JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF DIABETES MELLITUS USING MACHINE LEARNING TECHNIQUE AND CLINICAL NOTES

NAME - HIRALKUMAR DILIPBHAI PATEL

COLLEGE NAME – AMITY DIRECTORATE OF DISTANCE & ONLINE EDUCATION

ABSTRACT

The healthcare business deals with a lot of sensitive and enormous volumes of data that must be treated with care. Diabetes Mellitus is a fatal illness that is growing more frequent across the world. Doctors and nurses desire a dependable way to predict whether or not a patient has Diabetes. Using various machine learning techniques, you may examine the data from many perspectives and distil it into useful knowledge. Using data mining techniques on the massive volumes of data that we have access to and can obtain will allow us to learn useful things. The key objective is to discover new patterns and then figure out how to use these patterns to provide significant and useful information to consumers. Diabetes can cause issues with the heart, kidneys, nerves, and vision. It is critical to develop efficient methods for mining diabetes data. It will be taught how to utilise data mining methods and techniques to determine the best techniques and tactics for effectively categorising the Diabetes dataset and detecting valuable patterns. This study included a medical bioinformatics analysis to predict diabetes. Diabetes was identified using the WEKA computer software as a mining tool. The Pima Indian diabetes database was obtained from the UCI repository and then analysed. The dataset was reviewed and assessed in order to construct a sophisticated model that predicts and identifies the diabetes state. The goal of this study is to compare the accuracy of Naive Bayes, Decision Trees, and (KNN) after using the bootstrapping resampling technique.

INTRODUCTION

People's lives are too hectic these days, and most do not consider their health or how to safeguard it. It can lead to a variety of lifestyle diseases, including diabetes mellitus, one of the most frequent diseases that is strongly connected to the lifestyle we live. If no one knows what it is, it might be the most devastating sickness [2] [8]. Our bodies require energy to function. Blood glucose, which originates from the food we eat, is the primary source of energy. The pancreas is a critical organ in our bodies. It produces insulin. Insulin is a hormone that aids in the level of body sugar (glucose). Glucose is derived from carbs in food and is required by the body to function effectively. Insulin regulates blood sugar levels, preventing them from becoming dangerously low or excessive. Diabetes mellitus is a terrible disease that occurs when the pancreas in the body does not produce enough insulin or does not produce any insulin at all, or when the body is unable to respond to insulin because it is not functioning properly. As a result, the body's blood sugar levels get out of balance, potentially leading to eye disease, stroke, renal disease, high blood pressure, and dyslipidemia. According to the Centers for Disease Control and Prevention, 30.3 million Americans have diabetes, of which 23.1 million have been tested and 7.2 million have not (CDC). There are also 30 million people in India who have this disease, reaching an estimated 80 million by 2030. More than 5 million people died from diabetes in 2015, according to the International Diabetes Federation (IDF). This disease now affects over 415 million people worldwide, including approximately 50 million living in India.

Most algorithms for machine learning are either supervised or unsupervised. A supervised learning algorithm uses what it has learned in the past to make predictions about new or previously unseen data. Unsupervised algorithms, on the other hand, can draw conclusions from datasets. Classification is another name for supervised learning. This study uses the classification technique, which is one of the most commonly used machine learning techniques. It looks at the training data and creates an inferred function that can be used to map new or unseen examples. The primary purpose of the classification approach is to properly forecast the target class for each example in the data. Classification algorithms often rely on the values of data attributes to define the classes. They usually define these classes by examining the attributes of data that is already known to belong to the class. Another name for this process of discovering useful information and patterns in data is Knowledge Discovery in Databases (KDD). There are various phases in it, including data selection, transformation, classification, and evaluation. These phenomena have an impact on a variety of real-world applications, including medical diagnosis, fraud detection, network disruption detection, fault monitoring, pollution detection, biomedical, bioinformatics, and remote sensing. The classification algorithms were tested using the National Institute of Diabetes and Digestive and Kidney Diseases' PIMA Indians Diabetes Dataset, which contains information on female diabetes patients.

LITERATURE REVIEW

Tsao, H. Y., Chan, P. Y. et al,(2018) Diabetic retinopathy (DR) is the major cause of new incidents of blindness in persons of working age in the United States. Evidence from the Diabetes Control and Complications Trial (DCCT) in people with type 1 diabetes mellitus and the United Kingdom Prospective Diabetes Study (UKPDS) in people with type 2 diabetes mellitus suggests that tight control of blood sugar can help prevent the development of microvascular complications like diabetic retinopathy. Prior studies mostly focused on glucose levels, even though there are many other potential risk factors for DR that have been ignored. Diabetic retinopathy, in its most basic definition, is a microvascular consequence of diabetes mellitus. To what extent you've had issues with diabetes in the past is a determining factor. Diabetic retinopathy is an ocular vascular disease that affects nearly all people with diabetes. One can distinguish between non-proliferative and proliferative diabetic retinopathy (PDR). Lack of proliferative alterations in the retina caused by diabetes (NPDR) (BDR). NPDR describes the initial stages of DR, while PDR describes the later stages. Micro aneurysms, hematomas, hard exudates, cotton-wool patches, intraregional microvascular anomalies, and venous beading all contribute to NPDR.

Ye, J., Yao, L. et al,(2020) Diabetes mellitus, a common metabolic disorder, is characterized by chronic hyperglycaemia. The rising incidence and prevalence of Type 2 Diabetes mellitus in adult populations can be related to an increase in the number of ICU admissions (ICU). Diabetes patients receive more than 45 percent of intensive care unit resources, higher than any other category of patients. A impaired immune cell response to disease is another well-established risk factor for diabetic patients admitted to the ICU. Also, these risks can have a big effect on how long diabetic patients in the ICU will live. Few studies have looked into why people with diabetes mellitus die, and the ones that have focused on the intensive care unit haven't found much. The Cox regression model and linear regression models were used in previous forecasting methods. When data, such as cohort characteristics and diabetes duration, are available, these models work optimally.

Kavakiotis, I., Tsave, O. et al,(2017) Recent developments in biotechnology and the health sciences have resulted in an explosion in the volume and variety of data types created in these fields (EHRs). Machine learning and data mining techniques are more important than ever before in the biosciences as researchers strive to intelligently translate all available information into usable knowledge. Diabetes mellitus (DM), an umbrella term for a collection of metabolic diseases, poses a serious threat to global health. Massive volumes of information have been accumulated as a result of in-depth studies of every facet of diabetes (diagnosis, etiopathophysiology, therapy, etc.). Focusing first on Prediction and Diagnosis, then on Diabetic Complications, and finally on Health Care and Management, this study aims to perform a systematic evaluation of the uses of machine learning, data mining techniques, and tools in diabetes research. The use of a number of machine learning methods.

Sun, Y. L., & Zhang, D. L. (2019) In recent years, diabetes has risen to become a leading global health problem. More than 415 million people around the world were diagnosed with diabetes in 2015. Since diabetes is a metabolic disorder with many potential causes, its diagnosis is fraught with difficulty due to the complexity of its etiology, severity of its damage, and complexity of its pathophysiology. Researchers have found that machine learning has become increasingly significant in the study of diabetes as data mining has progressed. Importantly for the diagnosis and treatment of diabetes, machine learning approaches can identify diabetes risk factors and a reasonable threshold of physiological parameters buried in a mountain of data. The purpose of this study is to provide a comprehensive overview of the machine learning methods that have been used for diabetes data screening and diagnosis. This research describes the use of both traditional machine learning methods and more recent, deep learning methods for the early detection and diagnosis of diabetes. These methods have significant biomedical implications.

Basu, S., Johnson, K. T. et al,(2020) Machine learning is a set of techniques used to increase the performance of a learning algorithm ("learner") through iterative model fitting and error correction in order to create predictions, categorize data, or aid in decision-making. Because machine learning requires iterative and repeated data sampling, it works effectively with huge datasets, often incorporating many covariates and a large sample size. Machine learning algorithms, on the other hand, may be capable of detecting complicated, nonlinear, and even subtle connections between variables and outcomes, as well as traits that cluster patients or other forms of data into sub-groups. Clinical epidemiologists in the field of diabetes are increasingly turning to machine learning techniques for assistance when faced with difficult decision-making in the areas of risk stratification, therapy escalation, or the interpretation of complex input data from sources such as images and continuous glucose monitors. Here, we present an overview and critical perspective on the current state of the field of machine learning as it relates to the clinical epidemiology of diabetes.

Xiao, M. X., Lu, C. H. et al,(2021) Early detection and prevention of diabetes mellitus (DM) in high-risk persons, such as those with metabolic syndrome, are critical challenges in preventive medicine and public health. Diabetes complications, according to a World Health Organization research, have far-reaching consequences that go far beyond the scope of the disease itself. Diabetes is a severe threat to community health since it is so prevalent. HDL cholesterol, for example, has previously been related to a lower risk of all three types of stroke in patients with type 2 diabetes: ischemic, haemorrhagic, and total. Furthermore, a negative relationship was established between BMI and the incidence of total, ischemic, and hemorrhagic stroke in type 2 diabetes patients. Patients with type 2 diabetes were more likely than those with type 1 diabetes to have peripheral neuropathy (42.2% vs. 29.1%); additionally, the prevalence of diabetic peripheral neuropathy (DPN) increases with age and duration of diabetes, and is already high (35.0%) after a diagnosis of type 2 diabetes. Diabetic peripheral neuropathy is associated with foot amputation and ulceration, aberrant gait, and fall-related accidents (DPN).

Tripathi, G., & Kumar, R. (2020) Despite the abundance of modern medical technologies, early diabetes mellitus diagnosis remains a tough challenge. In this study, we employ machine learning classification methods to build a model that is both robust enough to handle the challenges of diabetes diagnosis and practical enough to be used in its early phases. The tests are run on the PIDD dataset using four machine learning algorithms: LDA, KNN, SVM, and RF. The dataset contains 768 records, with 8 important features labelled as "diabetic" or "non-diabetic," showing the difference in outcome between the two groups. While we are primarily concerned with improving the model's accuracy, we have also examined its F-score, recall, and specificity to assure its efficacy. To assess the usefulness of these performance indicators, confusion measures (true positive, true negative, false positive, and false negative) are utilized. According to the statistics, Random Forest (RF) had the highest accuracy (87.66%) among the classifiers examined. Because this is the case, we modified the RF classifier to work with our data. We intend to harness the most recent developments in machine learning and artificial intelligence to increase our capacity to foresee other diseases in the future, such as psoriasis and cancer.

Saru, S., & Subashree, S. (2019) In the healthcare industry, there is a lot of sensitive information that must be treated with care. Diabetes mellitus, a potentially fatal condition, is becoming more common in all parts of the world. Scientists and medical practitioners are both looking for an accurate means of predicting who will get diabetes. Multiple machine learning algorithms can examine data from multiple perspectives and synthesize the results into useful conclusions. Massive data sets that are widely available online may provide useful information if the correct data mining technologies are used. The basic goal of providing meaningful insights to users is to discover new patterns and make sense of them. Diabetes consequences include heart disease, kidney failure, nerve damage, and blindness. It is critical to create effective strategies for mining diabetes-related data. The appropriate data mining methods and procedures will be discovered in order for Diabetes data to be efficiently categorised and noteworthy patterns to be identified. Diabetes prognosis was achieved in this study using medical bioinformatics approaches. WEKA, a data mining algorithm, was utilized to establish diabetes diagnoses. The Pima Indians' diabetes data came from a database at the University of California, Irvine. A reliable model for diabetes diagnosis and prognosis was established by studying and evaluating this data set. The goal of this study is to examine the performance of Naive Bayes, Decision Trees, and Random Forests using the bootstrapping resampling method (KNN).

METHODOLOGY

CONVENTIONAL MACHINE LEARNING TECHNIQUES

Diabetes is diagnosed using a variety of epidemiology and genetic factors. Smoking status, eating habits, physical activity, BMI, and other risky epidemiological characteristics are examples. Pathogenic genes, also known as genetic factors, are inherited from parents. As a result, clinicians strive to include all aspects of these factors in order to successfully anticipate and diagnose diabetes; yet, medical researchers found that they were unable to explain the pathophysiology of diabetes. As a consequence of the continual development of artificial intelligence technology, machine learning techniques have been found to be exceptionally suitable for evaluating the tolerable threshold of harmful factors and physiological parameters impacting diabetes. Why is machine learning so promising in the medical field? To begin with, diabetes is a chronic condition, and the course of treatment will generate a large amount of clinical treatment information. Diabetes data may be handled and analysed using machine learning techniques due to the significant benefits of machine learning in solving massive data concerns. The extraction of accurate and relevant information from a large body of data for decision-making is the second goal shared by machine learning and medical diagnostics. Simultaneously, machine learning techniques can prevent human specialists from making errors when they are inexperienced or weary. They are also very stable and accurate when screening for and diagnosing diabetes. Furthermore, machine learning techniques can assist patients in gaining a comprehensive image of their health and the progression of their diabetes. This allows them to arrange their own lifestyles in order to prevent the disease's progression. So, we want to apply machine learning to discover diabetes treatments that are not currently available in the medical industry. This would be critical for early diabetes treatment, proper medication administration, and early recovery. Both supervised and unsupervised learning will be utilised in this research to demonstrate how typical machine learning techniques may be used to screen for and diagnose diabetes mellitus early.

Classification techniques are widely used in pattern recognition or predictive analysis for classifying the data into different classes. Machine learning and artificial neural network technology are beneficial technology that can do so due to the strength of their various classification algorithms supported by these technologies. These technologies are very frequently used in the medical field where predictive analysis is a challenging task; the cause of this is more imbalances and missing values in the data set. Human beings always learn from past experiences and machine always follows the instruction given by human being. So to make a model and train this model in a particular domain, develop a valuable amount of dataset, develop the set of algorithms and check the accuracy of the model using various statistical measurements over the correctly and incorrectly classified instances. In this study, we aim to develop a model in the healthcare application using machine learning for predictive analysis of diabetes using significant features that are closely related to this disease.

The procedures that are used in the building of a model contain several useful steps that are described one by one that explores the logic run behind this study.

The procedures that are used in the building of a model contain several useful steps that are described one by one that explores the logic run behind this study.

B. Preparation of data Because data is the most critical component that permits model training, machine learning approaches are completely reliant on it. The dataset may initially contain many divergences since it was assembled from various sources in a crude manner, which the model may not be able to handle. Pre-processing is thus essential to establish a clean data collection and eliminate any divergences. This entailed filling in the gaps created by missing values, generating new features, subdividing the train-test set's data, encoding the data (which means converting non-numerical data to numerical data), normalising the data, and so on. Another issue that develops during the pre-processing step is data imbalance, which means when there are more samples of one class than the other.

LOGISTIC REGRESSION

The logistic model, often known as the logit model, is a prominent statistical model in statistics that uses a logistic function to model a binary dependent variable in its most basic form; several more complicated variants are possible. Logistic regression, often known as logit regression, is a kind of binomial regression used in regression analysis to estimate logistic model parameters. From a mathematical standpoint, a binary logistic model has a dependent variable with two potential values, such as pass/fail, win/lose, alive/dead, or healthy/ill; they are represented by an indicator variable, with the two values labelled by "0" and "1." The log-odds (the logarithm of the odds) for the value labelled "1" in the logistic model is a linear combination of one or more independent variables (predictors); the independent variables can each be a binary variable (two classes, each coded by an indicator variable) or a continuous variable (any real value). The logistic function, as the name implies, is the function that converts log-odds to probability. The corresponding probability of the value labelled "1" might be somewhere between 0 (certainly the value "0") and 1 (certainly the value "1"); therefore, the labelling. The variant names are taken from the logit, or logistic unit, which is the log-odds scale's unit of measurement. The defining characteristic of the logistic model is that raising one of the independent variables multiplicatively increases the odds of the given result at a constant rate, with each dependent variable having its own parameter. Analogous models with a different sigmoid function in place of the logistic function, such as the probity model, can also be utilised.

SUPPORT VECTOR MACHINE

Support-vector machines (SVMs, also known as vector networks) in machine learning are supervised learning models with corresponding learning algorithms that examine data used for regression and classification analysis. An SVM training algorithm creates a linear model that assigns new examples to one of two categories, making it a non-probabilistic binary linear classifier, given a collection of training examples designated as belonging to one or the other of two categories (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). A SVM model is a representation of the examples as points in space that has been mapped in such a way that the examples of the various categories are separated by as wide of a gap as feasible. Then, based on which side of the gap they fall, new examples are projected into that same space and predicted to belong to a category.

Artificial Neural Network

ANNs are a powerful tool for analysing complex clinical data and are especially well suited for predicting disease diagnostic outcomes. During the training phase, ANNs use known data to discover the complex relationships between the input and the output. After training, ANNs may be used to predict the output value of given input data. ANNs are frequently successful for a variety of complex, non-linear, or missing data. Metabolic syndrome (Mets) refers to a range of clinical syndromes caused by protein, lipid, and carbohydrate imbalances in the body, as well as other metabolic irregularities. Mets is a well-known risk factor for the development of chronic diseases such as cancer, type 2 diabetes, cardiovascular disease, and chronic renal disease. As a result, it is critical to use ANN to diagnose metabolic syndrome in order to determine the cause of diabetes. Hirose et al. used ANN to predict the 6-year incidence of MetS using clinical data. The ANN is chosen above the standard logistic regression technique for estimating the risk of MetS based on sex, age, BMI, waist circumference, waist-to-height ratio, hip circumference, and systolic and diastolic blood pressure. Darko Ivanovic proposed employing a feed-forward ANN with back propagation as the training approach to anticipate MetS. The paper's contributions include the simple input vector and a more comprehensive search for the appropriate ANN design.

RESULTS AND DISCUSSION

Table 1: Accuracy of classifiers in % after GA-based Features Selection

Classifiers	Accuracy	ROC Area	F-Measure	ROOT Mean Squared Error
Bayes Net	78.58	0.823	0.829	0.4008
Multilayer Perceptron	69.594	0.644	0.658	0.5617
Simple Logistic	68.35	0.658	0.683	0.5617
Decision Table	74.86	0.791	0.743	0.4216
Random Tree	76.29	0.691	0.711	0.4814

Table 2: Accuracy of classifiers in % after PSO-based Features Selection

Classifiers	Accuracy	ROC Area	F-Measure	ROOT Mean Squared Error
Bayes Net	77.474	0.829	0.769	0.4008
Multilayer Perceptron	71.4844	0.649	0.699	0.534
Simple Logistic	76.8229	0.711	0.757	0.4814
Decision Tree	74.98	0.777	0.791	0.4216
Random Tree	72.84	0.639	0.649	

Reference	Proposed model / Method	Dataset Used	Purpose	Accuracy Achieved (%)
N. Gupta et.al (2013)	Decision Tree	PIMA Indian Diabetes Data	To predict diabetes	81.33%
P. Yasodha, M.Kannan (2011)	Bayes Net	A hospital repository	To predict diabetes	66.2%
A.Iyer et.al (2015)	Decision Tree	PIMA Indian Diabetes Data	To predict diabetes	To predict diabetes
K.Rajesh, V.Sangeetha (2015)	Decision Tree	PIMA Indian Diabetes Data	To predict diabetes	87%
Lee (2014)	Decision Tree	National Health and Nutrition Examination Survey	To predict diabetes	67%

CONCLUSION

Many data mining approaches and applications have been studied or reviewed. Machine learning techniques have been used to an application of medical data sets, most notably the machine Diabetes dataset. Machine learning approaches benefit from varied data sets. UCI provided us with a diabetes data set of 768 records. Individual algorithms and the proposed method are compared in this study. We are utilising 10 cross validation tests to assess the performance of these machine learning categorization methods. In this study, the proposed method has a high accuracy .however decision Stump has a lesser accuracy (78.58% accuracy value) than the other methods.

REFERENCES

1. Tsao, H. Y., Chan, P. Y., & Su, E. C. Y. (2018). Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. BMC bioinformatics, 19(9), 111-121.
2. Ye, J., Yao, L., Shen, J., Janarthanam, R., & Luo, Y. (2020). Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Medical Informatics and Decision Making, 20(11), 1-7.
3. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.
4. Sun, Y. L., & Zhang, D. L. (2019). Machine learning techniques for screening and diagnosis of diabetes: a survey. Tehnički vjesnik, 26(3), 872-880.
5. Basu, S., Johnson, K. T., & Berkowitz, S. A. (2020). Use of machine learning approaches in clinical epidemiological research of diabetes. Current Diabetes Reports, 20(12), 1-19.
6. Xiao, M. X., Lu, C. H., Ta, N., Wei, H. C., Haryadi, B., & Wu, H. T. (2021). Machine learning prediction of future peripheral neuropathy in type 2 diabetics with percussion entropy and body mass indices. Biocybernetics and Biomedical Engineering, 41(3), 1140-1149.
7. Tripathi, G., & Kumar, R. (2020, June). Early prediction of diabetes mellitus using machine learning. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1009-1014). IEEE.
8. Saru, S., & Subashree, S. (2019). Analysis and prediction of diabetes using machine learning. International journal of emerging technology and innovative engineering, 5(4).
9. 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, 2013
10. Yashoda and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in Waikato", International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2011.
11. A. Ayer, J. S and R. Sumbala, "Diagnosis of Diabetes Using Classification Mining Techniques", IJDKP, vol. 5, no. 1, pp. 01-14, 2015
12. . Sharmila and S. Manickam, “Efficient Prediction and Classification of Diabetic Patients from big data using R,” International Journal of Advanced Engineering Research and Science, vol. 2, Sep 2015.
13. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors. Big Data. 2015 Dec;3(4):277-87. doi: 10.1089/big.2015.0020.
14. Zhao Ming, Wang Xiaoxia, & Zhu Xiaowei. (2014). Understanding diabetes from the diagnosis of diabetes mellitus. Journal of Diagnostics Concepts & Practice, 2, 226-228.
15. Rajesh, K. & Sangeetha, V. (2012). Application of Data Mining Methods and Techniques for Diabetes Diagnosis. International Journal of Engineering and Innovative Technology (IJEIT), 2(3). 224-229.