Diabetes Prediction Using Machine Learning Algorithm

[1]Premkumar P Reddy, [2]Dhanamma Jagli

[1]PG STUDENT, [2]Assistant Professor
Department of MCA,
Vivekanand Education Society Institute of Technology, Mumbai, India

***Abstract*: Nowadays, Diabetes Miletus has become a common disease to the mankind from young to the old persons. The growth of the diabetic patients is increasing day-by-day due to various causes such as bacterial or viral infection, chemical substance in food, bad diet, change in lifestyles, eating habit, environment factors, etc. Hence, it is necessary to diagnos the diabetes to save the human life from diabetes. The analytics is a process of identifying the hidden patterns from large amount of data to derive conclusions. In health care, this analytical process is carried out using various machine learning algorithms for examing medical data to build the machine learning models to carry out medical diagnoses. This paper presents a diabetes prediction system to treat diabetes. Moreover, this paper identifies the various approaches to improve the accuracy in predicting diabetes using medical data with various machine learning algorithms.**

***Keywords*: Medical diagnosis, Medical data analytic, Diabetes disease, Prediction, Neural networks, Machine learning algorithm.**

I. .INTRODUCTION

Diabetes is a metabolic disease that causes high blood sugar. The hormone insulin moves sugar from the blood into your cells to be stored or used for energy. In that case body doesn't make enough insulin it leads to diabetic. Diabetes can be divided into two class such as class 1 and class 2. Class 1 diabetes is an autoimmune disease. In diabetes, the body destroys the cells that are producing insulin to absorb the sugar to produce energy. This class can be caused due to regardless of obesity. This is the increase of body mass index (BMI) than the normal level of BMI of an individual [2]. Class 1 diabetes can occur in childhood or adolescence age. Class 2 diabetes usually affects the adults who are obese. In this class, the body resists observing insulin or fails to produce insulin. Class 2 generally occurs in the middle or aged groups [1]. Moreover, there are other causes for diabetes such as bacterial or viral infection, toxic or chemical contents in food, obesity, bad diet, change of lifestyles, eating habit, environment pollution, etc. Diabetes leads various diseases such as kidney damage, nerve damage, heart disease, foot ulcers, etc.
Data analytic is a process of analyzing and identifying the hidden patterns from large amount of data to derive conclusions. In health care, this analytical process is carried out using machine learning algorithms for analyzing the medical data to build machine learning models to carry out the medical diagnoses. Machine learning is a type of Artificial Intelligence (AI) that enables a system to learn by itself and develop the knowledge models to make decision by predicting the unknown data. The machine learning algorithms can be divided into three types such as supervised learning, unsupervised learning and semi-supervised learning. In supervised learning, labelled data helps the model to learn from data. Supervised learning is used when human expertise does not exist such as (navigating on Mars), & where humans are unable to explain their expertise such as (speech recognition). A Solution that changes in time series (routing on a computer function) and to solution needs to be adapted to particular cases (user biometrics). The supervised learning algorithms are classified into various types such as probability-based, function-based, rule-based, tree-based, instance-based, etc. The unsupervised learning is the descriptive type of learning. This type of learning is used to describe or summarize the data. The examples of the unsupervised learning algorithms are association rule mining, clustering etc. The semi-supervised learning is the combination of supervised learning and unsupervised learning. This paper presents a diabetes prediction system to treat the diabetics. However, the supervised learning algorithm is used to learn the data and to develop diabetes prediction system for treating diabetes. The accuracy of this system is improved using pre-processing technique.

2. LITERATURE REVIEW

THIS SECTION REVIEWS VARIOUS RESEARCH WORKS THAT ARE RELATED TO MY PROPOSED WORK. MOHAMMED ABDUL KHALEEL ET AL CONDUCTED A SURVEY ON DATA MINING TECHNIQUES ON MEDICAL DATA FOR IDENTIFING LOCALLY FREQUENT DISEASES. THE MAIN FOCUS OF THIS SURVEY IS TO ANALYZE THE DATA MINING TECHNIQUES REQUIRED FOR MEDICAL DATA ANALYSIS THAT IS ESPECIALLY USED TO DISCOVER LOCALLY FREQUENT DISEASES SUCH AS BREAST CANCER, HEART LUNG CANCER, AILMENTS USING CLASSIFICATION AND REGRESSION TREE (CART) ALGORITHM AND THE DECISION TREE ALGORITHMS SUCH AS ID3, C4.5. VAISHALI AGGARWAL *ET AL*, HAS DEMONSTRATED A PERFORMANCE ANALYSIS OF THE COMPETITIVE LEARNING ALGORITHMS ON GAUSSIAN DATA FOR AUTOMATIC CLUSTER SELECTION AND ALSO STUDIED AND ANALYZED THE PERFORMANCE OF THESE ALGORITHMS AND RANDOMIZED RESULTS HAVE BEEN ANALYZED ON 2-D GAUSSIAN DATA WITH THE LEARNING RATE PARAMETER KEPT SIMPLE FOR ALL ALGORITHMS. ALGORITHMS USED IN THIS WORK INCLUDE COMPETITIVE LEARNING ALGORITHM, CLUSTERTING ALGORITHM AND FREQUENCY SENSITIVE COMPETITIVE LEARNING ALGORITHM. SUPERVISED LEARNING MACHINE ALGORITHMS ARE USED FOR CLASSIFICATION OF THE GAUSSIAN DATA K. SRINIVAS DEVELOPED APPLICATIONS OF DATA MINING TECHNIQUES IN HEALTHCARE AND PREDICTION OF HEART

ATTACKS. IN THIS RESEARCH MEDICAL PROFILE USED SUCH AS AGE, SEX, BLOOD PRESSURE AND BLOOD SUGAR AND PREDICTED THE LIKELIHOOD OF PATIENTS GETTING A HEART AND KIDNEY PROBLEMS .

DARCY A. DAVIS INDIVIDUAL PROPOSED DISEASE RISK PREDICTION BASED ON MEDICAL HISTORY. THIS PAPER ALSO PREDICTS EACH PATIENT'S DISEASE RISKS BASED ON THEIR OWN MEDICAL HISTORY DATA. IN THIS DATASET ARE USED FOR MEDICAL CODING AND COLLABORATIVE ASSESSMENT AND RECOMMENDATION ENGINE (CARE) INFORMATION TECHNIQUE. FROM THIS LITERATURE, IT CAN BE OBSERVED THAT THE MACHINE LEARNING ALGORITHMS PLACE A SIGNIFICANT ROLE IN KNOWLEDGE DISCOVERY FORM THE DATABASES ESPECIALLY IN MEDICAL DIAGNOSIS WITH THE MEDICAL DATA.

3. PROPOSED WORK

This section presents the diabetes prediction system for diabetes treatment. Figure illustrates the flowchart chart representation of the system model. Initially, the dataset is given into the data pre-processing module. The pre-processing module removes the irrelevant features from the dataset and gives the pre-processed dataset with relevant features to the machine learning algorithm. Then, this machine learning algorithm develops a learning model from the pre-processed dataset. This model is known as knowledge model. Furthermore, the diabetes is predicted for a person's medical data using the learning model.

Figure 1 Flowchart representation of diabetes predication system

3.1 PROPOSED ALGORITHM

NAÏVE BAYES

IT IS BASED ON BAYES' THEOREM WITH THE INDEPENDENCE ASSUMPTIONS BETWEEN PREDICTORS. A NAIVE BAYESIAN MODEL IS SIMPLE TO BUILD, WITH NO COMPLEX PARAMETER WHICH MAKES IT VERY USEFUL FOR LARGE DATASETS. BEING ITS SIMPLICITY, THE NAIVE BAYESIAN CLASSIFIER IS WIDELY USED ALGORITHM BECA USE IT OUTPERFORMS MORE SOPHISTICATED CLASSIFICATION METHODS. THIS ALGORITHM WORKS ON THE ASSUMPTIONS THAT ARE CLASSIFYING CATEGORICAL DATA, OCCURRENCES OF AN EVENT INDEPENDENT AND PREDICT ACCURATELY ON HIGH DATASET.

RANDOM FOREST

RANDOM FOREST IS KNOWN TO BE FLEXIBLE, USER FRIENDLY MACHINE LEARNING ALGORITHM THAT PRODUCES, EVEN WITHOUT HYPER-PARAMETER TUNING, A GREAT RESULT MOST OF THE TIME. IT IS A WIDELY USED ALGORITHMS, BECAUSE OF ITS SIMPLIFIED AND DIVERSIFIED USES..

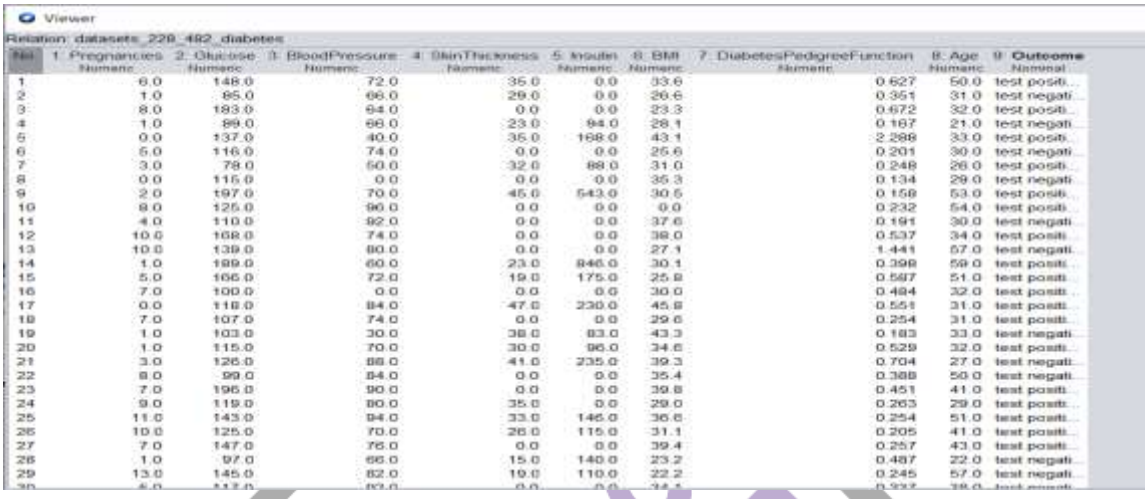
RANDOM FOREST PRODUCES RELEVANT RESULTS MOST OF THE TIME. RANDOM FOREST BUILDS VARIOUS DECISION TREES AND COMBINES THEM TO GET MORE SUITABLE RESULTS. THE THEORY BEHIND RANDOM FOREST IS THE OVERLAPPING OF RANDOM TREES, AND IT CAN BE SCRUTINIZED EASILY.

3.2 IMPLEMENTATION

This experiment is conducted using NETBEANS, WEKA software with the configuration of computer system 4 GB RAM, Intel(R) Core (TM)2 CPU 1.73 GHz Processor, Windows 7 64-bit operating system. For the conduction of this experiment, the medical dataset has been collected from kaggle (website). This dataset contains medical data of 768 persons. This medical data (dataset) includes 8 features of the persons such as age, plasma glucose concentration, number of times pregnancies, blood pressure, skin

fold thickness, insulin level, body mass index (BMI), diabetes pedigree function, and the results such as whether the person has diabetes (positive) or not (negative).

The correlation-based selection technique is employed for data pre-processing so as to get rid of the irrelevant features. differing types of supervised machine learning algorithms namely probabilistic-based naïve Bayes (NB), decision tree-based random forests (RF). The test methods like 10-fold cross validation (FCV), use percentage split with 66% (PS), and use training dataset (initially, the diabetes dataset is given into the machine algorithms (NB, RF) and therefore the accuracy with different test methods (FCV, PS, UTD) is noted. Then, the dataset is given into the correlation-based feature selection to perform the pre-process and therefore the refore the irrelevant feature are faraway from the dataset and the dataset is given to the machine learning algorithm (NB, RF) and therefore the accuracy is noted with different test methods (FCV, PS, UTD).


Viewer

Relation: datasets_228_482_diabetes

No.	1 Pregnancies Numeric	2 Glucose Numeric	3 BloodPressure Numeric	4 SkinThickness Numeric	5 Insulin Numeric	6 BMI Numeric	7 DiabetesPedigreeFunction Numeric	8 Age Numeric	9 Outcome Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	test positi...
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	test negati...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	test positi...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	test negati...
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	test positi...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	test negati...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	test positi...
8	0.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	test negati...
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	test positi...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	test positi...
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	test negati...
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	test positi...
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	test negati...
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	test positi...
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	test positi...
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	test positi...
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	test positi...
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	test positi...
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	test negati...
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	test positi...
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	test negati...
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	test negati...
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	test positi...
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	test positi...
25	11.0	143.0	94.0	33.0	146.0	36.6	0.254	51.0	test positi...
26	10.0	125.0	70.0	26.0	115.0	31.1	0.205	41.0	test positi...
27	7.0	147.0	76.0	0.0	0.0	39.4	0.257	43.0	test positi...
28	1.0	97.0	66.0	15.0	140.0	23.2	0.487	22.0	test negati...
29	13.0	145.0	82.0	19.0	110.0	22.2	0.245	57.0	test negati...

Figure 2 DATASET

4. RESULTS AND ANALYSIS

Table 1 shows the accuracy of machine learning algorithms (NB, PRF) on the diabetes dataset with respect to different test methods (FCV, PS, UTD) with pre-processing method (WPP) and without pre-processing method (WOPP). Figure 2 shows the accuracy of ML algorithms (NB and RF) on the diabetes dataset with respect to different test methods (FCV, PS, UTD) with pre-processing method (WPP) and without pre-processing method (WOPP).

From Table 1 and Figure 3, it is observed that NB machine learning algorithm, For PS test method gives better accuracy compared to other methods without pre-processing method. However, the pre-processing method increases the accuracy for the NB machine learning algorithm. For RF machine learning algorithm, UTD test method gives better accuracy compared to other methods in without pre-processing method. Moreover, the accuracy of RF machine learning algorithm increase with pre-processing method except FCV test method. From Figure 3, it is observed that the pre-processing technique produces better average accuracy for NB compared to other machine learning algorithm.

Test method	WOPP		WPP	
	NB	RF	NB	RF
FCV	76.30	75.78	77.47	74.73
PS	77.01	78.54	79.69	80.07
UTD	76.30	100.00	77.60	100.00
Average	76.53	84.77	78.25	84.93

Table 1: Accuracy of ML algorithms (NB,RF) on diabetes dataset with respect to different test methods (FCV, PS, UTD) with pre-processing method (WPP) and without pre-processing method (WOPP).

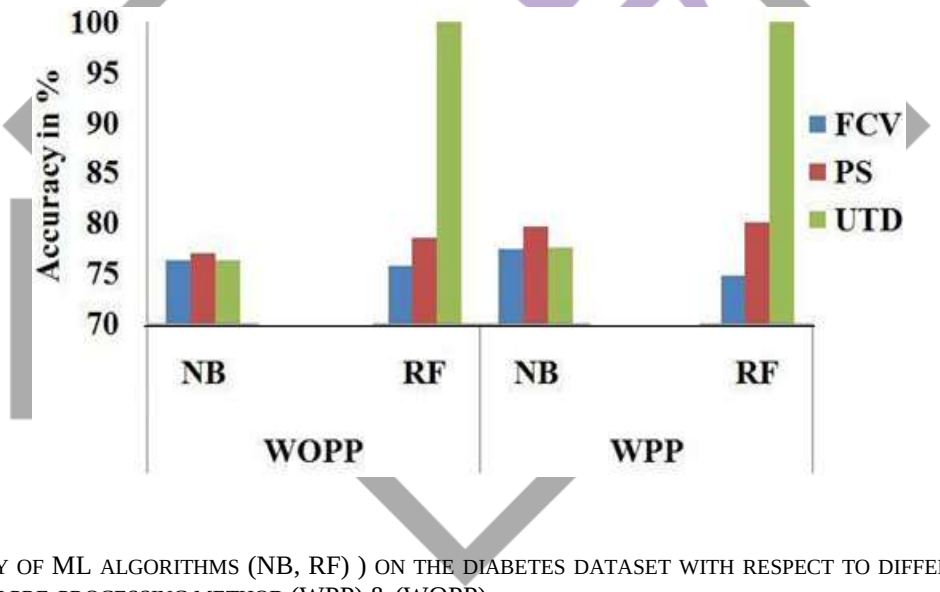


FIGURE 3 ACCURACY OF ML ALGORITHMS (NB, RF)) ON THE DIABETES DATASET WITH RESPECT TO DIFFERENT TEST METHODS (FCV, PS,UTD) WITH PRE-PROCESSING METHOD (WPP) & (WOPP).

5.1 CONCLUSIONS

This paper presented a diabetes prediction system for diabetes treatment. so as to develop this technique , the dataset is collected from kaggle (website). Different machine learning algorithm namely probabilistic-based naïve Bayes (NB), decision tree-based random forests (RF) are used to build the machine learning model to hold out the treatment of diabetes. Furthermore, the machine learning model is tested with different testing methods like 10-fold cross validation (FCV), use percentage split with 66% (PS), and use training dataset (UTD) to measure the performance of the model in terms of accuracy. The pre-processing technique is employed to extend the accuracy of the model. From the results, it's observed that the pre- processing technique increases the accuracy of the machine learning algorithm except two cases. The pre-processing technique gives better average accuracy for NB than other machine learning algorithm..

5.2 FUTURE WORK

Future work should be done on improving the accuracy of the prediction by increasing the extent of coaching data. Its performance are often further improved by identifying and incorporating various other parameters and increasing size of coaching.

REFERENCES

[1] Kaveeshwar, S.A., and Cornwall, J., 2014, "The current state of diabetes mellitus in India". AMJ, 7(1), pp. 45-48.

[2] Dean, L., McEntyre, J., 2004, "The Genetic Landscape of Diabetes [Internet]. Bethesda (MD): National Center for Biotechnology Information (US);. Chapter 1, Introduction to Diabetes. 2004 Jul 7.

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

[4] Chunhui, Z., Chengxia, Y., 2015, "Rapid Model Identification for Online Subcutaneous Glucose Concentration Prediction for New Subjects with Type I Diabetes", IEEE Transactions on Biomedical Engineering, 62 (5), pp. 1333 – 1344

[4] Vaishali, A., Harsh, K., Anil, K.A, 2016, "Performance Analysis of the Competitive Learning Algorithms on Gaussian Data in Automatic Cluster Selection", 2016 Second International Conference on Computational Intelligence & Communication Technology
.

[5] Srinivas, K., Kavihta, R.B., Govrdhan, A., 2010 "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" International Journal on Computer Science and Engineering", 2(2), pp. 250-255

[6] Durairaj, M., Ranjani, V., 2013, "Data Mining Applications In Healthcare Sector: A Study", International Journal of Scientific & Technology Research, 2(10), pp. 31-35.

[7] Salim, D., Suzan Mishol., Daniel, S.K., Dina M., Anael S., 2013, "Overview Applications of Data Mining in Health Care: The Case Study of Arusha Region" International Journal of Computational Engineering Research, 3(8), pp. 73 -77.

[8] Darcy, A. D, Nitesh V.C., Nicholas B, 2008, "Predicting Individual Disease Risk Based on Medical History" CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management, pp. 769-778

[9] Eibe F, Mark A.H, Ian H.W., 2016, "The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

[10] Shanta, Kumar, .Patil, P and Kumaraswamy, Y.S., (2011). "*Predictive data mining for medical diagnosis of heart disease prediction"* IJCSE, 17.

[11] Srinivas, K., (2010). "*Analysis of Coronary Heart Disease and Prediction of Heart Attack in coal mining regions using data mining techniques",* IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349).

[12] Witten, Ian, H., Frank, Eibe. and Mark A., (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd Ed.)

[13] Yoo, Illhoi., Alafaireet, Patricia., Marinov, Miroslav., Pena- Hernandez, Keila, Gopidi, Rejitha., Chang, Jia-Fu and Hua, Lei, (2011). *Data Mining in Healthcare and Biomedicine: A Survey of the Literature,* Med Syst DOI, Springer.