International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 07, July-2019

Data Analysis and Prediction of Diabetic Disease by Supporting Data Mining

Getaneh Ashenafi, Parma Nand, Eshetu Tesfaye

[1]M. Tech, [2]Professor, [3]M.Tech
Computer Science and Engineering, School of Engineering and Technology
Sharda University, Greater Noida, India

Abstract*— Diabetes is the world's prevalent and fast-growing illnesses. For all nations, it is the greatest health problem. Diabetes is regarded one of the deadliest and most chronic diseases causing blood sugar to rise. Diabetes is regarded one of the deadliest and most chronic diseases causing blood sugar to rise. If diabetes remains untreated and unidentified, there are many complications. The tedious process is defining outcomes in a patient's visit to a diagnostic centre and doctor's advice. The increase in approaches to machine learning, however, solves this critical issue. This study's motivation is to develop a model that can predict the probability of peak precision in patients with diabetes. For this research, we use data mining tools that WEKA. WEKA is machine-learning software. Weka has its working procedure. Machine learning classification algorithms; are used in this experiment to identify diabetes at an early point, namely Decision Tree, logistic function and Naive Bays. All three algorithms ' performances were assessed on multiple measures such as Precision, Accuracy, F-Measure, and Recall. Accuracy measured over cases classified properly and wrongly. Results achieved shows outperform of logistical feature with the greatest precision of 78.26 percent compared to other algorithms. These findings verified correctly and systematically using Receiver Operating Characteristic (ROC) curves. At the end the predication of each are attribute are registered.

Keywords: Diabetics disease prediction, Data mining, Classification, Logistic function, Naïve Bays

I. INTRODUCTION

In this research, WEKA 3.6.13 is used. WEKA is the Waikato Environment for Knowledge Analysis and is created and freely circulated by Waikato University, New Zealand. WEKA is one of the most known for data processing and data analysis tools. Therefore, since WEKA software written in Java language, it operates on nearly every platform. It comprises of a variety of machine learning algorithms and can solve a multitude of issues related to data mining and machine leaning. WEKA supports numerous functions in machine learning and information mining such as regression, classification, prediction, choice of features and visualization. WEKA offers and manipulates a database link for accessing information. WEKA enables us to more appropriately generate, operate, alter and evaluate experiments. WEKA's most prominent benefits include its free availability, portability, a comprehensive collection of pre-processing and modelling methods, and the user-friendly graphical user interface make it simple to use. WEKA output is relatively better than TANAGRA and MATLAB other data mining instruments. Different methods of classification demonstrate far better outcomes on WEKA than other instruments [8] when comparing WEKA with other data mining instruments, the investigator selects WEKA from TANAGRA and MATLAB from other data mining instruments as they had better analyzed the outcome of WEKA output or data mining tools.

A. Dataset Selection

When it comes to data mining and machine learning, information choice is a method that selects the most appropriate information from a particular domain to derive informative values and to promote learning within that domain. In the research, they used diabetes dataset with eight characteristics used to predict a female patient's symptom of gestational diabetes. This dataset is a benchmark dataset from the UCI repository. Based on historical data stored in the dataset such as age, body mass index, blood pressure and number of pregnant classifiers taught to decide whether an individual's diabetes test is positive or negative. The PIMA Diabetes Dataset reflects only Indian National Women who are at least 21 years of age. All characteristics are numerically valued ongoing data type. The class label attribute is the dichotomous variable that follows each tuple of the dataset within the PIMA dataset (i.e. the binary response variable). UCI repository PIMA Indian Diabetes Dataset includes 768 instances. The PIMA dataset transformed from CSV to the format ".ARFF" adopted by WEKA 3.6.13. The full details of all the eight characteristics presented in the following table. [9].

Table 1.The PIMA Dataset Description.

S N	Attribute	Type
1	Number of times pregnant	Numeric
2	Plasma glucose concentration	Numeric
3	Blood pressure (Diastolic)	Numeric
4	Triceps skin fold thickness (mm)	Numeric
5	2-Hourseruminsulin	Numeric
6	Body mass in dex(kg/m2)	Numeric
7	Diabetes pedigree function	Numeric
8	Age (years)	Numeric
9	Class Variable (True or False)	Nominal

According to WEKA Ml instruments, the scientists organize and pick the information. The row information must transformed from CSV to the ".ARFF" format in order to use WEKA system software. Using updated fresh 3.9 WEKA ML tools, the investigator will do their job better. They use WEKA 3.6's ancient template. When WEKA updates itself, it adds extra information to its instruments for testing.

B. Analysing Methods of Select Classification Algorithm between two publisher

It is possible to do data analysis and predication diabetic disease by using different research method. For example; using data mining algorithm, by using Weka open source software for machine learning tools, python, etc. From that methodology, we do our paper by using Weka open source software for machine learning and data mining tools. Because this software is easy, it is machine-learning algorithm and it processed automatically.

C. Decision Tree J48 Algorithm and Naïve Bays Algorithm:

They make different method to evaluate their testing method Algorithms. From those testing method they select the method, which has high Value of accuracy and low value of error comparison. By using that method, they select Naïve Bayes Algorithm for testing method. Both the models are efficient in the diagnosis of diabetes using the percentage split of 70:30 of the data set. A developed model for diagnosis of diabetes will require more training data for creation and testing. At end there result, they agree both the models are efficient in the diagnosis of diabetes using the percentage split of 70:30 of the data set. A developed model for diagnosis of diabetes will require more training data for creation and testing. [10]

D. J48 Decision tree, Naïve Bays and SMO Support Vector Machine

At the next step, they Evaluate results of three classifier models and select higher accurate results. From those higher accurate results given by the J48 Decision Tree and SVM Support, vector. Although all the methods have given more than 75% accuracy, the Decision Tree and the SMO Support Vector Machine give more accurate results than the Naïve Bayes algorithm. However, the ensemble method gives the highest accuracy from all due to the voting process of all the algorithms.

This publisher also have planned to gather more data from different locales over the country and develop more precise and general prescient model because increasing the data set also cause to increase the accuracy of the results. [11]. Both researcher select testing method algorithm and on the second researcher paper deeply annualize about their testing method algorithm by different analogizing methods.

The rest of this paper work arranged as follow: Section 2 describes literature review. Section 3 explains the proposed methodology. In Section 4, experiment analysis and results explained and lastly summarizations of this paper and are recommended in section 5.

II. LITERATURE REVIEW

Diabetes is a situation where your body is unable to generate the quantity of insulin needed to control our body's sugar content [1]. There are two mean reasons for diabetes in particular: the first is when the pancreas does not generate enough insulin or the body does not generate enough insulin and the second is when cells do not react to the insulin generated. Insulin is the hormone principle that controls blood glucose acceptance in most cells. Glucose is a nice Greek word. Glucose is a sort of sugar that you consume from food, and your body to get energy uses it. Blood glucose Insulin is a hormone that carries sugar from our blood into the cells for energy and storage as it travels through your bloodstream to your cells we call. People with diabetes have higher than normal blood glucose concentrations. Either they do not have enough insulin to move through it or their cells do not react as well as they should to insulin. Long-term high blood glucose may harm your kidneys, eyes, and other organs. Nearly 70% of the world's fatalities in which type II diabetes mellitus is most prevalent in all [2]. It proposes that in using various data mining techniques, a hybrid diagnostic model could predict type 2 diabetes. It also enables patients to undergo multiple tests such as blood tests, diastolic and systolic blood pressure checks etc. by using this model to make choices and enhance diagnostic accuracy. It is sufficient to diagnose whether or not they are suffering from Diabetes Mellitus after getting this system patient itself. These estimates

are based on the symptoms that occur in diabetes mellitus early phases. One of the significant contributors to the mortality rate is diabetes mellitus. It needs to be detect and diagnosed with diabetes disease on the list one day. A major classification problem is the diagnosis and interpretation of diabetes disease data [3]. A classifier needed and cost-effective, convenient and precise design is need. Artificial intelligence and soft computing techniques provide many human ideologies and are engaged in areas of implementation linked to human beings. In the medical diagnosis, these systems discover a location. A diagnosis of medicine is a method of classification. Before diagnosing diabetes, a doctor has to analyse many variables, which makes the work of a doctor hard. The value-based hospital therapy and world digitization prefers computerized information rather than hard copy form. Health care data include Patient Data Electronic Health Reports, doctor prescription, clinical reports, diagnostic records, medical pictures, pharmacy information, data associated to health insurance, social media data and medical journals. Machine learning and data mining methods recently had regarded in the design of an automatic diabetes diagnostic scheme [4].

Data mining has played a significant role in diabetic studies from these methods. For diabetic scientists, data mining would be a precious asset because it can uncover hidden knowledge from an enormous quantity of data-related diabetics. Data mining has played a significant role in diabetic studies from these methods. For diabetic scientists, data mining would be a precious asset because it can uncover hidden knowledge from a enormous quantity of data-related diabetics. Data mining is the method that used enormous amounts of information to find unknown values. As the population of patients increases the number of medical databases, it also rises daily. [5] Without a computer-based analysis scheme, communicating and investigating these medical information is hard. The computer-based scheme of assessment shows the automatic system of medical diagnosis. For example, data mining is capable of extracting hidden knowledge from multifaceted information repositories; medical records, reports, flow charts and tables of proof, etc. These repositories where transformed into helpful decision-making data. [6] This mechanized diagnostic system assists the medical specialist in making excellent treatment and disease decisions. Data mining is the huge area for physicians to handle the enormous amount of patient data sets in many ways, such as making sense of complex diagnostic testing, interpreting earlier results, and combining the dissimilar data. Traditionally nursing choice formed by the observations and foreknowledge of the medical practitioner rather than the understanding obtained from the enormous quantity of information. This mechanized diagnostic scheme improves the quality of service supplied to patients and reduces the cost of medicine. Pre-processing information is one of the techniques of information mining, which includes transforming raw information into a comprehensible format. Real-world information is often incomplete, inconsistent and/or missing in certain behaviours or trends and is probable to contain many errors [7]. Pre-processing data is a proven way to solve these problems. Pre-processing of information prepares raw information for further processing. In other words, the information that we want to analyse using data mining methods is loud, incomplete and inconsistent, which is why we need to do data washing, data integration, data transformation and data reduction. Predicated implementation allows branch removal when transformed into straight-line sections of conditional activities in branching code sections. A significant side effect of this conversion, which is generally unnoticed, is that the compiler must allocate separate resources at a specified moment to all predicted activities to guarantee that these resources are accessible at runtime.

III. PROPOSED METHODOLOGY

The current work expects to create a mining model based on two classification algorithms in order to provide a simpler solution to the problem of diagnosis of diabetes disease in women. Those classification algorithms are decision tree and logistic function. The results have been analyzed using statistical methods. Mainly our working methodology supported with weka tools. Those are Data Pre-processing and data classifier. At the first time when we start our work, we change our row data file format to CSV and arff file format. Then we have selected logistic regression after comparison of each model with each other, which are as explained below.

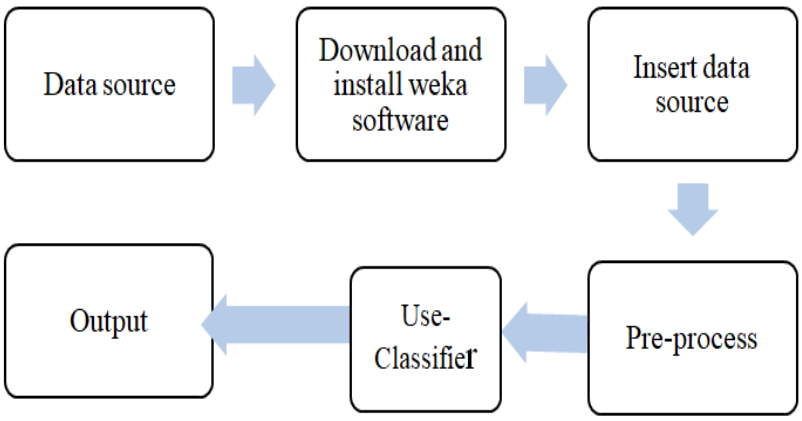


Fig. 1 The frame of design and development data prediction

IV. EXPERIMENT RESULTS AND ANALYSIS

A. Pre-processing

The step of pre-processing by WEKA explorer of different result showed as below:

- ✓ Open Weka GUI Chooser.
- ✓ Click the "Explorer" button to open the Weka Explorer.
- ✓ Click the "Open file…button, navigate to the *data/* directory
- ✓ Select the tested data; Click on Open button".
- ✓ After this process, we got the result on figure 3.

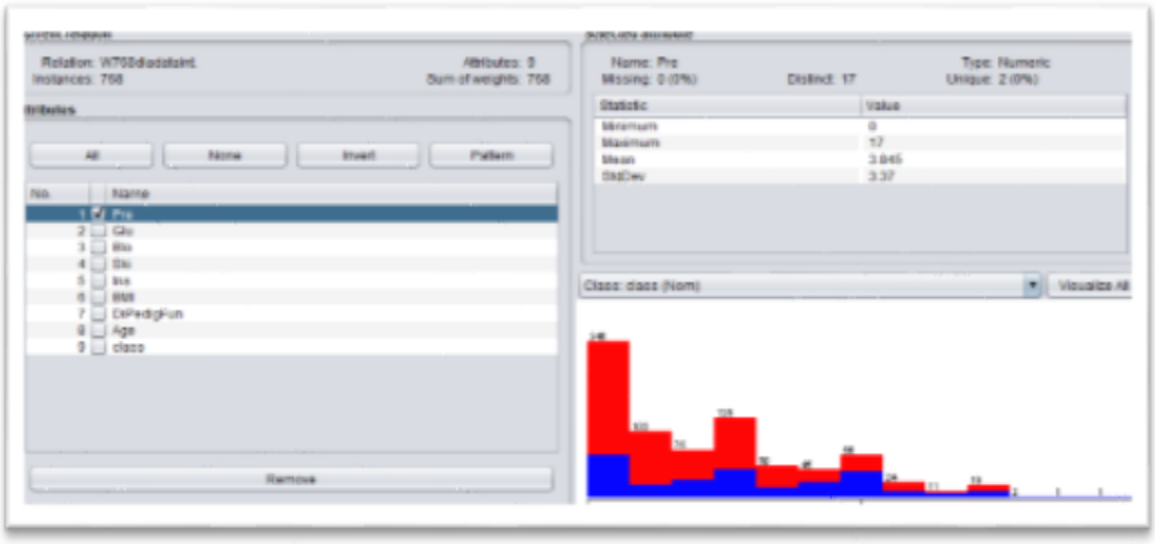


Figure 2: At the begging of Pre-processing Result of WEKA

rom dataset, the data loaded into the WEKA about the data of 768 diabetic disease instance and eight attribute, checked the first attribute, and got full result done by weka software automatically indicated on Figure 2 shown as below. At the same as the first attribute registered on figure two, to check each attribute then, we listed the Pre-processing Statistical Evaluation of each attribute on the table 2 below.

Table 2: Statistic value of each selected attribute of input

Name	Statistic value			
	Minimum	Maximum	Mean	Std Dav
Pregnancies	0	17	3.845	3.37
Glucose	0	199	120.895	31.973
Blood Pressure	0	122	69.105	19.356
Skin Thickness	0	99	20.356	15.952
Insulin	0	846	79.799	115.244
BMI	0	67.1	31.993	7.884
BIP edigFun	0.078	2.42	0.472	0.331
Age	21	81	33.241	11.76

Table 3: The result of each selected attribute at first pre-processing time.

Name	Missing	Distinct	Unique	Type
Pregnancies	0(0%)	17	2(0%)	Numeric
Glucose	0(0%)	136	19(2%)	Numeric
Blood Pressure	0(0%)	47	8(1%)	Numeric
Skin Thickness	0(0%)	51	5(1%)	Numeric
Insulin	0(0%)	186	93(12%)	Numeric
BMI	0(0%)	248	76(10%)	Numeric
BIP edigFun	0(0%)	517	346(45%)	Numeric
Age	0(0%)	52	5(1%)	Numeric
Class/out put	0(0%)	2	0(0%)	nominal

Table 4: ***Selected attribute of output at pre-processing time.***

No.	Label	Count	Weight
1	tested –positive	268	268.0
2	tested-negative	500	500.0

The above result automatically processed by WEKA Ml system software.

A. Classifier

At Classification, training set used to learn a model that can classify the data samples into known classes. The Classification process involves by the steps of Create training data set, Identify class attribute and classes, Identify useful attributes for classification (Relevance analysis), learn a model using training examples in Training set and at the end Use the model to classify the unknown data samples. On this paper when the analysis start and predicate the data set by using two way of training and test sets.

Training sets

The training set used to build the model Classification process consists of training set that are analyzed by a classification algorithms and the classifier or learner model is represented in the form of classification rules. For treating data set used different steps and "arff" format file. To provide test data to the training data. Then indicate how the training set in weka automatically by using by research dataset (diabetic dataset). From weka explorer; classifier tools differentiate the type of technique, which used for diabetic disease data research. For training data set test, select the techniques Algorithms Logistic and Diction tree, J48 tools from WEKA explorer classifier.

B. Decision trees (J48)

To use the technique on WEKA system software:

- ✓ Open Weka GUI Chooser.
- ✓ Click the Explorer button to open the Weka Explorer.
- ✓ Click the "Open file…button, navigate to the *data/* directory
- ✓ Select the tested dataset
- ✓ Click the "classify" button to open the Weka classier.
- ✓ Select cross-validation folds and change its folds 10 to 12
- ✓ Select percentage split and change its percentage split 66 to 70
- ✓ Select the type of test option from the indicated test options

Table 5: Performance Results from J48 Classification Algorithm

	No. Of Instances	By percentage
Correctly classified Instances	646	84.1146
Incorrectly classified Instances	122	15.8854

From the above table the accuracy of model performed by weka automatically, that is 84.1146 %. It used to evaluate with other technique of weka tools. Similarly, incorrectly classified instances means the sum of FP and FN. The total number of correctly instances divided by total number of instances gives the accuracy. In weka, percentage of correctly classified instances gives the accuracy of the model.

Table 6: The additional results from J48 Classification Algorithm

Kappa statistic	0
Mean absolute error	0.4545
Root mean squared error	0.4766
Relative absolute error	100 %
Root relative squared error%	100
Total Number of Instances	768

Table 7: The additional results from J48 Confusion Matrix

	a = tested _positive	b = tested _negative
Actual true	178 TP	32 FN
Actual false	90 FP	468 TN

TP- Positive tuples that correctly labelled by the classifier

TN-True Negative tuples that correctly labelled by the classifier.

FP- False Positive tuples that incorrectly labelled as positive.

FN- False Negative tuples that were mislabelled as negative.

The formula, which used to accuracy manually.

Accuracy = (TP+TN)/ (TP+FP+TN+FN)

C. The other selected WEKA technique is logistic.

The step to do it is the same as j48 from classifier choose function-logistic

Table 8. Performance Results from logistic Classification Algorithm

	No. of Instances	By percentage
Correctly Classified Instances	601	78.2552 %
Incorrectly Classified Instances	167	21.7448 %

At the above table we got the accuracy of model which done by weka automatically that is 78.2552 % It used to evaluate with other technique of weka tools. Similarly, incorrectly classified instances means the sum of FP and FN. The total number of correctly instances divided by total number of instances gives the accuracy. In weka, % of correctly classified instances give the accuracy of the model.

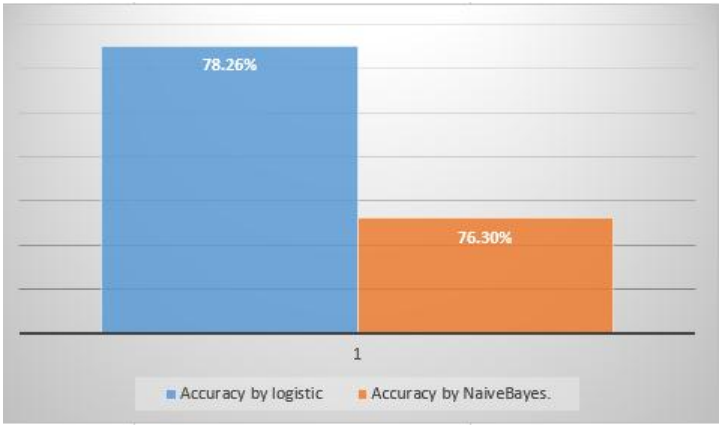
Table 9. The additional results logistic Classification Algorithm

	a = tested _positive	b = tested _negative
a = tested _positive	156 (1)	55 (2)
b = tested _negative	112 (3)	445 (4)

Where the numbers indicate:

1. Number of correct forecasts that the instance tested positive
2. Number of incorrect forecasts that the instance tested negative
3. Number of incorrect forecasts that the instance tested positive
4. Number of correct forecasts that the instance tested negative

The Comparison accuracy between logistic function and Bayes Naive Bayes, the accuracy logistic function is better than the accuracy Bayes Naive Bayes. That is from logistic function we got 78.26% and from Naive Bayes we got 76.30%



and graphically shown from the above.

V. CONCLUSION

This research study understands how to use "WEKA" information-mining instruments in short time to obtain a outcome prediction of enormous amounts of any dataset. The research demonstrates the use of diabetic disease dataset to analyse and obtain the outcome of future prediction.

Various instruments for data mining and ML support this paper. As an instance, it shows information analysis and prediction. This document uses various ML technology to create predication easy. When comparing the updated technology to the old / manual/, the updated technology used to consume time, avoid errors, etc. This study's key goal is to improve predictive model precision. Precision increased by enhancing information efficiency, algorithms, or even by tuning algorithms. This study improve the precision by enhancing the information that works really well in the pre-processing stage. Applying the bootstrapping resampling method to this dataset will increase the precision of nearly all classifiers, but logistic function are leading above others. It found a model's precision depends heavily on the dataset.

REFERENCE

1. Muni kumar N, Manjula R,"Role of Big Data Analytics in Rural Health Care – A Step Towards Svasth Bharath", International Journal of Computer Science and Information Technologies, vol 5(6), pp 7172-7178, 2014
2. De Silva, L. H. S., Pathirage, N., & Jinasena, T. M. K. K. (2016). Diabetic Prediction System Using Data Mining.
3. Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kauffman Publishers, 2001
4. Barakat, et al. "Intelligible Support Vector Machines for diagnosis of Diabetes Mellitus." IEEE Transactions on Information Technology in Biomedicine, 2009.
5. T.Mitchell, Machine Learning, McGrawHill, New York, 1997.
6. H.S De Silva1#, Nandana Pathirage2 and T.M.K.K Jinasena3 Proceedings in Computing, 9th International Research Conference-KDU, Sri Lanka 2016.
7. Kaur, H., & Wasan, S. K. (2006). Empirical study on applications of data mining techniques in healthcare. Journal of Computer science, 2(2), 194-200.
8. Zia, Uswa Ali, and N. Khan. "May, Predicting Diabetes in Medical Datasets Using Machine Learning Techniques." *International Journal of Scientific and Engineering Research*8.5 (2017).
9. Zia, Uswa Ali, and N. Khan. "May, Predicting Diabetes in Medical Datasets Using Machine Learning Techniques." *International Journal of Scientific and Engineering Research*8.5 (2017).
10. Iyer, Aiswarya, S. Jeyalatha, and Ronak Sumbaly. "Diagnosis of diabetes using classification mining techniques." *arXiv preprint arXiv:1502.03774* (2015)
11. De Silva, L. H. S., Pathirage, N., & Jinasena, T. M. K. K. (2016). Diabetic Prediction System Using Data Mining.
12. Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013