E-ISSN 2581 –7957
P-ISSN 2277 – 3916

CVR Journal of Science and Technology, Volume 16, June 2019
DOI: 10.32377/cvrjst1617

Machine Learning Approaches to Classify Diabetes Patients based on Age, Obesity level and Cholesterol level

Satyanarayana Nimmala
Assoc. Prof, CVR College of Engineering/ CSE Department, Hyderabad, India
Email: satyauce234@gmail.com

Abstract: **Nowadays finding the root cause of some diseases and their effect on different organs of the human body is challenging rather than treatment of the disease. Diabetes stands first in that category. Diabetes is a condition in which the body is incapable of producing insulin or it is not in a situation to make use of the produced insulin, and sometimes both. It is also called as Diabetes Mellitus. In this paper, we experimented machine learning algorithms to find the impact of age obesity level (O), and cholesterol level (C) of a person on diabetes. We have collected real-time data of 50 patients, where 34 are nondiabetic and 16 are diabetic. Each record consists of AOC of a person along with class label attribute. Experimental results unfold that, there is a significant effect of obesity and cholesterol in diabetic patients. Results are compared using rule-based classifier JRip, probability-based classifier Naïve Bayes, and decision tree based classifiers J48, Random forest. All experiments are conducted using 10 fold cross validations by considering the random blood sugar levels of patients.**

Index Terms: **Diabetes, Stress, diet, Obesity, Blood Cholesterol, classification, machine learning.**

I. Introduction

In Diabetes people, the body impairs to get the sugar from the blood into the tissues of the human body, raising the sugar levels in the bloodstream [1]. The excess blood sugar within the blood vessels can harm the blood vessels, this situation leads to different problems like cardiovascular diseases, kidney damages, nerve damage, eye damage and stroke [2]. Most of the time people are victims of diabetes either because of genes or because of lifestyle modifications. But the present day's research reveals that lifestyle changes are one of the leading causes of diabetes prevalence. Beta cells in the pancreas produce insulin to unlock the tissues of the human body to receive sugar in the blood. The received sugar is used for energy or for storage. In diabetic people either insulin is not produced by the pancreas or insulin produced is not able to unlock tissues of the human body to receive sugar in the blood, in such cases sugar levels in the blood raises, and in this situation if a person consumes more carbohydrates from which sugar is generated, then the sugar levels in the blood are at their peak. According to the American Diabetic Association, general symptoms of diabetes include loss of weight, increased thirst, hunger, frequent urination, drowsiness, blurry vision, foot problems, anxiety, and erectile dysfunction in men. According to the World Health Organization (WHO), in the last two decades, diabetes has risen from 108 million to 422 million [7]. There are mainly 3 types of diabetes.

A. Type 1 Diabetes: Beta cells in the pancreas stop production of insulin or produce a small amount of insulin. It may happen because of damaged Beta cells or because of the death of the Beta cells. Most of the times it happens because of the attack of the autoimmune system on Beta cells mistakenly. It is an altered response of the human body to lifestyle changes. In this case, there is no other alternative apart from taking insulin externally, if not patient may die. Type 1 diabetic happens normally in children or in adults.

B. Type 2 Diabetes: This is also called insulin resistance condition. In this pancreas produces insulin but tissues of the human body resist the insulin and they won't get unlocked to receive sugar in the blood. So, pancreas keeps on producing more insulin to unlock tissues, over the period of time as it over functions, Beta cells may get damaged and lead to under insulin. Generally, Type 2 diabetes happens when people age over 40 [3]. The exact cause of Type 2 diabetes is unknown, but genetics, lack of exercise, being overweight, sedentary lifestyle may be the reason.

C. Type 3 Gestational diabetes: It occurs due to insulin-blocking hormones produced during pregnancy. This type of diabetes only occurs during pregnancy. After the pregnancy period, in most cases, the patient may be back to her original health. The differences between these three types are as shown in Table 1.

TABLE 1.
Different Types Of Diabetes

Feature	Type 1 diabetes	Type 2 diabetes	Gestational Diabetes
Age	Children	Adults	During pregnancy time
Prevalence	Rare	More Common	Becoming Common
Treatment	Insulin Injections	Pills, exercise, diet, life style change	Exercise and diet control medications
Cause	Autoimmunity	Insulin resistance	Hormonal Imbalance

CVR Journal of Science and Technology, Volume 16, June 2019
DOI: 10.32377/cvrjst1617

II. Machine Learning and Classification

Machine learning is the concept where a machine or computer is trained to learn a model based on training data. Later on, this model is used by the machine for prediction, decision making, or solving a task. It includes many concepts from statistics, probability, databases, etc. It is broadly divided into 2 types [3][6], one is supervised learning another is unsupervised learning, in supervised learning we study about classification algorithms, like decision tree based, function-based, probability based, rule-based, etc. In unsupervised learning, we study about the association, clustering, anomaly detection and etc. Classification is a supervised learning approach, in which we divide the population into two parts. One part is the training data set by using which a classification model is built. Another is training data set tested against the model being built. We use various performance measures to validate the performance of the classification model, such as accuracy, precision, recall, and f –measure, etc [8].

TABLE II.
Data Set Details

	Minimum	Maximum	Mean	Standard Deviation
Age	18	65	39.1	11.051
Obesity level	15.1	37	24.74	4.42
Cholesterol level	102	258	173.08	35.499

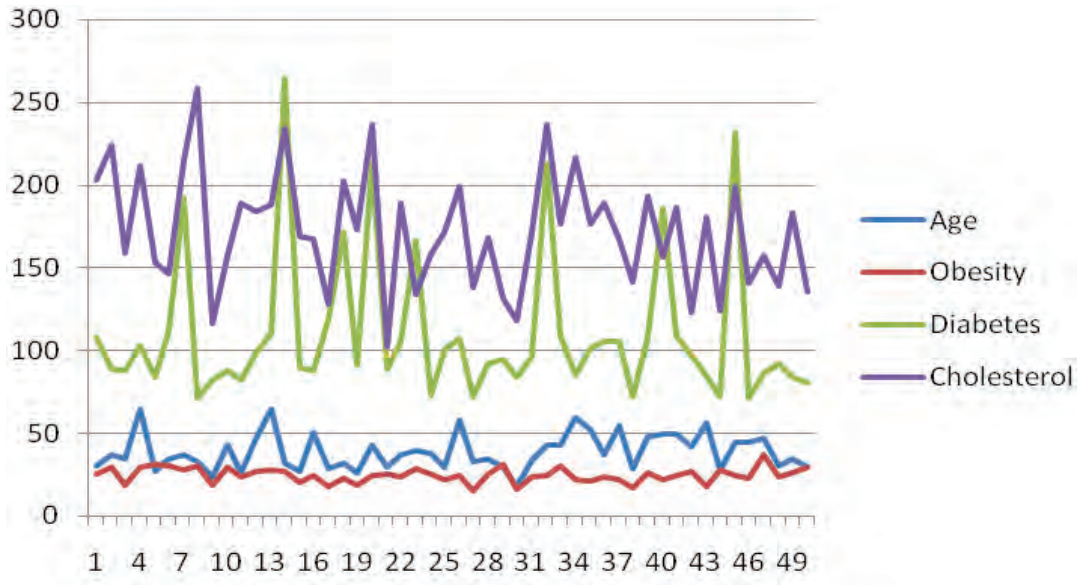

Figure 1. Distribution of Data Set

TABLE III.
Sample Data Set Copy

S.No	Age	Obesity Level	Cholesterol level	Diabetic
1	31	25.1	203	no
2	37	29.2	224	no
3	35	19	159	no
4	65	29.4	211	no
5	27	31.4	200	yes
6	35	30.5	146	no
7	37	28..3	215	yes
8	33	30.2	258	no
9	24	18.7	116	no

CVR Journal of Science and Technology, Volume 16, June 2019
DOI: 10.32377/cvrjst1617

10	43	29.4	157	no
11	26	23.9	189	no
12	47	27.4	184	no
13	65	27.6	188	no
14	32	26.7	234	yes
15	27	20.5	169	no

In the current scenario machine learning techniques are playing a vital role in the bio-medical data processing. Exclusively classification is being used in disease diagnosis and treatment. Classification algorithms are extensively used in biomedical engineering systems to help medical professionals to diagnose the problem with more accuracy. In this paper, we have used different classification algorithms to classify data of diabetic patients. We collected real-time data from a medical diagnostic center. Data set consists of diabetes and nondiabetes people. The details of the data set are as shown in table 2 and Table 3 shows the sample data set copy. Figure 1 shows the distribution of total data set, X-axis represents the total number of records and Y-axis represents the values of selected attributes.

III. Results And Discussions

Experimentation is done using 10 fold cross-validations. In each fold, 90% data is used for training and leftover 10% data is used for testing. For comparative analysis, we conducted experiments using JRip, Naïve Bayes, J48, and Random forest algorithms. **JRip:** It is a rule-based classifier supported by WEKA. It uses two stages to learn optimized rules. The first stage is building stage, rules are generated, second is the pruning stage, rules are pruned. **Naïve Baysed Classifier:** It is a probability-based classifier; it uses conditional probability to build a classification model. The model is built using training records, and then the same model is tested against the test records whose class label has to be predicted. **J48 Classifier:** It is a decision tree based classifier, it builds a decision tree, where each node is a decision node, and each branch is the outcome of the decision taken at the selected node. It uses information gain and entropy to find the best split attribute. **Random Forest Algorithm:** It is also a supervised learning algorithm. It initially builds random forest of trees using a training data set and selects the best tree to classify the test samples. The performance measures of different classifiers considered for experimental analysis is shown in Table 4. Although, Naïve Baysed classifier is good at accuracy, it is only good at predicting negative class tuples, but it is not good at predicting positive class tuples. JRip is reasonably good at predicting positive class as well as negative class tuples compared to all other classifiers. Although, the accuracy of the existing algorithms is not well performed, the experiments are interesting because so far, no research is found in the literature to predict diabetes patients based on their age, obesity and cholesterol levels. However, we are proposing a novel priority based rule-based classifier as part of the future research work to classify diabetes patients with improved accuracy. The relative absolute, the root relative squared errors are more because as we considered only 50 records for experimentation.

TABLE IV.
Accuracy and Error Rates

S. No	Algorithm used	Accuracy	Mean absolute error	Relative absolute error	Root relative squared error
1	JRIP	74%	0.3682	83.4766 %	97.7523 %
2	Naïve Bayesed	76%	0.328	89.9363 %	99.4892 %
3	Random Forest	72%	0.328	80.1247 %	95.067 %
4	J48	74%	0.3536	86.3722 %	101.0526 %

CVR Journal of Science and Technology, Volume 16, June 2019
DOI: 10.32377/cvrjst1617

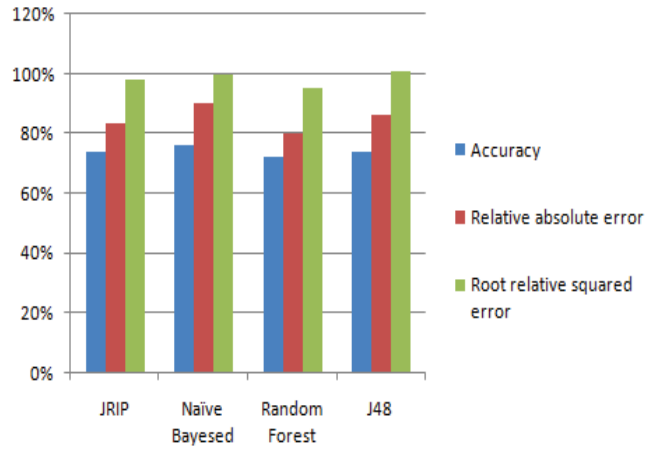

Figure 2. Details of Accuracy and Error Rates

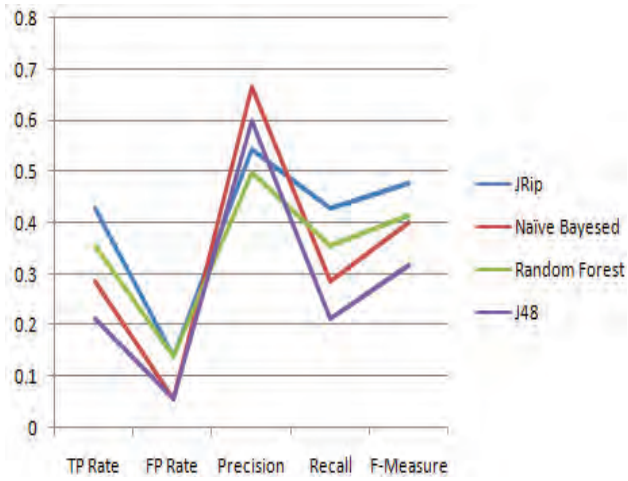

Figure 3. Performance Details of Yes Class

Figure 2 represents the accuracy details of different classifiers, Figure 3 represents the performance of each classifier in predicting diabetic patients, from the figure we can understand that Naïve Bayes is good at Precision and JRip is good at TP rate while predicting yes class tupples. Figure 4 represents the performance of each classifier in predicting nondiabetic patients, J48 is good at TP rate and JRip is good at precision while predicting no class tupples.

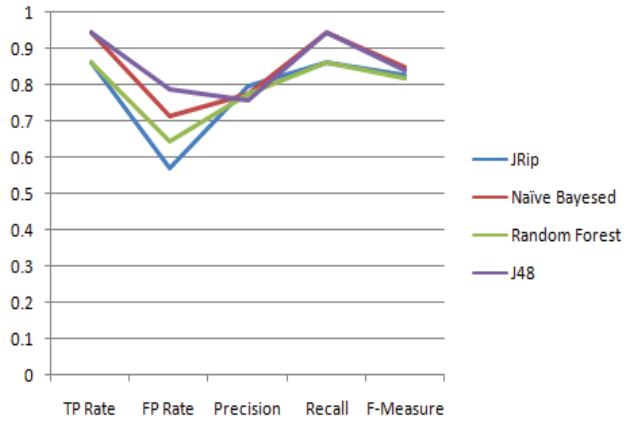

Figure 4. Performance Details of No Class

Here, TP rate represents True positive rate of the classifier means how good classifier is good at predicting positive class tuples, Fp rate represents False Positive rate means a number of tuples actually negative are predicted as positive. F-measure is a harmonic mean of precision and recall. Figures 3, 4 unfold that some classifiers are good at predicting positive class tuples, some are good at predicting negative tuples. Table 5 represents the confusion matrix of each classifier, Table 6 represents the Tp rate, Fp rate Precision, recall, and F – a measure of each classifier with respect to each decision class.

TABLE V.
CONFUSION MATRIX

Algorithm used		Predicted class	
	Actual class	YES	NO
JRip	YES	6	8
	NO	5	31
Naïve Bayesed	YES	4	10
	NO	2	34
Random Forest	YES	5	9
	NO	5	31
J48	YES	3	11
	NO	2	34

TABLE VI..
PERFORMANCE DETAILS OF DIFFERENT CLASSIFIERS WITH RESPECT TO YES CLASS AND NO CLASS

Algorithm used	Class	TP Rate	FP Rate	Precision	Recall	F-Measure
JRip	YES	0.429	0.139	0.545	0.429	0.480
	NO	0.861	0.571	0.795	0.861	0.827
Naïve Bayesed	YES	0.286	0.056	0.667	0.286	0.400
	NO	0.944	0.714	0.773	0.944	0.850
Random Forest	YES	0.357	0.139	0.500	0.357	0.417
	NO	0.861	0.643	0.775	0.861	0.816
J48	YES	0.214	0.056	0.600	0.214	0.316
	NO	0.944	0.786	0.756	0.944	0.840

IV. CONCLUSIONS

In this paper, we used age, obesity and cholesterol levels of a person to classify whether a person is Diabetic or not. Experiments revealed that, there is a significant level of impact of obesity and cholesterol levels in diabetic patients. Although the accuracy of Naive Bayes is more in classifying the records, but JRip is outperformed well in classifying the positive class records. The experimental analysis concludes that the patients are prone to diabetes if their cholesterol level is more than 200 and obesity is more than 30. Results also reveal that the people age over 45 are also more likely prone to diabetes. In the future, more parameters may be considered like family history, diet, smoking, drinking, and stress levels in classifying diabetic patients. A new model (priority-based rule-based classifier) may be built to improve the overall accuracy of the prediction system.

REFERENCES

[1] S.A Kaveeswar, and J Cornival J, "The current state of diabetes milletus in India", AMJ, 7(1), pp.45-48, 2014.
[2] M Durairaj, V Ranjani, "Data Mining Applications In Healthcare Sector: A Study", International Journal of Scientific & Technology Research, 2(10), pp. 31-35, 2013.
[3] D.A Kumar, R Govindasamy,"Performance and Evaluation of Classification Data Mining Techniques in Diabetes", International Journal of Computer Science and Information Technologies, vol 6, pp.1312–1319.
[4] S Perveen, M Shahbaz, A Guergachi, K Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Procedia Computer Science 82,115–121, 2016.
[5] P.S Kumar, V Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and ResearchPublications7,705–709, 2017.
[6] A Tarik, S.M.A Rashid, R.M Abdullah,Abstract, "An Intelligent Approach for Diabetes Classification ,Prediction and Description,Advances in Intelligent Systems and Computing 424,323–335, 2016.
[7] WHO: https://www.who.int/mediacentre/factsheets/fs138/en/, https://www.who.int/diabetes/en/.
[8] V.V ijayan, C Anjali, "Prediction and diagnosis of diabetes mellitus A machine learning approach.2015 IEEE Recent Advances in Intelligent Computational Systems(RAICS),122–127,2015.