International Journal of Engineering and Techniques - Volume 3 Issue 4, July-Aug 2017

RESEARCH ARTICLE OPEN ACCESS

DESIGN AND DEVELOPMENT OF NOVEL EFFECTIVE HYBRID PREDICTION FRAMEWORK FOR DIABETES USING SIMPLE K-MEANS AND M-TREE

[1*]AMIREDDY SRINISH REDDY, SANJAY PACHOURI[2]

[1]RESEARCH SCHOLAR, DEPT OF CSE, SUNRISE UNIVERSITY, ALWAR, RAJASTHAN, INDIA
[2]RESEARCH SUPERVISOR, DEPT OF CSE, SUNRISE UNIVERSITY, ALWAR, RAJASTHAN,INDIA

Abstract—

Diabetes has a very high rate everywhere throughout the world. For the prevention and treatment of diabetes, early detection is demanded. These days, data mining methods are increasing expanding significance in the medicinal diagnosis field by their classification ability. In this paper, a hybrid prediction demonstrates is proposed to help the diagnosis of diabetes. In the proposed model, K-means is utilized for data reduction with the J48 M-TREE as a classifier for classification. With the end goal to get the exploratory outcome, we utilized the Pima Indians Diabetes Dataset from the UCI Machine Learning Repository. The outcome demonstrates that the proposed model has achieved better precision contrasted with different past investigations that mentioned in the writing. Based on the outcome, it very well may be demonstrated that the proposed model would be useful in diabetes diagnosis.

Keywords-diabetes diagnosis; data mining; classification; K-means; M-TREE

I. INTRODUCTION

Diabetes is one of the most common diseases as of late, and its worldwide predominance is developing quickly. It is a general term for heterogeneous aggravations of digestion for which the fundamental finding is chronic hyperglycemia. The reason is either debilitated insulin secretion or weakened insulin action or both [1]. The chronic hyperglycemia of diabetes is related with long-term harm, dysfunction, and disappointment of various organs, particularly the eyes, kidneys, nerves, heart, and veins. By far most of diabetes can be separated into two classifications, viz. Type 1 and. The reason for Type 1 diabetes is an outright lack of insulin secretion. On the other hand, diabetes is considerably more pervasive, and the reason is a combination of protection from insulin action and an insufficient compensatory insulin secretory response [2]. The most common form of diabetes will be diabetes [3].

As indicated by the six edition of IDF (International Diabetes Federation) Diabetes Atlas, a surprising 382 million individuals are evaluated to have diabetes, with sensational increments found in nations everywhere throughout the world and diabetes constitutes the greater part of all diabetes [4]. As a consequence, diabetes is a serious medical problem for the entire world. On the off chance that we could analyze and anticipate diabetes as ahead of schedule as would be prudent, millions of lives may be spared.

Along with the incredible advancement of information innovation, we could make utilization of the tremendous measure of data in the medicinal services industry to help specialists diagnosing diabetes. Data mining could anticipate the future by displaying. As of late, a substantial number of computational techniques and apparatuses for data investigation are accessible. Data mining has been generally connected to the medicinal field and assumed a critical job in restorative research. Subsequently, this paper proposes a hybrid diagnosis display which could foresee diabetes in using different data mining techniques. This model could help specialists and restorative professionals in making decisions and enhance symptomatic exactness.

II. BACKGROUND

In this section of the paper, we will examine data mining and a portion of the data mining instruments and strategies.

A. Data mining

We face a daily reality such that immense measures of data are gathered every day. The traditional strategy for transforming data into knowledge depends on manual data investigation. As data volumes develop quickly, this form of data examination is moderate, costly, and abstract. The traditional technique is winding up totally unrealistic in numerous fields and couldn't address the issue for data examination [5]. Data mining, otherwise called knowledge revelation in databases (KDD) could address this issue by giving instruments to find knowledge from data. Data mining is the way toward finding intriguing examples and knowledge from a lot of data. The data sources can incorporate databases, data distribution centers, the Web, other information stores, or data that are gushed into the framework powerfully [6].

Amid the previous decades, data mining has been connected to an assortment of territories, for example, marketing, back (particularly venture), misrepresentation detection, assembling, telecommunications and numerous logical fields, including the investigation of therapeutic data [5]. As medicinal data volumes develop drastically, there is a developing weight for efficient data examination to remove valuable, task-situated information from the gigantic measures of data [7]. Such information may assume a vital job in future medicinal decision-making.

B. Data mining instruments

For the implementation of the proposed model, it is important to make utilization of a few data mining instruments. An efficient data mining device could help us in transforming the immense data Into valuable information. In the previous couple of years, there are many open - source data mining devices and software accessible for utilize, for example, Waikato Environment for Knowledge Analysis (WEKA), TANAGRA, Rapid excavator, Orange, KNIME and so forth. Among every one of these data mining devices, WEKA is one of the most prevalent and completely functional devices [8]. In this way, we chose to utilize WEKA as our data mining device.

WEKA is a Java-based PC program for data mining and machine realizing which was initially created at the University of Waikato in New Zealand. WEKA offers four options for data mining, viz. The experimenter, command-line interface (CLI), Explorer and Knowledge stream. WEKA contains a huge collection of the most up to date data mining and machine learning algorithms written in Java. It bolsters a decent variety of standard tasks for data

mining: data preprocessing, bunching, classification, regression, visualization and highlight selection [9].

C. Data mining techniques

Data mining is anticipated to be one of the most revolutionary developments of the following decades. In actuality, it was picked as one of 10 developing advancements that will change the world by the MIT Technology Review [10]. Specialists have been vivaciously growing new data mining procedures. Data mining approachs ought to consider issues, for example, data vulnerability, clamor, and inadequacy. A few data mining strategies investigate how client indicated measures can be utilized to survey the intriguing quality of found examples and also control the revelation procedure [11]. In this section, two of the common data mining strategies that would be utilized in the proposed model are examined.

K-means grouping calculation: K-means has a rich and various history as it was freely found in various logical fields by Steinhaus (1956), Lloyd (proposed in 1957, distributed in 1982), Ball and Hall (1965) and McQueen (1967). In spite of the fact that K-means was first proposed more than 50 years prior, it is as yet one of the most generally utilized bunching algorithms. Simplicity of implementation, effectiveness, straightforwardness and observational achievement are the fundamental reasons for its fame [12]. The methodology of K-means pursues a simple method to arrange a given data set through a specific number of bunches (accept K groups) settled apriori. K-means calculation randomly picks K objects, speaking to the K beginning bunch focus. The accompanying advance is to take each guide belonging toward a given data set and partner it to the closest focus dependent on the closeness of the question with the bunch focus using Euclidean separation. At the point when every one of the items are disseminated, the time has come to recalculate new K bunch focuses. The procedure would be rehashed until there is no adjustment in K group focuses. K-means goes for limiting a target function known as the squared mistake function that is given by the accompanying [13].

$$J(C) = \sum_{k=1}^{k} \sum_{x_i \in C_k} \left\| x_i - \mu_k \right\|^2 \qquad (1)$$

M-TREE calculation: Over the most recent couple of years, an incredible number of algorithms have been created for classification based data mining. A M-TREE is a critical classification calculation in data mining. The fundamental preferred standpoint of M-TREE algorithms is that they are anything but difficult to construct and the subsequent trees are promptly interpretable. It is commonly utilized in various territories. Analysts have built up an assortment of M-TREE algorithms over some undefined time frame with upgrade in performance and capacity to handle diverse kinds of data. Prevalent M-TREE algorithms including ID3, CART, C4.5, C5.0, J48 and so forth. C4.5 is produced by Ross Quinlan. It is an extension of Quinlan's prior ID3 calculation [14]. C5.0 and J48 are the enhanced versions of C4.5 algorithms. In the WEKA data mining apparatus, a J48 calculation is an open source Java implementation of the C4.5 calculation. WEKA furnishes various options related with tree pruning. J48 classifier makes a parallel tree. By using this strategy, a tree

is constructed to show the classification procedure. Once the tree is assembled, it is connected to each tuple in the database and results in the classification for the tuple [15].

III. LITERATURE REVIEW

As of late, prescient classification is one of the most fundamental and essential tasks in data mining and machine learning. Its application to the restorative diagnosis has gotten a strong lift because of sincere research exercises in the medicinal enormous data field. Numerous specialists have featured the capability of prescient classification to give decision support to specialists and therapeutic professionals. In the course of the most recent couple of years, a lot of research has been conducted on various datasets to prescient diabetes. A considerable lot of them demonstrated great classification exactness.

Diabetes has transformed into an overall pandemic that puts a grave load on human services frameworks, especially in creating nations [4]. On a worldwide stage, the aggregate number of diabetic patients is assessed to increment from 171 million to 366 million out of 2000 and 2030, individually [39]. T2D is an advanced stage in which the body turns out to be hardened to ordinary impacts of insulin and gradually loses the ability to produce enough insulin in the pancreas. It is vital for persons 45 years, with a BMI 25 kg/m2, to encounter screening to identify pre-diabetes and diabetes [2]. Besides, hypertension, which is typically alluded to as systolic circulatory strain 140 mmHg and diastolic pulse 90 mmHg, is an ordinary long-term disease that by and by effects 77 million Americans [40–42]. It is a critical risk component for the deadly cardiovascular diseases, creating heart disappointment in 91% of cases; it is available in 69% of persons who endure their first heart attack and in 77% of those having their first stroke [41]. Past investigations have demonstrated firm positive relationships among pulse, the risk of cardiovascular diseases, and mortality [43,44]. Together, hypertension and diabetes are stroke risk factors, yet they can be maintained a strategic distance from if people take a sound eating routine and also physical exercise each day [45]. Therefore, later on, a prediction show that advises individuals on the possibility of diabetes and hypertension is required, and it would allow them to take preemptive action. The machine learning algorithms can be utilized to analyze diabetes and hypertension that depends on the current condition of patients.

A few examinations have demonstrated a positive effect of the application of machine learning for diabetes classification. Patil et al. proposed HPM for T2D [13]. The proposed model consists of a K-means calculation to expel inaccurately characterized case and C4.5 to group the diabetes dataset. The Pima Indian dataset and k-overlap cross-validation are used. The outcome uncovered that the HPM demonstrated the most elevated precision, as high as 92.38%, among different strategies. Wu et al. used a HPM for anticipating T2D [14]. The model consists of an enhanced K-means and the calculated regression display. The enhanced K-means calculation was utilized to confiscate off base grouped data, later the strategic regression calculation was utilized to characterize the rest of the data. The discoveries shown that the proposed model demonstrated more noteworthy prediction precision as contrasted and past work. Past writing thought about the performance of calculated regression, counterfeit neural networks (ANNs), and M-TREE models for foreseeing diabetes or prediabetes utilizing common risk factors [15]. The dataset was accumulated from Guangzhou, China, and 735 patients were approved as having diabetes or prediabetes, while 752 were ordinary controls. The discoveries demonstrated that the best classification precision when contrasted with other model is appeared by M-TREE display C5.0. At last, past writing proposed a machine learning model to anticipate the predominance of diabetes and hypertension, with a dataset having 13,647,408 restorative

records for different ethnicities in Kuwait [16]. The classification models, for instance, calculated regression, K-Nearest Neighbors (KNN), Multi-factor Dimensionality Reduction (MDR), and Support Vector Machines (SVM), were utilized and shown critical finding on anticipating diabetes and hypertension. Additionally, the study gathered that ethnicity is a basic element for foreseeing diabetes.

Moreover, a few examinations have been conducted and uncovered that the machine learning algorithms give early prediction and in addition treatment for hypertension. Koren et al. researched the benefit of machine learning for treatment of hypertension [17]. They utilized machine learning strategies to recognize determinants that add to the achievement of hypertension tranquilize treatment on an enormous arrangement of patients. The outcome demonstrated that a completely connected neural network could accomplish AUC as much as 0.82. The consequence of their study demonstrated that machine learning algorithms can give the as of late, prescient classification is one of the most fundamental and essential tasks in data mining and machine learning. Its application to the medicinal diagnosis has gotten a strong lift because of sincere research exercises in the therapeutic enormous data field. Numerous analysts have featured the capability of prescient classification to give decision support to specialists and restorative professionals. In the course of the most recent couple of years, a lot of research has been conducted on various datasets to prescient diabetes. Huge numbers of them indicated great classification exactness.

Diabetes has transformed into an overall pandemic that puts a grave load on medicinal services frameworks, especially in creating nations [4]. On a worldwide stage, the aggregate number of diabetic patients is evaluated to increment from 171 million to 366 million out of 2000 and 2030, individually [39]. T2D is an advanced stage in which the body turns out to be solid to typical impacts of insulin and gradually loses the ability to create enough insulin in the pancreas. It is fundamental for persons 45 years, with a BMI 25 kg/m2, to encounter screening to recognize pre-diabetes and diabetes [2]. Moreover, hypertension, which is typically alluded as systolic pulse 140 mmHg and diastolic circulatory strain 90 mmHg, is an ordinary long-term disease that by and by effects 77 million Americans [40– 42]. It is a critical risk component for the deadly cardiovascular diseases, creating heart disappointment in 91% of cases; it is available in 69% of persons who endure their first heart attack and in 77% of those having their first stroke [41]. Past examinations have indicated firm positive relationships among circulatory strain, threat of cardiovascular diseases, and mortality [43,44]. Together, hypertension and diabetes are stroke risk factors, yet they can be maintained a strategic distance from if people take a sound eating routine and in addition physical exercise each day [45]. Therefore, later on, a prediction demonstrate that informs individuals on the shot of diabetes and hypertension is required, and it would allow them to take preemptive action. The machine learning algorithms can be utilized to analyze diabetes and hypertension that depends on the current condition of patients.

A few investigations have demonstrated a positive effect of the application of machine learning for diabetes classification. Patil et al. proposed HPM for T2D [13]. The proposed model consists of a K-means calculation to evacuate erroneously arranged case and C4.5 to characterize the diabetes dataset. The Pima

Indian dataset and k-overlay cross-validation are used. The outcome uncovered that the HPM demonstrated the most noteworthy exactness, as high as 92.38%, among different strategies. Wu et al. used a HPM for anticipating T2D [14]. The model consists of an

enhanced K-means and the calculated regression demonstrate. The enhanced K-means calculation was utilized to confiscate off base bunched data, later the strategic regression calculation was utilized to order the rest of the data. The discoveries shown that the proposed model demonstrated more noteworthy prediction exactness as contrasted and past work. Past writing looked at the performance of calculated regression, counterfeit neural networks (ANNs), and M-TREE models for foreseeing diabetes or prediabetes utilizing common risk factors [15]. The dataset was assembled from Guangzhou, China, and 735 patients were approved as having diabetes or prediabetes, while 752 were ordinary controls. The discoveries demonstrated that the best classification exactness when contrasted with other model is appeared by M-TREE show C5.0. At last, past writing proposed a machine learning model to anticipate the pervasiveness of diabetes and hypertension, with a dataset having 13,647,408 therapeutic records for different ethnicities in Kuwait [16]. The classification models, for instance, strategic regression, K-Nearest Neighbors (KNN), Multi-factor Dimensionality Reduction (MDR), and Support Vector Machines (SVM), were utilized and demonstrated important finding on anticipating diabetes and hypertension. Plus, the study derived that ethnicity is a basic element for foreseeing diabetes.

Besides, a few investigations have been conducted and uncovered that the machine learning algorithms give early prediction and in addition treatment for hypertension. Koren et al. researched the benefit of machine learning for treatment of hypertension [17]. They utilized machine learning strategies to recognize determinants that add to the achievement of hypertension tranquilize treatment on a huge arrangement of patients. The outcome demonstrated that a completely connected neural network could accomplish AUC as much as 0.82. The consequence of their study demonstrated that machine learning algorithms can give the hypertension treatment combinations of three or four medications. Tayefi et al. developed a M-TREE model to recognize the risk factors that are identified with hypertension [18]. A dataset involving 9078 subjects was part to 70% as preparing set and 30% as the testing dataset to evaluate the performance of the M-TREE. Two models are proposed dependent on various risk factors. The outcome demonstrated that the exactness of the M-TREE for the two models could be as much as 73% and 70%, separately. The finding is accepted to recognize the risk factors that are identified with hypertension that might be used to make

Asma A. AlJarullah conducts a diabetes prediction show by using the M-TREE calculation. In this study, Weka's J48 M-TREE classifier was connected to the dataset to construct the M-TREE show. The precision of the subsequent model was 78.1768% [20]. Wei Yu presents a possibly helpful elective methodology dependent on help vector machine (SVM) strategies that can be utilized to order persons with and without diabetes. The study utilized the data from the U.S. National Health and Nutrition Examination Survey to create SVM models two classification plans, one is analyzed or undiscovered diabetes versus pre-diabetes or no diabetes, the other is undiscovered diabetes or pre-diabetes versus no diabetes. The outcomes demonstrate the territory under the collector working trademark (ROC) bend were separately 83.5% and 73.2%. The outcome shows SVM demonstrating is a promising classification approach for identifying common diseases like diabetes [21]. Mira Kania Sabariah, Aini Hanifa and Siti Sa'adah join Classification and Regression Tree technique (CART) and Random Forest (RF) to manufacture the classification show that can be utilized in the early detection of diabetes. The study demonstrates that the normal exactness of the proposed model is 83.8%, which is higher than the single classifier CART [22].

IV. DATA SOURCE

With the end goal to conduct the examination, we utilized the Pima Indian Diabetes Data (PIDD) set, which is openly accessible from UCI storehouse [23]. The dataset contains females with something like 21 years of age of Pima Indian legacy living around Phoenix, Arizona. There are 768 records in the dataset, out of which 268 cases in class "tried positive for diabetes" and 500 cases for "tried negative for diabetes" with 376 records contain missing qualities. The reason for this exploration is to foresee whether a person would test constructive by using the eight physiological estimations and restorative test outcomes given in the dataset. It is a two-class issue with class esteem 1 being deciphered as "tried positive for diabetes" while class esteem 0 being translated as "tried negative for diabetes". The characteristic information present in the dataset has been given in following Table i .

TABLE
I. ATTRIBUTE INFORMATION

Number	Attribute	Mean	Standard Deviation	Type
1	Number of times pregnant	3.8	3.4	Numeric
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	120.9	32.0	Numeric
3	Diastolic blood pressure (mm Hg)	69.1	19.4	Numeric
4	Triceps skin fold thickness (mm)	20.5	16.0	Numeric
5	2-Hour serum insulin (mu U/ml)	79.8	115.2	Numeric
6	Body mass index (weight in kg/(height in m)^2)	32.0	7.9	Numeric
7	Diabetes pedigree function	0.5	0.3	Numeric
8	Age (years)	33.2	11.8	Numeric

V. NOVEL EFFECTIVE HYBRID PREDICTION FRAMEWORK

A. Working Principle

For the purpose of prediction, a prediction model was defined. The working principle of the proposed model has been shown in **Fig. 1**. It comprises four steps:

1) Data preprocessing: Replace the missing qualities and unthinkable qualities with mean.

2) Data reduction: Remove the erroneously characterized data by using the K-means calculation to bunch the dataset.

3) Classification: Constructing a M-TREE by using the lessened data.

4) Performance evaluation: Evaluate the performance by using a portion of the classifier evaluation measurements.

B. Data Preprocessing

The nature of the data is the key to the entire prediction display as it could impact the prediction result from the investigation. Subsequently, data preprocessing must be done before data examination. The PIDD set contains various missing qualities and unthinkable qualities, for example, 0 weight file and 0 plasma glucose [17]. In this study, data preprocessing is done by supplanting the missing qualities and unthinkable qualities with mean.

C. Data Reduction

Before the application of the classification calculation, the bunching calculation K-means actualized by WEKA is connected to expel the inaccurately arranged examples. From the bunching result, we found 236 cases were inaccurately arranged.

D. Classification

After the second step, it could be seen that 236 occurrences are mistakenly ordered, these cases would be evacuated. At that point, the classification calculation could be connected. The J48 M-TREE calculation executed by WEKA is utilized to assemble a M-TREE with 10-overlap cross-validation strategy.

The dataset for testing. The M-TREE was developed using 532 examples (got after data reduction of the 768 occasions in which 236 were accurately ordered by using the K-means calculation) with 10-overlap cross-validation.

E. Performance Evaluation

In this section, various measures for surveying how great or how exact a classifier is at anticipating the class name of tuples will be presented [6].

1) Accuracy, affectability, and specificity: Firstly, there are four additional terms we have to know that are utilized in processing numerous evaluation measures.

a) True positives (TP): The positive tuples that were accurately marked by the classifier.

b) True negatives (TN): The negative tuples that were accurately named by the classifier.

c) False positives (FP): The negative tuples that were mistakenly named as positive.

d) False negatives (FN): The positive tuples that were mislabeled as negative.

In this study, the accompanying equations are utilized to quantify the exactness, affectability, and specificity.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$\text{Sensitivity} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

Confusion matrix: The confusion lattice is a valuable apparatus for breaking down how well a classifier can perceive tuples of various classes.

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Figure 2. Confusion matrix

TP and TN reveal to us when the classifier is getting things right, while FP and FN disclose to us when the classifier is misunderstanding things. For a classifier that has great exactness, in a perfect world a large portion of the tuples would be spoken to along the diagonal of the confusion lattice with whatever is left of the sections being zero or near zero [6]. A confusion network is appeared in Fig. 2.

3) ***k-fold cross-validation****:* Cross-validation is a valuable and by and large material strategy which is often utilized in machine picking up, including M-TREE. For a k-crease cross-validation, each single model happens precisely k-multiple times as a preparation precedent. Henceforth, the time expected to figure the insights of all test precedents is decreased by a factor k-1 contrasted with running the first calculation k times. The time expected to sort models into youngster hubs is decreased by k-1 if a similar test is chosen in all folds, generally a littler reduction happens. Other than this accelerate, there are no adjustments in the computational unpredictability of the calculation [24]. In this study, we utilized 10-overlap cross-validation in the proposed model. It can diminish the predisposition related with random inspecting strategy.

VI. Experimental Results

In the present study, we initially supplanted the missing qualities and incomprehensible qualities by mean and then expelled the mistakenly characterized tests by using a K-means bunching calculation. After this progression, we had 532 examples cleared out. At long last, we connected the J48 M-TREE calculation with 10-crease cross-validation to the dataset. Using this model we acquired the last outcomes. The confusion framework of the proposed model is appeared in Table

TABLE II. CONFUSION MATRIX OF THE RESULTS

Actual Class \ Predicted Class	Yes	No
Yes	144	21
No	32	335

According to the confusion matrix above, we could figure out the accuracy, sensitivity and specificity of the proposed model are 90.04%, 87.27% and 91.28% respectively.

Since we got the experimental results of the proposed model, we could compare it to some of the other former presented classification models with evaluation measures. From Table Ⅲ, it can be proven that the proposed model has a better accuracy [16] [17] [18] [19] [20] [22] [25] [26] [27].

TABLE III. . COMPARISON OF THE PROPOSED WORK WITH THE EXISTING WORKS

Method	Accuracy	Reference
J48	73.82%	Kandhasamy J P, Balamurali S (2015)
J48	81.33%	Rahman R M, Afroz F (2013)
ANFIS (MATLAB)	78.79%	Rahman R M, Afroz F (2013)
C5.0	76.13%	Meng X H (2013)
Amalgam KNN algorithm	>80%	V Vijayanv , A Ravikumar (2014)
ANFIS algorithm with adaptive KNN	80%	V Vijayanv , A Ravikumar (2014)
J48	78.1768%	Jarullah A A A (2011)
CART and Random Forest	83.8%	Sabariah M T M K (2015)
Predictive model based on H-TSVM	87.46%	Tomar D, Agarwal S (2014)
Naive Bayes	83.37%	K.R. Ananthapadmanaban, G. Parthiban (2014)
Agglomerative Hierarchical Clustering and J48	80.8%	Norul Hidayah Ibrahim (2013)
Proposed Model	90.04%	This Study

VII. CONCLUSION

As indicated by the outcomes appeared in Table iii, we can make sense of that the proposed model has preferred exactness over other classification models for diabetes in the related investigations we mentioned in this paper. Contrasting and the above outcomes, it is obvious to see the proposed model acquires very encouraging outcomes in characterizing the conceivable diabetes patients. With the quickly developing demand for medicinal data investigation, the proposed model can be genuinely valuable to the scientists and specialists for their decision-making on the patients as by using such an efficient model they can make more exact decisions. There are additionally couple of parts of this study could be enhanced further or stretched out later on. For example, the proposed model is proposed to apply to diabetes diagnosis which is a two-class classification issue. It is intriguing to see its conduct on multi-class classification issues. The proposed model is connected to numeric data only, we could enhance the model to see its conduct on various sorts of therapeutic data, for example, pictures and flags. In addition, for reasonable implementation, future work is required to evaluate the effectiveness of the proposed strategy with a bigger measure of data.

REFERENCES

[1] Kerner W, Brückel J. Definition, classification and diagnosis of diabetes mellitus. [J]. Experimental and clinical endocrinology & diabetes: official journal, German Society of Endocrinology [and] German Diabetes Association, 2014, 122(7):384.

[2] Malchoff C D. Diagnosis and classification of diabetes mellitus.[J]. Diabetes Care, 2011, 34(Suppl 1):S62-S69.

[3] Rajendra A U, Tan P H, Subramaniam T, et al. Automated Identification of Diabetic Subjects with and without Neuropathy Using Wavelet Transform on Pedobarograph[J]. Journal of Medical Systems, 2008, 32(1):21-29.

[4] Aguiree F, Brown a, Cho N H, et al. IDF Diabetes Atlas: sixth edition [J]. International Diabetes Federation, 2013.

[5] Fayyad, Usama M, PiatetskyShapiro, et al. From data mining to knowledge discovery: an overview[J]. Ai Magazine, 1996, 17(3):37-54.

[6] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2012, 5(4):1 - 18.

[7] Sumathi S, Sivanandam S N. Introduction to Data Mining and its Applications[J]. Studies in Computational Intelligence, 2006, 26(25):236-238.

[8] Hasim N, Haris N A. A study of open-source data mining tools for forecasting[C]// International Conference on Ubiquitous Information Management and Communication. ACM, 2015:79.

[9] Jovic A, Brkic K, Bogunovic N. An overview of free software tools for general data mining[C]// International Convention on Information and Communication Technology, Electronics and Microelectronics. IEEE, 2014:1112-1117.

[10] Larose D. Data mining methods and models[M]. Wiley-Interscience, 2006.

[11] Liang M. Data Mining: Concepts, Models, Methods, and Algorithms[J]. IIE Transactions, 2004, 36(5):495-496.

[12] Anil K. Jain. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31(8):651-666.

[13] Velmurugan T. Efficiency of k-Means and K-Medoids Algorithms for Clustering Arbitrary Data Points[J]. International Journal of Computer Technology & Applications, 2012, 03(05):1758-1764.

[14] Patel, B. R., & Rana, K. K. (2014). A Survey on M-TREE Algorithm For Classification. International Journal of Engineering Development and Research, 2(1), 1–5.

[15] Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, 6(2).

[16] Kandhasamy J P, Balamurali S. Performance Analysis of Classifier Models to Predict Diabetes Mellitus [J]. Procedia Computer Science, 2015, 47:45-51.

[17] Rahman R M, Afroz F. Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis[J]. Journal of Software Engineering & Applications, 2013, 06(3):85-97.

[18] Meng X H, Huang Y X, Rao D P, et al. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors[J]. Kaohsiung Journal of Medical Sciences, 2013, 29(2):93.

[19] Vijayanv V, Ravikumar A. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus[J]. International Journal of Computer Applications, 2014, 95(17):12-16.

[20] Jarullah A A A. M-TREE discovery for the diagnosis of type II diabetes[C]// International Conference on Innovations in Information Technology. IEEE, 2011:303-307.

[21] Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes[J]. BMC Medical Informatics and Decision Making, 2010, 10(1):16.

[22] Sabariah M T M K, Hanifa S T A, Sa'Adah M T S. Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART)[C]// Advanced Informatics: Concept, Theory and Application. IEEE, 2015:238-242.