A Machine Learning Perspective: To analyzed the diabetes disease

by Bhavini Jain

Submission date: 23-Nov-2020 01:52PM (UTC+0530)
Submission ID: 1454886304
File name: Bhavini_Nandeshwari_Research_paper.docx (4.13M)
Word count: 3294
Character count: 18344

Available online at www.sciencedirect.com

ScienceDirect

materialstoday: PROCEEDINGS

A Machine Learning Perspective: To analyzed the diabetes disease

Bhavini Jain[a],Nandeshwari Ranawat[a] , Pankaj Chittora[b]

[a]Student, Department of Computer Science Engineering, Techno India NJR Institute of Technology Udaipur-313001, India
[b]Assistant Professor, Department of Computer Science Engineering, Techno India NJR Institute of Technology, Udaipur-313001, India

Abstract

This exploration work was directed on the plan and execution of a diabetes forecast framework, a contextual investigation of Pima Indian Diabetes. This exploration will help in robotizing expectation of diabetes even before clinicians showed up. In this analysis, diabetes is foreseen using basic credits, and the correlation of the changing characteristics is moreover depicted.. The current cycle of conveying this action is physically which tends not to breaking down information adaptable for the specialists, and transmission of data isn't straightforward. Different algorithms had been used for better accuracy and the calculations are done on the basis of GINI coefficient. The neural network technique gave the best accuracy of 87.88% that can be useful for the doctors to treat this disease at an early stage.

Keywords: Diabetes:Data Mining:Diagnosis:Artificial Neural Natwork:Random Trees:Gini Coefficient.

1. INTRODUCTION

Diabetes Mellitus which is also commonly called as diabetes is a metabolic disease which causes high blood sugar level. This disease mainly occurs when the amount of insulin in our body is decreased by a certain level in pancreas. This reduction is caused because may be the insulin producing cells in our body are destroyed or damaged and stop responding may ne glucose not reaching the cells. Insulin which is used to move the blood sugar into the cells. Thus, this insul eficiency leaves too much sugar in the blood and not enough in the cell (for energy). Its symptoms may include weight loss, increased hunger, increased thirst, frequent urination, blurry vision and many more. There are major three categories of diabetes. First one is Type 1 Diabetes, which can cause at any age, but is very frequent in children and adolescence. This is caused due to low insulin production in our pancreas which can be balanced by taking insulin injections daily. Next comes the Type 2 Diabetes, which is common in adults and is one of the most frequent form of diabetes. Its only treatment is healthy life style. Last type of diabetes is Gestational diabetes (GDM) that consists of high blood glucose in pregnancies. A person's chance for this disease mainly depends on his genes, life style, age, history and ethnicity. Diabetes is mainly treated by a few different types of medication. One is by taking certain drugs and other by taking injections. Diabetes may lead to many serious other complications of heart diseases, so it's necessary to control the condition with proper medications and some lifestyle changes.

2. LITERATURE REVIEW

Shetty et al. [1] used KNN and the Naïve Bayes technique has been used for the expectation of diabetes. Their system was completed as a pro programming program, where customers offer commitment to terms of patient records and the findi that either the patient is diabetic or not.

Ahmed [2] utilized patient information and strategy for treatment estimations for the gathering of diabetes. Three computations were applied which were Naïve Bayes, vital, and J48 estimations.

Antony et al. [3] utilized clinical data for diabetes conjecture. Gullible Bayes, work based multilayer

perceptron (MLP), and decision tree-based sporadic boondocks (RF) computations were applied after pre-planning of the data. An association-based segment assurance method was used to kill extra features. A learning model by then envisioned if the patient was diabetic. Using a pre-planning technique, results were improved while using Naïve Bayes as differentiated and other AI estimations.

K. Rajesh and V. Sangeetha [4] utilized arrangement method. They utilized C4.5 choice tree calculation to discover concealed examples from the dataset for classifying effectively.

Aiswarya Iyer[5] utilized characterization method to examine shrouded designs in diabetes dataset. Naïve Bayes and Decision Trees were utilized in this model. Correlation was made for execution of the two calculations and viability of the two calculations was appeared therefore.

Saravana Kumar N M et al. S [6] actualized a framework utilizing Hadoop and Map Reduce method for investigation of Diabetic information. This framework predicts kind of diabetes and furthermore chances related with it. The framework is Hadoop based and is practical for any medical services association

Anuja Kumari et al. [7] used the SVM model to predict diabetes utilizing a high-dimensional dataset.

Mohammed Abdul Khaleel et al [8] drove an examination on data mining procedures on clinical data for finding locally relentless diseases. The essential point of convergence of this examination is to separate the data burrowing systems needed for clinical data examination that is especially used to discover locally visit sicknesses, for instance, heart cell breakdown in the lungs, illnesses, chest threat using plan and backslide tree (CART) figuring and he decision tree computations, for instance, ID3, C4.5.

Sajida et al. [9] in analyzes the piece of Adaboost and Bagging group AI methods using J48 decision tree as the explanation behind describing the Diabetes Mellitus and patients as diabetic or non-diabetic, in view of diabetes risk factors.

Orabi. [10] in arranged a system for diabetes desire, whose principal point is the gauge of diabetes a candidate is suffering at a particular age. The proposed system is arranged subject to the possibility of AI,by applying decision tree. Results were pleasant as the arranged structure works splendidly in envisioning the diabetes scenes at a particular age, with higher exactness using Decision tree.

Pradhan et al [11] in used Genetic programming (GP) for the arrangement and testing of the data base for forecast of diabetes by using Diabetes instructive record which is sourced from UCI storage facility. Results achieved using Genetic Programming gives ideal precision w n appeared differently in relation to other realized methodology. There can be significantly improve in precision by saving less exertion for classifier age. It winds up being useful for diabetes conjecture with ease.

Rashid et al. [12] in arranged a figure model with two sub-modules to foresee diabetes-steady ailment. ANN (Artificial Neural Network) is used in the essential module and FBS (Fasting Blood Sugar) is used in the ensuing module. Decision Tree (DT)is used to recognize the symptoms of diabetes on patient's prosperity.

3. METHODOLOGY

3.1. Dataset

This dataset for research is provided by National Institute of Diabetes and Digestive d Kidney Diseases. Main purpose of using this dataset is to making the machine learning model for prediction whether a patient has Diabetes or not, based on factors such as blood pressure, glucose level, age etc. The type of data set and question is in form of binary classification (0 and 1). This dataset includes 9 columns with one predicting the result. It has 768 records with both positive and negative instances. The dataset description has been shown in the table 1.

Table 1 Description of Dataset

S.No.	Attribute Name	Attribute Description

1.	Pregnancies	Count of pregnancies
2.	Glucose	glucose level in the body
3.	Blood Pressure	blood pressure level (mm/Hg)
4.	Skin thickness	thickness of triceps skin fold
5.	insulin	Insulin serum
6.	BMI	Body-mass-index
7.	Diabetes pedigree function	pedigree function of diabetes
8.	Age	Patient age
9.	Outcome	Class variable (0 or 1)

3.2 *Construction of references*

3.2.1 *Neural Networks*

A neural network can be defined as a sequence of algorithms designed to identify potential relationships in a set of data by imitating the operation of the human brain. It has three units input unit, output unit and transfer unit. Neural networks can make alternations to constantly changing inputs. Therefore, the network can produce the best results without redesigning the output standard. A neural network consists of interrelated nodes where each node is a perceptron, similar to multiple linear retrogression. In a multilayer perceptron (MLP), the perceptron's are arranged in inter associated layers. The input layer fetch input signals. The output layer has categorization of output signals to which input signals can be mapped.

3.2.2 *Discriminant*

Fisher Linear Discriminant Analysis is a reduction technique which is used to recognize patterns of linear combination that classifies or separates two or more than two entities of events based on k variables. The end result for this algorithm can be used for forecasting of group membership. This analysis can be used when groups are already determined. Discriminant analysis is used for analyzing differences in groups. It can be used for classification of new entities.

3.2.3 Random trees

Random forest is a easy to use and most convenient type of algorithm where it constructs various decision trees and then all this combine together to obtain more accurate prediction. This can be used for both categorized and regression problems. This algorithm adds more randomness to the model which results in making of a better model. Its only disadvantage is that if many numbers of trees are used than it can slow down the speed of the model and cause clashes too.

3.2.4 CHAID

Chi-square Automatic Interaction Detector is a technique in which nominal, ordinal and continuous data are used. This algorithm discovers correlation between variables and objects. In such consumer research, algorithms include both descriptive analysis (methods not based on standard variables) and predictive analysis (based on standard variables). However, CHAID established predictive analysis and established standard variables associated with the remaining variables of the configuration variables. These variables are configured to configure segments through the correlation relationship proved by important chi-square.

3.2.5 Support Vector Machine

Support vector machine also known as SVM is machine learning Classification algorithm, mainly used in Classification as well as regression issues. SVM most used profit strategy changed to solve complex problems. Second programming problem due to the high-performance classification of support vector machines, a wide variety of applications. The SVM choose the extreme point (known as support vectors) that help in making the hyperplane.

This algorithm mainly aims on creating the best decision boundary (known as hyperplane) that can separate space into classes. There are two types of SVM: Linear and Non-Linear SVM.

3.2 Performance Evaluation Measures

Building AI models deals with a useful criticism standard.. Model is created, get analysis from estimations, cause improvements and continue until you to achieve an appealing exactness. Evaluation estimations explain the introduction of a model. A critical piece of evaluation estimations is their capacity to isolate among model results.

3.2.1 Confusion Matrix:

A confusion matrix is a A X A framework, where A is the quantity of classes being anticipated. Some terms to be defined for confusion matrix. Positive Predictive Valueis the extent of positive cases that were accurately recognized. Negative Predictive Value is the extent of negative cases that were accurately distinguished.

Table 2: Confusion Matrix's Parameters

		Predicted	
		Positive	Negative
Actual	True	a	b
	False	c	d

Positive predicted value = a/(a+b)
Negative Predicted value = d/(c+d)

Coincidence Matrix for $N-Outcome (rows show actuals)

'Partition' = 1_Training	0	1
0	265	29
1	42	126
'Partition' = 2_Testing	**0**	**1**
0	193	13
1	25	75

Fig 1: Eaxmple of Confusion Matrix of Neural Network Algorithm

3.2.2 *Classification Accuracy:*

Generally speaking, accuracy shows execution of the grouping framework. It can be calculated as:

Accuracy: (a+d)/(a+b+c+d)

3.2.3 Classification error

It can be defined as the overall incorrect classification of a model. It can be formulated as:

Error: (c+d)/(a+b+c+d)

3.3.4. *Precision*

The extent of genuine negative cases which are accurately distinguished. It can be formulating as follows:

Precision= d/(b+d)

3.3.5. Recall

The extent of real certain cases which are effectively distinguished. It can be formulating as:

Recall= a/(a+c)

3.3.6. *AUC*

Area Under the Roc bend gives a complete extent of execution over all possible portrayal limits. One strategy for unraveling AUC is as the probability that the model positions a subjective positive model more significantly than a sporadic negative model.

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$N-Outcome	0.846	0.693	0.902	0.804

Fig 2: Evaluation Metrics showing AUC and GINI cofficient of Neural Network Algorithm

3.3.7. *GINI*

Gini coefficient is at times utilized in characterization issues. Gini coefficient is taken from the AUC ROC number. Gini is only proportional between zone between the ROC bend and the diagonal line and zone of the above triangle. Its's formula is as follows:

GINNI=2*AUC – 1

GINNI whose value is above 60%is considered as good model.

4. Result

Distinctive characterization calculations were applied on our dataset, and results of all methods were somewhat extraordinary as working standards of every calculation is unique. The outcomes were assessed based on exactness and the GINNI coefficient. The exactness of models was anticipated with the assistance of a confusion matrix as described above. Firstly, we used random trees algorithm. Random forest calculation builds choice trees on information results and then after gets the projection from every one of them and then lastly selects the best system by methods for setting down a ballot. It is a group plan which is superior to a unique choice tree since it reduces the incompleteness by averaging the outcome. We got 63.51accuracy by testing 195 right and 108 wrong. Right off the bat we went after for the 70-30 proportion yet the outcomes were most exceedingly awful, at that point 75-25 proportion we got the exactness as 63.51% which is the most elevated of all. In this we got the AUC as 0.71 and GINI coefficient as 0.42. The performance evaluation was 0.172 for positive results and 0.341 for negative results. The counts for precision and recall were done and acquired 0.785 and 0.619 values individually. CHAID: Using this algorithm we obtained relationship between predictable variable and response variable. From this we visualized many relationships of categorized data. The accuracy obtained was 83.71%, by testing 277 correct and 49 wrong. Its sensitivity was 0.885 and specificity as 0. 863.The AUC calculated was 0.89 and GINI coefficient 0.77.SVM: By using this algorithm we obtained hyperplane and support vector. In this we created extreme point that helped in obtaining the hyperplane. In this accuracy obtained was 86.55% by testing 276 positive and 48 are negative, with precision 0.892 and recall 0. 897.The GINI coefficient was 0. 823.The Neural Network algorithm,

applied to acquire advance outcomes. The model was tuned based on number of covered up neurons, number of learning cycles just as estimation of starting learning loads. We got 87.1%accuracy by testing 297 correct and 44 wrong. Firstly, we tried for the 60-40 proportion but the results were worst, then 70-30 proportion we got the accuracy as 87% which is the highest of all. In this we got the AUC as 0.892 and GINI coefficient as 0.784. The performance evaluation was 0.857 for positive outcomes and 0.322 for negative outcomes. The calculations for precision and recall were done and obtained 0.880 and 0.943values respectively. Next, we used diriment algorithm in which model tunning is based on independent variable as a group and dependent variables as predictors. From this we predicted memberships of such groups. This function follows two step process. One initially plays out the multivariate test, and, if measurably critical, continues to see which of the factors have essentially various methods over the groups. We got the 67.5%accuracy.

	Precision	Recall	F-Measure	AUC	GINI Coefficient	Acuuracy
Neural Network	0.880	0.943	0.910	0.89	0.781	87.88%
Discriminant	0.687	0.928	0.790	0.59	0.172	67.51%
Random Trees	0.785	0.619	0.692	0.70	0.401	63.51%
CHAID	0.863	0.885	0.874	0.89	0.770	**83.71%**
SVM	0.892	0.897	0.895	0.91	0.823	86.55%

Fig 3: Shows Performance evaluation of different classifiers on diabetes disease on full features.

The Gini coefficient is the proportion of the region between the line of wonderful balance and the observed Lorenz bend to the zone between the line of amazing uniformity and the line of wonderful disparity. The higher the coefficient, the more inconsistent the circulation is. In the chart on the right, this is given by the proportion A/(A+B), in which A and B are the zones of areas as set apart in the outline.
Graph of GINI coefficient with respect to classifiers with obtained values is shown as below:

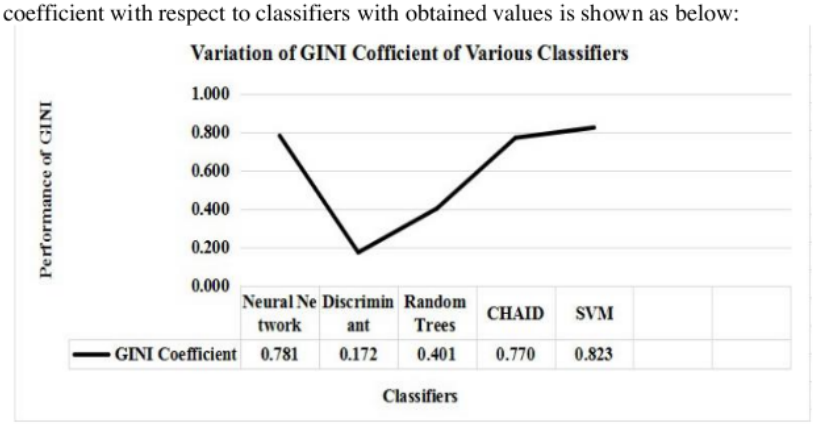

Fig 4: GINI Cofficient Graph

Graph below shows the comparison with of precision , recall and accuracy of various classifier.

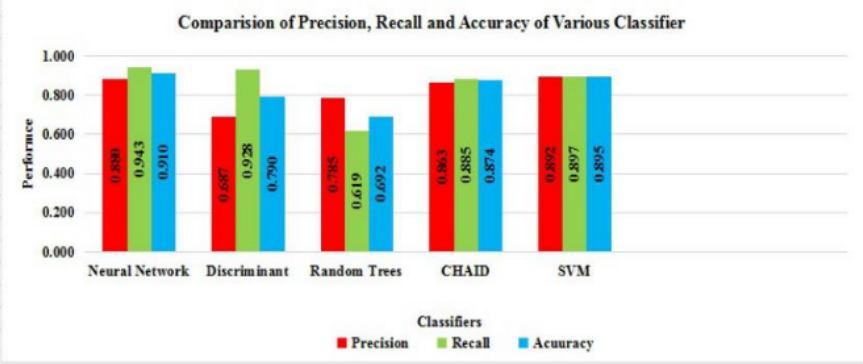

Fig 5: Graph of Precision, Recall and Accuracy

As per the observed value neural network algorithm gives the highest accuracy precision and recall whereas random trees have the lowest recall value while the CHAID and SVM were almost similar.

The below graph depicts the variation of F-measure of various classifiers.

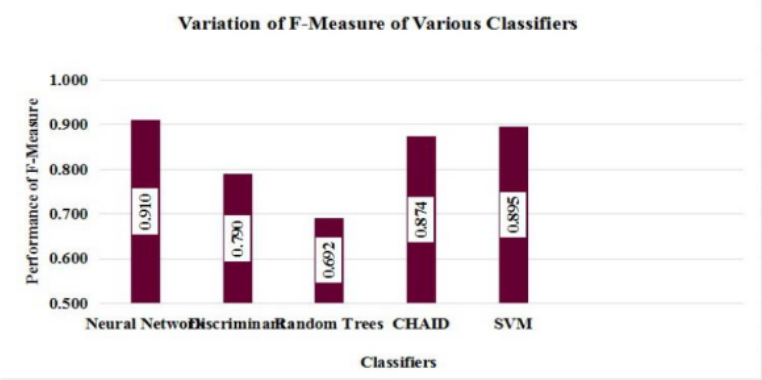

Fig 6: F-Measure Graph

F-Measure gives an approach to join both accuracy and recall into a solitary measure that catches the two properties. Neural network has the highest f-measure value while random tree gives the lowest.

5. Conclusion and Future Work

ML and data mining procedures are critical in disorder examination. The capacity to predict diabetes early, acknowledges a significant part for the patient's reasonable treatment methodology. In this paper, a barely any current course of action strategies for clinical investigation of diabetes patients have been analyzed dependent on precision. AI procedures were applied on the Pima Indians diabetes dataset, just as prepared and approved against a trial dataset. The outcome of our model operations has indicated that Neural Network Algorithm outflanks various models. By observing the results using affiliation rules, there found firm connection of BMI and Glucose with diabetes. The obstruction of this experimentation is that a coordinated dataset has been chosen at this point later on, unstructured data will in like manner be considered, and these methods will be applied to other clinical regions for desire, for instance, for different sorts of danger, psoriasis, and Parkinson's contamination. Various credits including real idleness, family foundation of diabetes, and smoking affinity, are moreover proposed to be viewed as later on for the investigation of diabetes. An Application utilizing an information mining calculation of classes' correlation has been created to foresee the event of or repeat of diabetes chances. This application would be a gigantic resource for specialists who can have organized explicit and significant data about their patients/others so they can guarantee that their determination or derivations are right and proficient. Future work ought to be done on improving the

precision of the expectation by expanding the degree of preparing information. Its execution can be additionally improved by recognizing and consolidating different boundaries and expanding size of preparing.

References

[1] Shetty D et al, "Diabetes sickness expectation utilizing information mining. Developments in data, inserted and correspondence frameworks", (ICIIECS), worldwide meeting on. 2017. p. 1–5.

[2] Singh A at al., Effect of various information types on classifier execution of irregular timberland, credulous Bayes, and K-closest neighbors calculations. Int J Adv Comput Sci Appl 2017;8:1–10.

[3] Ahmed TM. Utilizing information digging to create model for arranging diabetic patient control level dependent on chronicled clinical records. J Theor Appl Inf Technol 2016;87.

[4] Singh DAAG et al., Di s expectation utilizing clinical information. J Comput Intell Bioinform 2017;10:1–8

[5] K. Rajesh et al., "Use of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Designing and Innovative T nology (IJEIT) Volume 2, Issue 3, September 2012

[6] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Information Mining and Know e Management Process (IJDKP) Vol.5, No.1, January 2015

[7] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", second Worldwide Sym um on Big Data and Cloud Computing,2015.

[8] mari VA, Chitra R. Arrangement of diabetes illness utilizing support vector machine. Int J Eng Res Afr 2013;3:1797–801

[9] Mohammed, A.K., Sateesh, K. P., Dash G. N., 2013, "A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Dise s" International Journal of Advanced Research in Computer Science and Software Engineering, 3(8), pp. 149-153.

[10] Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Execution Analysis of Data Mining Classification Techniques to Predict Dia es. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016

[11]Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Spring Springer. pp. 420–427.

[12] Pradhan, M.P., G.R., 2014. Plan of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Fram orks and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[13] Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract, 2016. An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424, 323–335. doi:10.1007/978-3-319-28031-8.

A Machine Learning Perspective: To analyzed the diabetes disease

ORIGINALITY REPORT

28%	16%	24%	16%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Talha Mahboob Alam, Muhammad Atif Iqbal, Yasir Ali, Abdul Wahab et al. "A model for early prediction of diabetes", Informatics in Medicine Unlocked, 2019 Publication	7%
2	www.irjet.net Internet Source	2%
3	www.ijitee.org Internet Source	2%
4	Deepti Sisodia, Dilip Singh Sisodia. "Prediction of Diabetes using Classification Algorithms", Procedia Computer Science, 2018 Publication	2%
5	www.ijetr.org Internet Source	1%
6	Submitted to University of Sheffield Student Paper	1%
7	Submitted to KDU College Sdn Bhd Student Paper	

		1%
8	ijcsn.org Internet Source	1%
9	Submitted to RMIT University Student Paper	1%
10	Submitted to October University for Modern Sciences and Arts (MSA) Student Paper	1%
11	"Diabetes Prediction and Analysis using Machine Learning Methods", International Journal of Innovative Technology and Exploring Engineering, 2020 Publication	1%
12	Submitted to Université Internationale de Rabat Student Paper	1%
13	www.ijert.org Internet Source	1%
14	Submitted to University of Bradford Student Paper	1%
15	Submitted to Wilmington University Student Paper	1%
16	B.L. Ahuja, Veera Raykar, Ritu Joshi, Shailja Tiwari, Sonal Talreja, Gopal Choudhary. "Electronic properties and momentum densities	1%

	of tin chalcogenides: Validation of PBEsol exchange-correlation potential", Physica B: Condensed Matter, 2015 Publication	
17	docplayer.net Internet Source	1%
18	Submitted to essex Student Paper	<1%
19	"New Trends in Computational Vision and Bio-inspired Computing", Springer Science and Business Media LLC, 2020 Publication	<1%
20	Submitted to Georgia Institute of Technology Main Campus Student Paper	<1%
21	Submitted to Sheffield Hallam University Student Paper	<1%
22	www.mdpi.com Internet Source	<1%
23	www.icaen.uiowa.edu Internet Source	<1%
24	"Innovations in Bio-Inspired Computing and Applications", Springer Science and Business Media LLC, 2016 Publication	<1%

25	hdl.handle.net Internet Source	<1%
26	www.ijiset.com Internet Source	<1%
27	Nagaraj V. Dharwadkar, Shivananda R. Poojara, Anil K. Kannur. "chapter 14 Risk Analysis of Diabetic Patient Using Map-Reduce and Machine Learning Algorithm", IGI Global, 2021 Publication	<1%
28	"Soft Computing: Theories and Applications", Springer Science and Business Media LLC, 2020 Publication	<1%
29	www.conductdisorders.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		