International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET2293173

Survey Paper on Diabetes Risk Prediction using Machine Learning Algorithm

Shalinee Bhondekar*[1], Dr. Shalini Sahay[2]
*[1]M. Tech. Scholar, Department of ECE, SIRT, Bhopal, India
[2]Associate Professor, Department of ECE, SIRT, Bhopal, India

Article Info
Volume 9, Issue 3

Page Number : 544-550

Publication Issue :
May-June-2022

Article History
Accepted : 10 June 2022
Published: 24 June 2022

ABSTRACT

Diabetes Mellitus (DM) is a chronic, lifelong metabolism disorder. It affects the ability of the body system to use the energy found in food. The improper management of the disease will lead to Heart disease, kidney disease, eye disease, nerve disease and pregnancy complications. Classification model helps physicians to improve their prognosis, diagnosis or treatment planning procedures. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

Keywords : Diabetic Dataset, Classification, Machine Learning

I. INTRODUCTION

Diabetes has been most prevalent disease in the world which is not confined to any specific age group.[1] Threatening all age group make it more worrisome for the doctor and Government all over the globe. We need pensive attention and research to eradicate this menace from the life of people. Deficiency of insulin in human body results in excess glucose (sugar) level because insulin the main agent to regulate the sugar level in our body system.[9] The fundamental reason explored by the medical science for such phenomenon is that the absence of insulin enhances the carbohydrate rich food conversion into glucose which is accumulated in the blood , enhancing the sugar level more than normal. Diabetes is responsible for other diseases like heart disease, kidney damage, blindness, blood vessel damage or it may even cause death. Presently there is no such permanent treatment in medical science for diabetes; however it can be controlled by proper control over diet and regular physical exercise [1].

Detection of diabetes in advance and in early stages is very important for better prevention and cure. The

researchers have been founded several techniques to diagnose the diabetes by contributing their continuous efforts and knowledge There have been many techniques used in Neural Networks that were applied over diabetes diagnosis to help physicians. An uncommon approach utilized by Polat and Gunes is Principal element Associate in Nursingalysis and reconciling Neuro Fuzzy reasoning System (ANFIS). In initial stage PCA reduces options for higher accuracy and in second stage ANFIS is employed for classification of polygenic disorder knowledge set [4]. Equally LDA is employed for feature reduction of polygenic disorder knowledge with classification by ANFIS within the second stage [3]. Manjeevan Seera et al have planned 3 methods- Fuzzy Min-Max Neural network (FMM), Classification and Regression Tree (CART) and Random Forest Model (RF) over 3 totally different medical datasets- carcinoma Wisconsin, Pima Indians polygenic disorder and Liver Disorders. Among them FMM-CART-RF generates higher results compared to FMM and FMM-CART results. The conception of fuzzy modeling has been used effectively in diagnosis systems. within the work planned by Sean N. Ghazavi et al 3 strategies namely- fuzzy k-nearest neighbor rule, a fuzzy clustering-based modeling and reconciling Network-based Fuzzy reasoning System (ANFIS) were used over Wisconsin carcinoma and Pima Indians polygenic disorder dataset. The classification accuracy obtained is ninety seven.17% and 77.65% severally. many different mechanisms like k-means agglomeration and C4.5 rule are used for polygenic disorder prediction. K-means agglomeration enforced by maori hen tool is used for pattern extraction. The resultant output is given to C4.5 call tree used for classification exploitation k-fold cross validation is employed to separate samples into totally different categories. Preprocessing of knowledge is incredibly vital to enhance the accuracy and to scale the info in same vary of values, that is achieved by Principal element Analysis. The classification of resultant knowledge is completed by Feed forward Higher Order Neural network exploitation k-fold cross validation. Support Vector Machines have proved exceptional success within the space of diagnosis. Within the work planned by V. Anuja kumari et al have classified the polygenic disorder knowledge exploitation SVM classifier with RBF kernel. The classification accuracy obtained is seventy eight.

The classification task consists in distribution instances from a given domain, delineated by a group of distinct price attributes, into a group of categories, which might be thought-about values of selected distinct target attribute known as target conception. The proper category labels area unit typically unknown, however area unit provided for a set of the domain. It is wont to produce the instance from identical domain, describes by identical set of attributes. This follows the overall assumption of inductive learning of the classification task is that the commonest mental representation [2].

The classification Techniques disperse a category to assortment of knowledge records having specific attributes and its values, therefore the classification techniques in aid is applied for nosology functions. A classification model receives a group of connected attribute values, like clinical measurements and offers a category of knowledge records as output.

In our study we tend to propose a brand new technique by adding Particle swarm improvement with Multi layer Neural Network (PSONN) to classify designation of Pima Indians polygenic disorder dataset taken from UCI machine learning repository named as Particle swarm improvement Neural Network (PSONN). the basic plan for this rule is that at the start stage of finding out the optimum parameter of PSO by meted out random tuples from the input dataset, then PSO is used to accelerate the coaching speed. once the fitness operate price has not amendment for a few generations, the looking method is switched to gradient downhill looking per heuristic information. Finally PSO combined with Neural Network is employed to classify the samples that be 2 classes either '0' for negative (Not Diabetic) or '1' for positive (Diabetic).

II. LITERATURE REVIEW

Nikos Fazakis et al. [1], asteady rise has been seen in the level of older individuals whowant and are as yet ready to add to society. Consequently, exiting the workforce or exit from the work market, because of wellbeing related issues, represents a signi cant issue. These days, because of mechanical advances and different information from various populaces, the gamble factors examination and medical problems screening are moving towards mechanization. In the setting of this work, a specialist driven, IoT empowered inconspicuous clients wellbeing, prosperity and practical capacity checking structure, engaged with AI apparatuses, is proposed. Diabetes is a high-predominance ongoing condition with hurtful ramifications for the personal satisfaction and high death rate for individuals around the world, in both created and non-industrial nations. Thus, its extreme effect on people's life, e.g., individual, social, working, can be significantly decreased assuming that early identification is conceivable, however most exploration works in this eld fall flat to give a more customized approach both in the demonstrating and forecast process. Toward this path, our planned framework concerns diabetes risk expectation in which speci c parts of the Knowledge Discovery in Database (KDD) process are applied, assessed and consolidated. Speci cally, dataset creation, highlights choice and classi cation, utilizing different Supervised Machine Learning (ML) models are thought of. The outfit WeightedVotingLRRFs ML model is proposed to work on the forecast of diabetes, scoring an Region Under the ROC Curve (AUC) of 0:884. Concerning the weighted democratic, the ideal loads are assessed by their comparing Sensitivity and AUC of the ML model in light of a bi-objective hereditary calculation. Likewise, a near report is introduced among the Finnish Diabetes Risk Score (FINDRISC) what's more, Leicester risk score frameworks and a few ML models, utilizing inductive and transductive learning. The tests were led utilizing information removed from the English Longitudinal Study of Aging (ELSA) data set.

Naveen Kishore et al. [2], diabetes is considered as one of the deadliest and persistent illnesses which causes an increment in glucose. Numerous entanglements happen if diabetes stays untreated and unidentified. The monotonous distinguishing measure brings about visiting of a patient to a symptomatic focus and counseling specialist. However, the ascent in AI approaches tackles this basic issue. The thought process of this investigation is to plan a model which can visualize the probability of diabetes in patients with greatest exactness. Thusly three AI characterization calculations specifically Decision Tree, SVM and Naive Bayes are utilized in this analysis to identify diabetes at a beginning phase. Analyses are performed on Pima Indians Diabetes Database (PIDD) which is sourced from UCI AI storehouse. The exhibitions of the multitude of three calculations are assessed on different estimates like Precision, Accuracy, F-Measure, and Recall. Exactness is estimated over effectively and erroneously ordered cases. Results got show Naive Bayes beats with the most elevated precision of 76.30% relatively different calculations. These outcomes are confirmed utilizing Receiver Operating Characteristic (ROC) bends in a legitimate and methodical way.

Muhammad Azeem Sarwar et al. [3], there are a few AI strategies that are utilized to perform prescient investigation over enormous information in different fields. Prescient examination in medical services is a difficult undertaking in any case can help professionals settle on huge information educated ideal choices about tolerant's wellbeing and therapy. This paper talks about the prescient investigation in medical care, six diverse AI calculations are utilized in this examination work. For analyze reason, a dataset of patient's clinical record is gotten and six distinctive AI calculations are applied on the dataset. Execution and precision of the applied calculations is examined and

analyzed. Correlation of the diverse AI methods utilized in this examination uncovers which calculation is most appropriate for forecast of diabetes. This paper intends to help specialists and experts in early expectation of diabetes utilizing AI strategies.

Rao G.A. et al. [4], extraction of complex head and hand developments alongside their continually changing shapes for acknowledgment of communication via gestures is viewed as a troublesome issue in PC vision. This paper proposes the acknowledgment of Indian communication through signing motions utilizing an incredible man-made consciousness device, convolutional neural organizations (CNN). Selfie mode ceaseless gesture based communication video is the catch technique utilized in this work, where a conference weakened individual can work the SLR versatile application freely. Because of non-accessibility of datasets on portable selfie communication through signing, we started to make the dataset with five distinct subjects performing 200 signs in 5 diverse survey points under different foundation conditions. Each sign involved for 60 edges or pictures in a video. CNN preparing is performed with 3 distinctive example estimates, each comprising of different arrangements of subjects and survey points. The leftover 2 examples are utilized for testing the prepared CNN. Distinctive CNN models were planned and tried with our selfie gesture based communication information to acquire better precision in acknowledgment. We accomplished 92.88% acknowledgment rate contrasted with other classifier models wrote about the equivalent dataset.

Quan Zou et al. [5], the astounding advances in biotechnology and wellbeing sciences have prompted a critical creation of information, for example, high throughput hereditary information and clinical data, produced from huge Electronic Health Records (EHRs). To this end, use of AI and information mining strategies in biosciences is as of now, like never before previously, imperative and essential in endeavors to change shrewdly all accessible data into significant information. Diabetes mellitus (DM) is characterized collectively of metabolic issues applying huge tension on human wellbeing around the world. Broad exploration in all parts of diabetes (determination, etiopatho physiology, treatment, and so on) has prompted the age of tremendous measures of information. The point of the current examination is to direct a methodical survey of the uses of AI, information mining strategies and instruments in the field of diabetes research regarding a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management with the main class having all the earmarks of being the most famous. Wide scopes of AI calculations were utilized. As a rule, 85% of those utilized were portrayed by directed learning draws near and 15% by solo ones, and all the more explicitly, affiliation rules. Backing vector machines (SVM) emerge as the best and broadly utilized calculation. Concerning the kind of information, clinical datasets were predominantly utilized. The title applications in the chose articles project the convenience of removing important information prompting new speculations focusing on more profound arrangement and further examination in DM.

III. METHODOLOGY

Learning

The main property of an ML is its capability to learn. Learning or preparing is a procedure by methods for which a neural system adjusts to a boost by making legitimate parameter modifications, bringing about the generation of wanted reaction. Learning in an ML is chiefly ordered into two classes as [9].

- Supervised learning
- Unsupervised learning

Supervised Learning

Regulated learning is two stage forms, in the initial step: a model is fabricated depicting a foreordained

arrangement of information classes or ideas. The model developed by investigating database tuples portrayed by traits. Each tuple is expected to have a place with a predefined class, as dictated by one of the qualities, called to have a place with a reclassified class, as controlled by one of the traits called the class name characteristic. The information tuple are dissected to fabricate the model all things considered from the preparation dataset.

Unsupervised learning

It is the kind of learning in which the class mark of each preparation test isn't knows, and the number or set of classes to be scholarly may not be known ahead of time. The prerequisite for having a named reaction variable in preparing information from the administered learning system may not be fulfilled in a few circumstances.

Data mining field is a highly efficient techniques like association rule learning. Data mining performs the interesting machine-learning algorithms like inductive-rule learning with the construction of decision trees to development of large databases process. Data mining techniques are employed in large interesting organizations and data investigations. Many data mining approaches use classification related methods for identification of useful information from continuous data streams.

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule differentiates the classification of unknown data point because of closest neighbor whose class is known. The nearest neighbor is calculated based on estimation of k that represents how many nearest neighbors are taken to characterize the data point class. It utilizes more than one closest neighbor to find out the class where the given data point belong termed as KNN. The data samples are required in memory at run time called as memory-based technique. The training points are allocated weights based on their distances from the sample data point. However, the computational complexity and memory requirements remained key issue. For addressing the memory utilization problem, size of data gets minimized. The repeated patterns without additional data are removed from the training data set.

Naive Bayes Classifier

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

Support Vector Machine

SVM are used in many applications like medical, military for classification purpose. SVM are employed for classification, regression or ranking function. SVM depends on statistical learning theory and structural risk minimization principal. SVM determines the location of decision boundaries called hyper plane for optimal separation of classes as described in figure 1.4. Margin maximization through creating largest distance between separating hyper plane and instances on either side are employed to minimize upper bound on expected generalization error. Classification accuracy of SVM not depends on dimension of classified entities. The data analysis in SVM is based on convex quadratic programming. It is expensive as quadratic programming methods need large matrix operations and time consuming numerical computations.

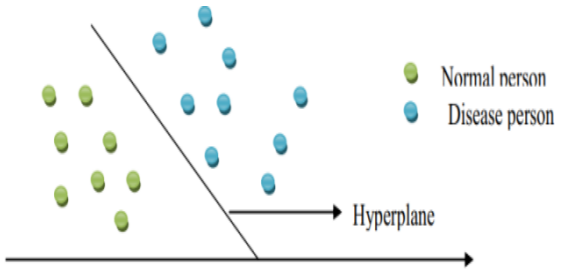

Fig. 1: Support Vector Classification

Particle Swarm Optimization

The notation is used in Particle swarm optimization are as follows:

Xid :Component in dimension d of the ith particle of swarm

Vid :The particle velocity of particle I in dimension d

PBi : the best position achieved so far by particle i

GB : The best global best position

C1,C2 : Constant weight factors

W: The inertia weight

r1,r2: Random factors in [0,1] interval

Vmin : The minimum velocity value of particle

Vmax: The maximum velocity value of particle

yi : The fitness value of particle i

PSO is a population base stochastic optimization technique inspired by the social behavior of swarm, such as bird flocking or fish schooling, to obtain a promising position to achieve certain objectives. The PSO algorithm works by having a population (called a swarm) of candidate solution (called particles).Each particle in a population has a fitness value computed from a fitness function and each particle has a position, and move based on an updated velocity according to few simple formula. The movements of the particles are guided by their own best known position in the search space as well as the entire swarm's best known position. The particle movements are directed by the position vector and velocity vector of each particle. In the n-dimensional space, the vector and velocity vector of the ith particle position are represented as Xi=[xi1,xi2,x3i,xi4,…….xin] and Vi=[vi1,vi2,……..,vin] respectively, where xid is a binary bit, i=1,2,…..,m (m is the number of particles). The record of the position of the previous best performance of the neighborhood is GBi=[gbi1, gbi2,…….gbin]. The particle velocity and position is updated and based on Eqs. (A1) and (A2), respectively.

Vidnew = w x vidold + c1r1(pbidold –xidold) + c2r2 (gbdold- xidold) , d=1,2 ely D

Xidnew = xidold + vidnew , d=1,2……elyN

Where c1 and c2 are the positive constant values between 0 and 4, indication the cognitive and the social learning factors, respectively. The inertia weight (w has a value between 0.4 and 0.9 and r1 and r2 are uniformly distributed with the numbers between 0 and 1. The values of the velocities are between vmin and vmax , N is the size of the swarms.

Feed Forward Back propagation Neural Network

Neural networks are predictive model that have ability to learn, analyses, organize the data and predict test results accordingly. Among several kinds of neural networks, feed forward neural network is usually employed in medical diagnosis applications and others. These networks are trained by a set of patterns called training set, whose outcome is already known. In our study Feed Forward NN consists of input, hidden and an output layer, and the data operates in forward direction, and the error is back propagated to update the weights at every epoch in order to reduce errors.

IV. CONCLUSION

Diabetes is metabolic disease that arises due to high blood glucose level in body. Insulin is not sufficient enough in body of diabetic patients to regulate the sugar level. Further several other diseases also arise from diabetes which is hazardous to health. It is necessary to detect such a serious health issue as early as possible Diabetes is cause of various diseases in human body. To make diabetes diagnosis easier for Physicians, there have been several methods

employed. The diabetic data set is tested with selected classification algorithm.

V. REFERENCES

[1] Nikos Fazakis, Otilia Kocsis , Elias Dritsas, Sotiris Alexiou, Nikos Fakotakis, and Konstantinos Moustakas, “Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction”, IEEE Access 2021.

[2] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy, “Prediction of Diabetes using Machine Learning Classification Algorithms”, International Journal of Scientific & Technology Research Volume 9, Issue 01, January 2020.

[3] Muhammad Azeem Sarwar, 2Nasir Kamal, 3Wajeeha Hamid,4Munam Ali Shah, “Prediction of Diabetes Using Machine Learning Algorithms in Healthcare”, 24th International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2019.

[4] Rao G.A., Syamala K., Kishore P.V.V., Sastry A.S.C.S. ., “Deep convolutional neural networks for sign language recognition”, 2018, International Journal of Engineering and Technology(UAE) ,Vol: 7, Issue 5, pp: 62 to 70.

[5] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju and Hua Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques”, Springer, 2018.

[6] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, “Machine learning on big data: Opportunities and challenges,” Neuro computing, vol. 237, pp. 350–361, May 2017.

[7] J. B. Heaton, N. G. Polson, and J. H. Witte, “Deep learning for finance: deep portfolios,” Appl. Stoch. Model. Bus. Ind., vol. 33, no. 1, pp. 3–12, Jan. 2017.

[8] Reddy S.S., Suman M., Prakash K.N. ., “Micro aneurysms detection using artificial neural networks”, 2018, Lecture Notes in Electrical Engineering ,Vol: 434 ,Issue 3, pp: 409 to 417.

[9] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., “Machine Learning and Data Mining Methods in Diabetes Research”, Computational and Structural Biotechnology Journal 15, 104–116, 2017.

[10] Majid Ghonji Feshki and Omid Sojoodi Shijan, “Improving the Heart Disease Diagnosis by Evolutionary Algorithm of PSO and Feed Forward Neural Network”, International paper on IEEE 2016.

[11] L. Hermawanti, "Combining of Backward Elimination and Naive Bayes Algorithm To Diagnose Breast Cancer", Momentum, vol. 11, no. 1, pp. 42-45, 2015.

[12] O.S. Soliman, E. Elhamd, “Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine”, IEEE 2014.

[13] K. Saxena, Z. Khan, S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbour Algorithm", International Journal of Computer Science Trends and Technology (IJCST), 2014.

[14] L. Hermawanti, S.G. Rabiha, "Combining of Backward Elimination and K-Nearest Neighbor Algorithms To Diagnose Heart Disease", Prosiding SNST Ke-5 Fakultas Teknik Universitas Wahid Hasyim, pp. 1-5, 2014.

[15] R.A. Vinarti, W. Anggraeni, "Identification of Prediction Factor Diagnosis of Breast Cancer Rates with Stepwise Binary Logistic Regression Method", Jurnal Informatik, vol. 12, no. 2, pp. 70-76, November 2014.

[16] Muhammad Waqar Aslam, Zhechen Zhu and Asoke Kumar Nandi, “Feature generation programming with comparative partner selection for diabetes classification”, “Expert Systems with Applications”, 5402-5412, IEEE 2013.

Cite this article as :

Shalinee Bhondekar, Dr. Shalini Sahay, "Survey Paper on Diabetes Risk Prediction using Machine Learning Algorithm", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 3, pp. 544-550, May-June 2022. Available at doi : https://doi.org/10.32628/IJSRSET2293173

Journal URL : https://ijsrset.com/IJSRSET2293173