Proceedings of the Third International Conference on Smart Electronics and Communication (ICOSEC 2022)
IEEE Xplore Part Number: CFP22V90-ART; ISBN: 978-1-6654-9764-0

2022 3rd International Conference on Smart Electronics and Communication (ICOSEC) | 978-1-6654-9764-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOSEC54921.2022.9952107

Diabetes Prediction based on Supervised and Unsupervised Learning Techniques – A Review

Sivaraman M
Ph.D. Research Scholar, Department of Computer Science,
Dr. SNS Rajalakshmi College of Arts and Science,
Coimbatore, Tamilnadu, India.
sivaramanranjith@gmail.com

Sumitha J
Assistant Professor, Department of Computer Science,
Dr. SNS Rajalakshmi College of Arts and Science,
Coimbatore, Tamilnadu, India.
sumivenkat2006@gmail.com

***Abstract*—Diabetes is a disorder that develops in the human body when blood glucose or sugar levels are extremely high. Machine Learning (ML) is subfield of Artificial Intelligence (AI) that is built on the idea that systems and machines will evaluate and interpret data, learn from it, and make a decision with no human participation. Medical fields that operate with massive amounts of data have recognized by the of machine learning technology. Health care systems can work more effectively and gain a competitive advantage by extracting insights from this data. The general goal of this study is to discover and observe the working and performance of various machine-learning algorithms. This paper work discusses the possibilities of employing machine-learning technology in healthcare.**

Keywords—Data Mining, Diabetes, Machine Learning, Health Care.

I. Introduction

A. Diabetes

Diabetes is split into three types based on the human body condition. The human body, pancreatic cells stop producing insulin. This type of diabetes is called type1. This type of patient needs to take insulin daily [5]. The human body produces insulin. But did not use produced insulin in the human body properly. This type is called type-2 [1]. It is the standard type and mainly develops in adults. Gestational diabetes only arises in pregnant females. This cause affects pregnancy and the baby's health [3]. The most common symptoms are frequent urination, increased thirst, slow healing sores, Weight loss, Extreme hunger, fatigue, Presence of ketoses in the urine, blurred vision, irritability, frequent infections, and vaginal infections, skin infections [18]. So, diabetes patients need to follow some habits in real life, such lose extra weight, stopping smoking, doing regular physical activity, drinking more water, eating healthy plant foods, and stopping taking sugar products [8].

B. Machine Learning

The process of Mining of data is to find the inherent or unseen data and / or knowledge which are possibly useful to the people who may not distinguish in advance[16]. This extraction is from the enormous, unfinished, missing, unclear and unsystematic data [2]. The very important change between the traditional data analysis and mining of data process is that in data analysis is performed with the help of queries, forms and reports and analysis of applications [11].

But mining of data extracts info and realize information on the principle of no strong assumptions. This process analyses data already exist in a database and finds the hidden patterns from it. These extracted patterns must be meaningful to the user in some or the other sense like may be economically beneficial to the patient or from a patient safe treatment point of view. The term "data mining" is also known as device learning, predictive analytics, and knowledge discovery in databases(KDD) [7]. KDD has some stepwise process where mining of data is considered one of the phases.

In this area of digitization the huge amount of data are generated and they are simply stored on hard disks of growing size. It is a great challenge to extract meaningful information from these data [4]. The challenges also include the time duration between the generations of collected data with the understanding of collected data. Therefore techniques of mining data can find possibly valuable hidden information with the help of huge amounts of data. Also the data mining opportunities are increased when the hard disk drive capacity is increased and more and more data is stored. The processing power of regular computers also increases according to Moore's law [15].

A traditional data analysis is performed by querying the data to answer specific questions or verify specific hypotheses. To get the information all possible combinations need to be checked with the number of possible column values. If the numbers of columns are increased in the query, then the numbers of queries also increase. While using data mining, data analysis does not use predefined theory on which the data analysis is based. Here data column names are made input and / or output. Then the data mining system will automatically show the relations between both input columns and output columns [13].

The chosen data mining algorithm performs the calculations and generates patterns. Because there is no predefined theorem for possible and unforeseen relationship, the new relationship may be generated and that becomes the advantage of data mining [5]. There are possibilities of missing those relationships because of the predefined theory.

C. Data Mining Steps

Mining of Data makes use of various algorithms to perform a variety of different tasks. These algorithms examine the sample data collected from original database of a problem and pre-processing of those data it determines a model that is closer to solve the problem [10].

Fig 1 shows the Knowledge discovery-process. For extraction of information or hidden knowledge, data mining include following steps:

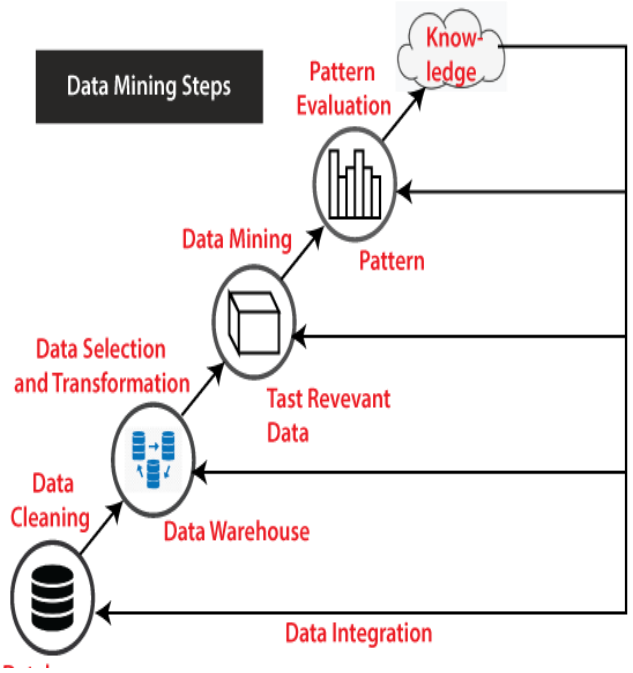

Fig.1. KDD-Process

a) Data Selection: This step identifies the needed datasets and relevant fields from the database. It extracts the data which are of concern. The selection of data is a subcategory of the attributes of data which were collected for the activity of data collection [8]. Example: During the data collection process, the researchers might choose to check or collect data from the physical checkups of a patient. But here, during this phase, some specific fields like height or weight values are selected for the use of operations of data mining.

b) Data Pre-processing and Cleaning: At this stage, pre-processed data are generated. It checks the data records for the wrong values like unacceptable data for characterized objects of data, and available data for statistical qualities. In reality, various attributes generally have lost values. So, users or experts reject these attributes or values from the sets of data. Also they add lost values of data with calculated values using some formula or predefined values [9].

*c) Data Transformation:*In this step many unusual kinds of modifications are made applicable to the data items and they are made suitable for the specific purpose of data mining. The changes are measured as data space is changed on the basis where all the records exist as facts. Example: the patient's weight in milligram has perhaps many precisions. But it may be converted into gram or kilogram. Also, the data mining experts may change the value of height for further simplification of data for mining process of data [9].Also, some more fields are also available which may result from other attributes of data stored. Example: Infection duration in a patient may be counted by deduction of the treatment dates from last date to first date [14].

*d) Model Building:*This step generates a relevant model. Any algorithm suitable from the list of data mining algorithms is made practical on the stored data [6]. This algorithm is chosen by the experts or researchers and it is dependent on the analysis type on which it is carried out. Many algorithms are available, but mainly two types of groups are available. First group include the algorithms that describe the data and second group include algorithms that implement the prediction for future cases [11]. These groups of algorithms may also be combined together on the base of the performance of clustered data, classified data, and mining using association rules between the same types of data.

e) Model Interpretation: In this step finally model interpretation or evaluation is done to achieve the knowledge. The necessity is that results will be checked or evaluated in terms of its correctness, or usefulness or it's meaningfulness [12]. The result evaluation becomes the basis for the experts to choose some steps backward and complete them again in different ways using some other way. Therefore, the process of knowledge discovery becomes an repetitive process. Whenever the discovery process has completed, the explored performance was measured as mined expertise by the professionals. Then the generated results are eventually deposited in some application with the help of a data miner tool and in a format of specifications for future access or use or specific purpose.

D. Machine Learning Types

To learn something is to get info on something by study, experience, or being taught. Artificial Intelligence has made the computer to think. Artificial Intelligence has made the computer system nearly as intelligent as human being. Machine learning is considered as the subtype for the study of AI [17]. Many researchers are in the opinion that when proper usage of computers in getting the knowledge is not done, we are unable to develop the intelligence. The different kinds of machine learning techniques are supervised, semi-supervised, unsupervised, evolutionary learning, reinforcement, and deep learning. Based on these techniques the classification of the data set is performed.

a) Supervised Learning: One of the tasks of data mining is supervised learning which assumes a job from categorized sample data set. The set of sample data include a set of examples of sampling. Therefore, in method of supervised learning, every sample is in a pair containing an input item which is normally a vector and the anticipated value for output that is known as the supervisory signal. An algorithm for supervised learning examines the sample set of data and generates an assumed task, through which new examples are mapped later on. At the top the algorithm allows to determine the class labels correctly. It also decides the not known instances which require generalization of the training data in a reasonable way [19].

The entire machine learning algorithms practically uses supervised learning. Supervised learning has two variables, namely output variables Y and input variables X. The methods learn from the following mapping equation, from the input variable into the output variable.

$$Y = f(X)$$

Here, the target is to forecast the mapping work so correctly that while the new input data item (x) is added-in, output variables (Y) are predicted from that data. The algorithm is known as supervised learning because the algorithm learning process works as a teacher and supervises the learning process. Here the right answers areidentified;thealgorithmrepetitivelypredictsonthetrainingdataandifneededcorrected by the teacher. Learning is stopped when a satisfactory performance level of the algorithm is achieved [19]. The problems arises from supervised learning are collected as classification and regression problems where,

Classification method predicts the answers either of Yes or No, Example. “Are reports positive for a cancerous tumor?”, “Does this medicine maintain the quality standards?” A classification problem arises when the prediction output variable falls into a category like “yellow” and “green” or “no disease” and “disease”.

Regression method gives the answer of “How many” or “How much”. A regression problem arises when the prediction output variable is a real value like “salary” or “sugar value”. All the procedures related to classifying and regressing fall under machine learning of supervised method.

- Decision-Tree
- Logistic-Regression
- Support-Vector-Machine
- Naïve-Bayes
- K-Nearest Neighbor
- Random-forest
- Polynomial-regression
- Linear -regression

b) Un Supervised Learning: The unsupervised learning method tries to discover unknown organization in unlabeled set of data. Because the training data patterns provided to the learner algorithm are unknown, error or any type of signal is not available to check the probable result. Unsupervised method has only input data (X) and corresponding output variables are not available. It creates a model which is the original structure or distribution of data so that learning more about the data is possible. They are known as unsupervised learning algorithms because no correct answers are available and also no teacher is available for correction of the solutions. Algorithms are dependent on their own resources to find and display the interesting structure in the data. In these algorithms true answers or goals are not delivered. It attempts to search out the relationships among the inputted data and classification done on those data [20]. It is identified as density estimation. The problems arose from unsupervised learning methods are further put together into clustering and association problems.

Clustering creates groups depending on relationships of the data. A clustering problem is where the groups or the clusters inherent data are learned. Example: the group of the customers who have the same purchasing behavior.

Association problem is where rules are discovered that describe the bigger part of the data. Example: the customers who buy item X will also like to buy item Y. All algorithms related to clustering fall under procedures of unsupervised learning.

- Hierarchical clusters
- K-means clusters
- Apriori algorithm

c) Semi Supervised Learning: These type of algorithms have the problems where a huge amount of input data (X) are available but only some of the output data are labeled (Y). These kinds of problems fit in between both supervised and unsupervised learning [20]. E.g. in an image some of the sub images are labeled like a dog, a cat, or a person but rest of all part of the image is unlabeled. Almost all actual machine learning problems in reality are this type of learning. These problems are time-consuming and expensive because labeling of data is required. This is actually done with the help of the field experts. Unlabeled data are easy to collect, store and cheap. Unsupervised learning techniques find and learn the structure of the input variables. Whereas supervised learning techniques guess the good predictions for the un-labeled data, put those data back as training data to the supervised learning algorithm. Then it uses the created model for the predictions on new not known unlabeled data. The technique of semi supervised learning is a part of technique of supervised learning. It uses unlabeled data for the purpose of training. Usually small sizes of labeled-data are used with a big sized unlabeled-data. Semi-supervised understanding falls within unsupervised- understanding method which uses unlabeled-data and supervised understanding method which uses labeled-data.Psychiatristsnormallyusethismethodofunderstanding.Thealgorithmadvisesabout the incorrect answer but does not advise about the methods to correct the answers.Various possibilities were tested and explored till it finds the correct answer. This learning method is considered as understanding from a critic. Enhancements are not declared. Reinforcement learning differs with learning method for supervised data where presentation of precise input data and output data sets is not done. Also, no optimum actions are mentioned clearly [20].

d) Evolutionary Learning: The evolution learning is measured as an understanding procedure of changes of living entities that grow at their own existence levels without having any kind of mechanisms. This model is used to check the accuracy of the solution [20].

e) Deep Learning: This understanding method uses different types of procedures. In mining of data, these methods create the advanced model construction. It uses different processing layers graphs that use many linear and non-linear transformation of data.

II. LITERATURE REVIEW

TABLE 1: The results of diabetes predictions using supervised learning presented

Ref	Comments	Algorithms
[1]	85% of the techniques were relevant to supervising-learning based, and 15% are relevant to un-supervised techniques. This was observed that SVM techniques is the most frequently applied mellitus classification method.	SVM
[2]	Different types of supervised ML algorithms were applied to determine which method is best suited for diabetes detection.	SVM
[3]	Three classification methods, including DT, SVM, and NB, are applied to detect diabetes during the initial stages.	NB
[4]	The author utilized RF, DT, and NB.	NB
[5]	LR, Gradient Boosting, and NB techniques are applied to calculate the future diabetic risk.	Gradient Boosting.
[6]	Involved various external characteristics responsible for diabetes in addition to usual aspects such as BMI, age, glucose, insulin, and etc.The Classification accuracy values is increased the new data set.	AdaBoost - 98.8%
[7]	The author used 4 famous ML-algorithms: NB, KNN, SVM, and DT(C4.5), on adult population dataset to predict the disease.	DT(C-4.5)
[8]	Prediction models employing LR method on Canadian-patients aged 18 to 90 years to identify people at risk of getting disease.	GBM and LR
[9]	The author collect the 952 samples from both offline and online survey with 18-questions on family history, style of life, health to predict Type II diabetes.	RF-94.10%
[10]	Type 2 diabetes detection through multiple regression analysis algorithms such as Glmnet, XGBoost, RF and LightGBM.	XGBoost-88.1%
[11]	Diabetic patients were monitored using J48, SMO, NB, RF, ZeroR, Simple logistic, and OneR.	SMO
[12]	CNN, LSTM, and their combinations were recommended for extracting intricate temporary dynamic features from the input HRV data.SVM was used for classification.	ML to Deep Learning
[13]	Analyze several articles for feature choices, various data samples, classification methods such as SVM, KNN, RF, and so on, as well as the Deep-Learning method.	ML to Deep Learning

TABLE 2: The results of diabetes predictions applying Unsupervised learning presented

Ref	Comments	Algorithms
[14]	A prediction model for diabetes was created using 85% parts is clustering techniques and 15% part is non-clustering techniques.	K-means-78%
[15]	K-Means and Hierarchical cluster techniques used to predicting diabetes	K-means

TABLE 3: The results of diabetes predictions using Both supervised learning and Unsupervised learning presented

Ref	Comments	Algorithms
[16]	Apriori-algorithm were applied to create a good relationship of diabetes blood glucose and BMI Level. RF, ANN, and K-means cluster methods are used for predict the disease.	ANN-75.7% accuracy
[17]	Outlier prediction- K-mean and Classification- SVM.	K-Mean and SVM
[18]	The author applied LR, K-Mean, and PCA algorithms.PCA algorithm is booting in to the K-Mean algorithm.	LR, PCA, and K-Mean
[19]	The author analyzed k-mean and SVM methods. Finally, SVM method applied.	SVM and K-mean methods 99.64%

III. Results and Discussions

Diabetes is among the most serious illnesses. Age, obesity, a physically inactive lifestyle, hereditary diabetes, poor food diet, blood pressure, etc. are the primary causes of this disease.

Table-1 reveals that versions of decision-trees, such as XG-Boost, RF, and Ada-Boost,are the most frequently applied supervised classification methods.The recent trend is changing deep learning methods from machine learning methods. ANN method is a recent and highly famous ML methodology that also performs in a range of distinct facets.

Table-2 is unsupervised learning algorithms such as LDA, PCA, are usually utilised for dimension reduction. Useless features in diabetic data samples are misguiding the accurateness of the classifier.

Hence, people can still have a mixture of both supervised and unsupervised methods for the good prediction and diagnosis of disease. This research could be further extended with block-chain and Internet in of things order to send the predictions in a secure and transparent manner to physicians, doctors, diagnosing labs, etc.

IV. Conclusion

This article reviewed and evaluated the state-of-the-art for predicting and identifying diabetes. Early diagnosis is necessary for this chronic disorder in order to avoid dangerous stages. Several supervised learning methods, including SVM, Naive Bayes, Decision trees, etc., were used. As a result, classifiers built on decision trees may be able to find out diabetes at an early days. It is evident that the model, when used in combination with an unsupervised learning technique like K-Mean and PCA increases the accuracy and precision of diabetes prediction. Diabetes has also been accurately identified and analyzed using K-Mean and SVM. For better outcomes in the diabetic care system, deep learning-based algorithms like ANN, CNN, etc., also perform well.

REFERENCES

[1] I. Kavakiotis, A.Salifoglou, O.Tsave, I.Vlahavas, I.Chouvarda, and N.Maglaveras, "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology journal, vol. 15, pp. 104-116, 2017.

[2] M. A. Sarwar, W. Hamid, N. Kamal, and M. A. Shah, "Prediction of diabetes using machine learning algorithms in healthcare", 2018 24th International Conference on Automation and Computing (ICAC), pp. 1-6, September 2018.

[3] D. S. Sisodia and D. Sisodia, "Prediction of diabetes using classification algorithms", Procedia computer science, vol. 132, pp. 1578-1585, 2018.

[4] T. Gangil and N. Sneha, "Analysis of diabetes mellitus for early prediction using optimal features selection", Journal of Big data, vol. 6, pp. 13, 2019.

[5] R. Birjais, R. Chauhan, A. K. Mourya,and H. Kaur, "Prediction and diagnosis of future diabetes risk using Machine Learning Approach", SN Applied Sciences, vol. 1, pp. 1112, 2019.

[6] V. Vaidehi and A. Mujumdar, "Diabetes prediction using Machine Learning Algorithms", Procedia Computer Science, vol. 165, pp. 292-299, 2019.

[7] I. H. Sarker and M. F. Faruque, "Performance analysis of Machine Learning Techniques to predict diabetes mellitus", International Conference on Electrical Computer and Communication Engineering (ECCE), pp. 1-4, February 2019.

[8] H. Lai, H. Huang, K. Keshavjee, A. Guergachi and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques", BMC endocrine disorders, vol. 19, pp. 1-9, 2019.

[9] S. Garg and N. P. Tigga, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods", Procedia Computer Science, vol. 167, pp. 706-716, 2020.

[10] L. Kopitar, L. Cilar, P. Kocbek, G. Stiglic, and A. Sheikh, "Early detection of type 2 diabetes mellitus using Machine Learning-based prediction models", Scientific reports, vol. 10, pp. 1-12, 2020.

[11] J. Lloret, A. Rghioui, A. Oumnad, and S. Sendra, "A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms In Healthcare", Multidisciplinary Digital Publishing Institute, vol. 8, pp. 348, September 2020.

[12] G. Swapna, K. P. Soman, and R. Vinayakumar, "Diabetes detection using deep learning algorithms", ICT Express, vol. 4, pp. 243-246, 2018.

[13] S. T. Ganesh, J. Chaki, S. K. Cidham, and S. A. Theertan, "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review", Journal of King Saud University-Computer and Information Sciences, 2020.

[14] D. M. Kumar, D. C. Sujatha, and M. C. Peter, "Building predictive model for diabetics data using K Means Algorithm", International Journal of Management IT and Engineering, vol. 8, pp. 58-65, 2018.

[15] M. T. Islam, M. Raihan, F. Farzana, H. S. Mondal, and M. G. M. Raju, "An Empirical Study to Predict Diabetes Mellitus using K-Means and Hierarchical Clustering Techniques", 10th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-6, July 2019.

[16] M. A. Iqbal, T. M. Alam, Y. Ali, A. Wahab, S. Ijaz, et al., "A model for early prediction of diabetes", Informatics in Medicine Unlocked, vol. 16, pp. 100204, 2019.

[17] S. T. A. Niaki, M. Alirezaei,and S. A. A. Niaki, "A bi-objective hybrid optimization algorithm to reduce noise and data dimension in diabetes diagnosis using Support Vector Machines", Expert Systems with Applications, vol. 127, pp. 47-57, 2019.

[18] C.U. Idemudia, C. Zhu, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques", Informatics in Medicine Unlocked, vol. 17, pp. 100179, 2019.

[19] O. Yildiz and S. Afzali, "An effective sample preparation method for diabetes prediction", Int. Arab J. Inf. Technol, vol. 15, pp. 968-973, 2018.

[20] S. Rawat, T. Chauhan, S. Malik, P. Singh, "Supervised and Unsupervised Machine Learning based Review on Diabetes Care", 7th International Conference on Advanced Computing and Communication Systems (ICACCS), June 2021.