. The proposed ensemble method produced the highest accuracy with 98.54%, while Naïve Bayes theorem produced the lowest accuracy and gave 93.82% accuracy.

Table 5: No. of Instances classified



Out of 768 instances of Dataset, which we obtained from the PIMA India dataset were classified using a given algorithm in the above Table 2. It can be seen that Naïve Bayes produced 517 correct instances and 34 incorrectly.

VI. CONCLUSION

This paper proposes a fast and accurate diabetes prediction method. For each of the PIMA India datasets, the proposed system used 768 instances in 8 attributes. In order to remove unwanted data and to speed up processing time, the used data are preprocessed. In comparison, the separation of the data set into sub-set culminated in an optimal grouping. A function description and grouping sections became the subject of the proposed program. The ideas of such pieces produce the most reliable possible outcomes. The experiment findings demonstrated the benefits of utilizing the method algorithms by reaching a higher rate of classification than the other methods. Non-linear algorithms or a combination of linear or non-linear algorithms would be ideal for very complicated prediction problems. A notable proportion of people in this contemporary world have Diabetes worldwide, and the most shocking thing is that most patients do not know this. You do not yet know what type of Diabetes you have. If diabetes types are identifiable in an early stage and can be treated properly, Diabetes may be controlled and cannot be alarmed. Several practical experiments have been done to enable people with ML to work out the best way for them to anticipate human diseases. So, ML can be used to perform this analysis very conveniently. Our task was to build the Model with Bagging and Stacking produced more accurate results than the existing models. There were several drawbacks during the analysis, such as restricted data. The most important thing is missing Information handled using the WEKA tool with ML techniques, instance restriction, and features.

The study was conducted right, given these limitations. However, the trial can also be carried out through other advanced ML algorithms, for example, ANFIS and advanced ensemble methods which combine the Neural Network, the Fuzzy System, etc. with more instances and more attributes.

REFERENCES

1. Insulin and Diabetes, Diabetes UK (2019). https://www.diabetes.org.uk/guide-todiabetes/ managing-your-diabetes/treating-your-diabetes/insulin. Accessed 30 May 2020
2. Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag.Big Data.Dec 2015.277-287.http://doi.org/10.1089/big.2015.00
3. T.G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, Multiple Classi_er Systems. First International Workshop,MCS 2000, Cagliari, Italy, volume 1857 of Lecture Notes in Computer Science,pages 1{15. Springer-Verlag, 2000. [48]
4. L.I. Kuncheva. Combining Pattern Classifers: Methods and Algorithms. Wiley-Interscience, New York, 2004. [116]
5. Han L, Diao L, Yu S, et al. The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. Cancer Cell. 2015;28(4):515-528. doi:10.1016/j.ccell.2015.08.013
6. Polat K, Güneş S. A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted preprocessing and AIRS. Comput Methods Programs Biomed. 2007;88(2):164-174. doi:10.1016/j.cmpb.2007.07.013
7. Duygu Çalişir and Esin Doğantekin. 2011. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Syst. Appl. 38, 7 (July, 2011), 8311–8315. DOI:https://doi.org/10.1016/j.eswa.2011.01.017
8. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data. 2015;3(4):277-287. doi:10.1089/big.2015.0020
9. Kavakiotis, Ioannis & Tsave, Olga & Salifoglou, Athanasios & Maglaveras, N. & Vlahavas, I. & Chouvarda, Ioanna. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal. 15. 10.1016/j.csbj.2016.12.005.
10. Eleni I. Georga, Vasilios C. Protopappas, Diego Ardigò, Demosthenes Polyzos, and Dimitrios I. Fotiadis.Diabetes Technology & Therapeutics.Aug 2013.634-643.http://doi.org/10.1089/dia.2012.0285
11. Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Comput Methods Programs Biomed. 2011;104(3):443-451. doi:10.1016/j.cmpb.2011.03.018
12. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques, 3rd edn, pp.370–382. Morgan Kaufmann, Burlington (2011)
13. Wolpert, David. (1992). Stacked Generalization. Neural Networks. 5. 241-259. 10.1016/S0893-6080(05)80023-1.
14. Witten, I., Frank, E., Hall, M.: Data Mining Practical Machine Learning Tools and Techniques, 3rd edn, pp. 166–580. Morgan Kaufmann, Burlington (2011).
15. Yue, C., Xin, L., Kewen, X., and Chang, S. (2008). "An intelligent diagnosis to type 2 diabetes based on QPSO algorithm and WLS-SVM," in Proceedings of the 2008 IEEE International Symposium on Intelligent Information Technology Application Workshops, Washington, DC. doi: 10.1109/IITA.Workshops.2008.36
16. N.H. Barakat, A.P. Bradley, M.N.H. Barakat. (2010). Intelligible SupportVector Machines for Diagnosis of Diabetes Mellitus. IEEE Transactions on Information Technology in Biomedicine. 14(4), pp.1114-1120.
17. Larabi-Marie-Sainte, S.; Aburahmah, L.; Almohaini, R.; Saba, T. Current Techniques for Diabetes Prediction: Review and Case Study. Appl. Sci. 2019, 9, 4604.
18. Nguyen, BK, Patel, NM, Arianpour, K, et al. Characteristics and management of sinonasal paragangliomas: a systematic review. Int Forum Allergy Rhinol. 2019; 9: 413– 426.
19. Bradley, P., & Fayyad, U. (1998). Refining initial points for k-means clustering. Proc. 15th International Conf. on Machine Learning.
20. Chris Ding and Xiaofeng He. 2004. K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 29. DOI:https://doi.org/10.1145/1015330.1015408

AUTHORS PROFILE

Dr. Jyoti Verma is currently working as Assistant Professor in the Department of Computer Science in J.C. Bose University of Science and Technology. She completed her Ph.D in the area of Information Retrieval in 2011. She has over 15 years of teaching experience with almost 10 years of research experience. She has over 30 publications in her name. her current areas of interest are Information Retrieval and Big Data Analytics.

Dr Jyoti, Assistant Professor (CE)
YMCA University of Science & Technology
Sector 6, Mathura Road, Faridabad - 121006
Phone: +91-9910341139

Peri Arjun is student at JC Bose YMCA University of Science and Technology, Faridabad where he is currently pursuing Master of Technology course in Computer Science with specialization in Computer Networking. His research interest lies in the area of Machine Learning and Artificial Intelligence with special focus on healthcare domain. He completed his Bachelor of Technology in Computer Science and Engineering from Lingaya's University Faridabad with first division.

Peri Arjun
periarjun@gmail.com
+91-9871303495