https://doi.org/10.46335/IJIES.2023.8.8.6 **e-ISSN: 2456-3463**
Vol. 8, No. 8, 2023, PP. 29-33

International Journal of Innovations in Engineering and Science, www.ijies.net

A Review on Diabetic Detection Using Machine Learning

Shirin S. Pinjari[1], Nilesh Vani [2]

[1]PG Student, [2]Associate Professor
GF's Godavari College of Engineering, Jalgaon, India, 425002

shirinrp8989@gmail.com

***Received on**: 28 April, 2023* ***Revised on**: 23 May, 2023* ***Published on**: 25 May, 2023*

Abstract – *Healthcare data typically consists of a variety of variable types, missing values, and is quite large, complex, and heterogeneous. These days, having access to such knowledge is required. By building models from healthcare data sets, such as those pertinent to patient data sets for diabetes, data mining can be utilized to extract knowledge. In order to predict the prevalence of diabetes using 18 risk factors, three data mining algorithms—Self-Organizing Map (SOM), C4.5, and Random Forest—are used to adult population data from the Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia. Random Forest performed better than other data mining classifiers in comparison.*

Keywords- *data mining, machine learning, Diabetes, Decision Trees, Healthcare, Logistic Regression, Naïve Bayes, Random Forrest, SVM.*

I. INTRODUCTION

Chronic hyperglycemia and aberrant protein, carbohydrate, and lipid metabolism are features of the metabolic illness known as diabetic disease (DA). The following are the three main types of diabetes: (DA). A person with type 1 diabetes must currently inject insulin or use an insulin pump because the body cannot produce insulin. Previously, this disorder was referred to as IDDA (insulin-dependent diabetic illness) or juvenile diabetes. The main contributing factor to type 2 diabetes mellitus is insulin resistance, a condition in which cells mistreat insulin. Insulin resistance and absolute insulin deficit (DA) rarely coexist. The previous titles for this type were "adult-onset diabetes" or NIDDA (non-insulin-dependent diabetic ailment). Pregnant women with high blood sugar who have not yet been diagnosed with diabetes can develop gestational diabetes, the third main kind. India (40.9 million cases), China (38.9 million cases), the US (19.2 million cases), Russia (9.6 million cases), and Germany (7.4 million cases) had the highest rates of diabetes diagnoses in 2007. It is required to structure and emphasize the amount of data into a nominal value using a workable approach due to the expanding unstructured nature of diabetic data from the health industry or any other sources. To expand patient access to healthcare services on all levels, reliable electronic communication technologies and data sharing on diabetes must be used in tandem. Such that each patient's data must be kept in one place. Using a health information exchange (HIE), it is possible to safely gather clinical data from numerous different repositories and merge that data into a single patient health record. In order to discover or predict specific future events, predictive analysis is a strategy that combines data mining, statistics, and game theory techniques [6]. It accomplishes this using models and techniques from statistics or other types of analysis. It uses both current

https://doi.org/10.46335/IJIES.2023.8.8.6 **e-ISSN: 2456-3463**
Vol. 8, No. 8, 2023, PP. 29-33

and historical data. Using big data analytics, significant predictions or choices in the healthcare industry can be produced. In this study, we predicted the most common forms of diabetes, its complications, and the type of treatment to be given using a predictive analysis algorithm in a Hadoop/Map Reduce context. According to the analysis, this approach to patient care and treatment works well and produces better results in terms of affordability and accessibility. To predict upcoming events or unknowable consequences, predictive analytics employs statistical or machine learning techniques [1]. The question "what's the next step?" is addressed using text mining with unstructured data. It forecasts conduct, trends, and activity for the present and the future. It accomplishes this by utilizing automated machine learning algorithms, analytical queries, and statistical analysis methods. The World Health Organization (WHO) forecasts that 350 million people will have diabetes worldwide by 2030 [2, 3]. Almost all of the food we consume is transformed into glucose or sugar. Now, this sugar or glucose is transformed into energy. Glucose is carried by insulin to the body's cells. The body's insufficient or improper usage of the hormone insulin causes diabetes.

To create predictive models for predictive analytics, experts are required. Prediction is accomplished using these models. Predictive analytics has a wide range of uses, particularly in the healthcare sector. The most prevalent illness right now is diabetes. Everyone is impacted, and the number of patients grows every day. It was said by Type 1, Type 2, gestational, and pre-diabetes are the four different kinds of diabetes. Type 1 diabetes, in which the pancreas does not generate the hormone insulin, is often referred to as insulin-dependent diabetes [4]. Type 2 diabetes, sometimes referred to as non-insulin-dependent diabetes, is characterized by normal insulin production and insulin resistance in the body [4]. Gestational diabetes is a form of diabetes that can affect pregnant women [5]. Blood sugar levels that are higher than normal but not high enough to be diagnosed as diabetes are referred to as pre-diabetes [6]. Diabetes is a disease that harms blood vessels, the kidneys, the heart, the eyes, the nerves, and other body parts [7]. Predictive analytics in the field of diabetes can be utilized for diabetes diagnosis, diabetes prediction, diabetes self-management, and diabetes prevention, according to a literature review.

II. LITERATURE REVIEW

• M. Kannan and P. Yasodha [2] Several datasets that can be used to diagnose diabetes in people are categorized in this study. The data set for the diabetic patient is created by combining the 239 instances and 7 attributes from the hospital warehouse. Both blood tests and urine tests are mentioned in these instances of the dataset.

• A. Rawal, N. NiyatiGupta, and V. Narasimhan [3] The study assesses the effectiveness of the same classifiers when applied using several additional tools, such as Rapidminer and Matlab, while keeping the same parameters. It makes an effort to evaluate the accuracy, sensitivity, and specificity of different classification methods in WEKA as well as seek and find the accuracy, sensitivity, and specificity percentages of various classification methods. The JRIP, BayesNet, and Jgraft algorithms were used.

• Paul Lee H [4] A classification model for undiagnosed diabetes was created using the Classification and Regression Tree (CART) method. According to WHO guidelines, the person showed signs of diabetes. The exposure variables were socioeconomic status and demographics. The evaluation dataset, which represented the remaining 30% of the sample, was utilized to determine the testing dataset's area under the receiver operating characteristic curve (AUC). CART models were created using the training dataset, which contained 70% of the data that was randomly chosen. The training dataset, the oversampled training dataset, the weighted training dataset, and the under sampled training dataset were all used to develop CART models. In addition, the case-to-control ratios of 1:1, 1:2, and 1:4 for resampling were looked at.

• Meryem SAIDI, Nesma SETTOUTI, and Mohamed Amine CHIKH [7]. The authors of this article suggested MAIRS2, a ground-breaking method for identifying diabetes. The size of the diabetes dataset was reduced in the first phase using the AIRS2 learning approach, and the resulting database was named Memory Cells Pool. Because classification is done using the k-nearest neighbor method, whose classification time depends on the amount of data points used to categories a previously unknown data item, the methodology benefits from any decrease in the overall number of produced memory cells. We employ the fuzzy k-nearest neighbor to classify each patient in order to get around the k-nn classifier's restrictions in the second stage. We use the diabetes dataset for Pima Indians to assess the efficacy of our MAIRS2 algorithm.

https://doi.org/10.46335/IJIES.2023.8.8.6 e-ISSN: 2456-3463
Vol. 8, No. 8, 2023, PP. 29-33

• Sampath P., Lavanya S., Eswari T., and Dr. Saravana Kumar N. M. All of the aforementioned studies were successful in producing accurate prediction models from the diabetic data set. In this study, we predict and classify the type of diabetes using a predictive analysis technique in a Hadoop/Map Reduce environment. The cost-effectiveness of this method of patient care is higher than it is in terms of accessibility and affordability. Data collecting, data warehousing, predictive analysis, and report processing are just a few of the functions that make up the predictive analysis system's architecture.

III. METHOLOGY

Many algorithms are used in the literature review to detect diabetes. On the basis of the survey data, Nave Bayes, Logistic Regression, J48, and AdaBoost outperform other diabetes diagnosis algorithms.

1) Naïve Bayes

Naïve Bayes is a classification algorithm. This algorithm depends upon the Bayes theorem. This is a simple and very powerful algorithm.

- Bayes theorem: Bayes theorem finds the probability of an event occurring given the probability of another event that has already occurred.

$$P(A/B) = (P(B/A)\,P(A)) / P(B)$$

Where P (A) – Priority of A

P (B) – Priority of B

P (A/B) – Posteriori priority of B

- Naïve Bayes algorithm is easy and fast. This algorithm needs less training data and is highly scalable

2) Logistic Regression

A supervised classification technique called logistic regression gives the likelihood that a binary dependent variable would be predicted from the independent variable of the dataset. The likelihood of an outcome with two possible values—zero or one, yes or no, and false or true—is predicted by logistic regression..

Although logistic regression and linear regression are similar, logistic regression yields a curve instead of a straight line. Based on the use of one or more predictors or independent variables, logistic regression generates logistic curves that plot values between zero and one. Regression is a regression model that examines the relationship between a number of independent factors and a categorical dependent variable. Binary logistic models, multiple logistic models, and binomial logistic models are only a few of the numerous varieties of logistic regression models. In order to determine the likelihood of a binary answer based on one or more predictors, the binary Logistic Regression model is utilized. ·

This algorithm is similar to the linear regression algorithm. But linear regression is used for predicting / forecast values and Logistic regression is used for the classification task.

· Linear regression is classified as

Ø Binomial – 2 Possible types (i.e. 0 or 1) only

Ø Multinomial – 3 or more possible types which are not ordered

Ø Ordinal – Ordered in category (i.e. very poor, poor, good, very good)

· This algorithm is easy for binary and multivariate classification tasks.

3) J48

A conditional control statement is used in the decision tree algorithm, which predicts the ultimate decision by using a tree-like graph or model of decisions and their potential outcomes. An algorithm for approaching discrete-valued target functions is called a decision tree, and it is represented by a learnt function. These kinds of algorithms are well-known and have been effectively used for a wide variety of inductive learning tasks. In order to determine whether a new transaction is legitimate or fraudulent for which the class label is unknown, first assign it a label. Next, the transaction value is checked against the decision tree, and finally, a path is established from the root node to the transaction's output or class label. The result of the content of the leaf node is decided by decision rules. In general rules have the form of 'If condition 1 and condition 2 but not condition 3 then outcome'. A decision tree enables the inclusion of additional potential scenarios and aids in determining the worst, best, and anticipated values for various circumstances.

https://doi.org/10.46335/IJIES.2023.8.8.6 **e-ISSN: 2456-3463**
Vol. 8, No. 8, 2023, PP. 29-33

International Journal of Innovations in Engineering and Science, www.ijies.net

· J48 algorithms used to generate a decision tree and it is for the classification task.

· J48 is an extended of ID3 (Iterative Dichotomieser 3). This algorithm has some special features such as rules derivation, continuous value range, decision tree pruning, etc.

· J48 algorithm is the most extensively analyzed area in machine learning. They analyze based on generated decision trees and understandable rules.

· This algorithm works on constant and categorical variables.

4) Ada Boost

Ada Boost is a machine learning algorithm created mostly for binary categorization. This approach is used to improve the decision tree's performance.

· For Ada Boost, Each instance in the training dataset is weighted. Initial weight is set To Weight (xi) = (1/n) Where, xi – ith training instance

n – Number of the training instance

This algorithm is mainly for classification rather than regression. So the Ada Boost algorithm is used in fraud detection because this classifies the transaction which transactions that are fraudulent and non-fraudulent.

IV. CONCLUSION

In this paper, a thorough explanation of predictive modeling is provided along with a combination of traditional and hybrid prediction models. Modeling, This study shown that hybrid models outperform traditional models in terms of accuracy of outcomes. A researcher who is interested in conducting research to develop a clinical prediction model would find this study to be helpful. Diabetes is a frequent disease in underdeveloped countries like India, so there is a lot of space for improvement in clinical prediction models in this area.

According to the analysis of the aforementioned studies, some of the numerous holes that need to be filled include the use of different datasets, the detection of outliers, strengthening the prediction model, and integrating optimization approaches into the hybrid prediction model.

REFERENCES

[1] *P. T. Katzmarzyk, C. L. Craig, and L. Gauvin, "Adiposity, physical fitness, and incident diabetes: The physical activity longitudinal study," Diabetologia, vol. 50, no. 3, pp. 538–544, Mar. 2007.*

[2] *P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WEKA Tool", International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011, ISSN 2229-5518.*

[3] *N. Niyati Gupta, A .Rawal, and V.Narasimhan ,"Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data", IOSR Journal of Computer Engineering, vol. 11, no. 5, pp. 70-73, 2014.*

[4] *[4] M. Chikh,M. Saidi, and N. Settouti, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K- nearest neighbor," Journal of medical systems,vol.36, no.5, pp. 2721-2729, 2015.*

[5] *R. N. Feng, C. Zhao, C. Wang, Y. C. Niu, K. Li, F. C. Guo, S. T. Li, C. H. Sun, and Y. Li, "BMI is strongly associated with hypertension, and waist circumference is strongly associated with type 2 diabetes and dyslipidemia, in northern Chinese adults," J. Epidemiol., vol. 22, no. 4, pp. 317–323, May 2012.*

[6] *A. Berber, R. G´omez-Santos, G. Fangh¨anel, and L. S´anchez-Reyes, "Anthropometric indexes in the prediction of type 2 diabetes mellitus, hypertension and dyslipidemia in a Mexican population," Int. J. Obes. Relat Metab. Disorders, vol. 25, no. 12, pp. 1794–1799, Dec. 2001.*

[7] *B. Balkau, D. Sapinho, A. Petrella, L. Mhamdi, M. Cailleau, D. Arondel, and M. A. Charles, D. E. S. I. R. Study Group, "Prescreening tools for diabetes and obesity-associated dyslipidemia: Comparing BMI, waist and waist-hip ratio. The D.E.S.I.R. Study," Eur. J. Clin. Nutr., vol. 60, no. 3, pp. 295–304, Mar. 2006.*

[8] *I. S. Okosun, K. M. Chandra, S. Choi, J. Christman, G. E. Dever, and T. E. Prewitt, "Hypertension and type 2 diabetes comorbidity in adults in the United States: risk of overall and regional adiposity," Obes. Res., vol. 9, no. 1, pp. 1–9, Jan. 2001.*

[9] *L. A. Sargeant, F. I. Bennett, T. E. Forrester, R. S. Cooper, and R. J. Wilks, "Predicting incident diabetes in Jamaica: the role of anthropometry," Obes. Res., vol. 10, no. 8, pp. 792–798, Aug. 2002.*

[10] *N. T. Duc Son le, T. T. Hanh, K. Kusama, D. Kunii, T. Sakai, N. T. Hung, and S. Yamamoto, "Anthropometric characteristics, dietary patterns and risk of type 2 diabetes mellitus in Vietnam," J. Amer. Coll. Nutr., vol. 24, no. 4, pp. 229–234, Aug. 2005.*

[11] *G. T. Ko, J. C. Chan, C. S. Cockram, and J. Woo, "Prediction of hypertension, diabetes, dyslipidemia or albuminuria using simple anthropometric indexes in*

Hong Kong Chinese," Int. J. Obes. Relat. Metab. Disorders, vol. 23, no. 11, pp. 1136–1142, Nov. 1999.

[12] *M. B. Snijder, P. Z. Zimmet, M. Visser, J. M. Dekker, J. C. Seidell, and J. E. Shaw, "Independent and opposite associations of waist and hip circumferences with diabetes, hypertension, and dyslipidemia: The AusDiab study," Int. J. Obes. Relat. Metab. Disorders, vol. 28, no. 3, pp. 402–409, Mar. 2004.*

[13] *B. J. Lee, B. Ku, J. Nam, D. D. Pham, and J. Y. Kim, "Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes," IEEE J. Biomed. Health Information. vol. 18, no. 2, pp. 555–561, Mar. 2014.*

[14] *L. de Koning, H. C. Gerstein, J. Bosch, R. Diaz, V. Mohan, G. Dagenais, S. Yusuf, and S. S. Anand, Epi DREAM Investigators, "Anthropometric measures and glucose levels in a large multi-ethnic cohort of individuals at risk of developing type 2 diabetes," Diabetologia, vol. 53, no. 7, pp. 1322–1330, Jul. 2010.*

[15] *I. S. Okosuna and J.M.Boltrib, "Abdominal obesity, hypertriglyceridemia, hypertriglyceridemia waist phenotype and risk of type 2 diabetes in American adults," Diabetes Metab. Syndrome, vol. 2, no. 4, pp. 273–281, Dec. 2008.*

[16] *Z. Yu, L. Sun, Q. Qi, H. Wu, L. Lu, C. Liu, H. Li, and X. Lin, "Hypertriglyceridemia waist, cytokines and hyperglycemia in Chinese," Eur. J. Clin. Invest. vol. 42, no. 10, pp. 1100–1111, Oct. 2012.*

[17] *T. Du, X. Sun, R. Huo, and X. Yu, "Visceral adiposity index, hypertriglyceridemic waist and risk of diabetes: The china health and nutrition survey 2009," Int. J. Obes. (Lond.), vol. 38, no. 6, pp. 840–847, Jun. 2014.*

[18] *M. Solati, A. Ghanbarian, M. Rahmani, N. Sarbazi, S. Allahverdian, and F. Azizi, "Cardiovascular risk factors in males with hypertriglycemicwaist (Tehran lipid and glucose study)," Int. J. Obes. Relat. Metab. Disorders, vol. 28, no. 5, pp. 706–709, May 2004.*

[19] *I. Lemieux, A. Pascot, C. Couillard, B. Lamarche, A. Tchernof, N. Alm´eras, J. Bergeron, D. Gaudet, G. Tremblay, D. Prud'homme, A. Nadeau, and J. P. Despr´es, "Hypertriglyceridemic waist: A marker of the atherogenic metabolic triad (hyperinsulinemia; hyper apolipoprotein B; small, dense LDL) in men?" Circulation, vol. 102, no. 2, pp. 179–184, Jul. 2000.*

[20] *L. B. Tank´o, Y. Z. Bagger, G. Qin, P. Alexandersen, P. J. Larsen, and C. Christiansen, "Enlarged waist combined with elevated triglycerides is a strong predictor of accelerated atherogenesis and related cardiovascular mortality in postmenopausal women," Circulation, vol. 111, no. 15, pp. 1883–1890, Apr. 2005.*