Accuracy of Classification Algorithms for Diabetes Prediction

Ambika Rani Subhash, Ashwin kumar UM

Abstract— The illness that happens in the human body because of enormous amounts of sugar in the blood, i.e., when the human body has elevated amounts of glucose in the blood is Diabetes Mellitus all the more ordinarily referred to just as diabetes. The diverse most usually happening assortments of diabetes are Prediabetes, Type2, Type 1 and Gestational Diabetes. Type 2 diabetes, is interminable and generally happens, when the human body does not usefully utilize the hormone, insulin, which is created by it. The Type 1 assortment happens when the pancreatic organ doesn't deliver enough insulin as is required by the human body.Prediabetes is one that occurs when the blood sugar levels are very high but not as much when compared to the Type 2 variety. Gestational diabetes usually affects pregnant women and here also the blood sugar levels are very high. According to the global report by the WHO (World Health Organization), around 422 million people suffer from the disease and a worrying 1.6 million odd deaths are credited only to diabetes every year. However, timely diagnosis of the disease and care of patients through simple lifestyle measures has proven to keep this deadly disease in check. The main challenge for doctors however, is the tedious process of identifying the factors that cause the occurrence of this disease, in an effective and timely manner. During the recent times this challenge is being addressed through Data Mining and Machine Learning techniques. The main aim of this experimentation is for designing a prediction model which can, with utmost accuracy, diagnose the occurrence of diabetes in patient. These training models have been designed using the WEKA tool and four supervised machine learning classification algorithms such as Naïve Bayes, J48, SVM and Neural Networks have been used to predict the onset of diabetes at an early stage. The dataset used here is the Pima Indian Diabetes training Dataset abbreviated as PIDD, which has been acquired from the UCI repository. Chi-squared tests have been applied on this dataset to obtain only those attributes that have the highest tendency of causing diabetes in patients. The performance of each of the classification algorithms have been compared and analyzed based on Accuracy, F-measure, Recall, Precision and ROC curves.

Keywords— Naïve Bayes, J48, SVM, Neural Networks/Multilayer Perception, Diabetes, Chi-squared test, WEKA, Accuracy

I. INTRODUCTION

Diabetes, a disease, that occurs in the human body when the levels of glucose in the blood i,e., the sugar levels are abnormally high. The glucose which is a main element in carbohydrates is present in the blood stream and is the main source of energy for our body and comes from the food that we eat. The human body, through the pancreas, produces a hormone called insulin, which extracts the glucose form the food and allows the beta cells to absorb the glucose and convert into energy. Unfortunately sometimes, the human body never produces required amounts of insulin or sometimes does not produce any insulin at all.

When this happens, overtime the body accumulates this glucose leading to excessive glucose in the blood stream, which in turn leads to health problems such as diabetes. The most commonly occurring types of the disease are: Type 2 – which occurs when the body either doesn't produce any or does not effectively make proper use of the insulin it produces. Type 1 – when the human body is not capable of making any insulin, and the cells in the pancreas that do make the insulin, are attacked and destroyed by the immune system. Gestational diabetes – this usually occurs in pregnant women, which sometimes might become normal post the delivery of the baby. However, persistence of this type of diabetes post pregnancy results in Type 2 diabetes in the patients. Diabetes is a major cause of health risks such as heart diseases, kidney diseases, blindness, strokes, nerve damages etc.

Nonetheless, even though diabetes is a life threatening disease, it can be brought under control by early detection, correct diagnosis along with proper medical care and simple measures taken for improved life style changes.

Recent trends in the medical field has seen the use of data mining techniques to analyze medical datasets to help doctors with early diagnosis of diseases to help save human life. Since diabetes affects a large population across the globe, it is a hard disease to diagnose and for centuries doctors have been basing their diagnosis on medical datasets. The collection of these datasets is a continuous process and it comprises of various patient related attributes such as age, gender, symptoms, insulin levels, blood pressure, blood glucose levels, weight etc. Since the available information is very vast, data mining techniques are used to extract and use only the most useful information that would help with correct diagnosis.

Machine learning algorithms such as Decision Trees, J48, Naïve Bayes, Logistic Regression, Support Vector Machines (SVM), Neural Networks /Multilayer Perception, Ensemble methods etc., are used on the mined datasets pertaining to the disease to help with accurate diagnosis.

Since data mining techniques coupled with machine learning algorithms help make sense of large amounts of data by putting them into predefined

Revised Manuscript Received on April 25, 2019.
Ambika Rani Subhash, School of C&IT, REVA University, India.
Dr. Ashwinkumar UM, School of C&IT, REVA University, India.

groups, these methods are gaining immense popularity in the medical field.

This work focuses on the training Dataset of the Pima Indians diagnosed with Diabetes (PIDD) extracted from the University of California, Irvine (UCI) machine learning repository. Instead of considering all the attributes pertaining to diabetes given in the dataset, chi-squared test methodology has been used to obtain only those attributes with the highest likelihood of predicting the occurrence of the disease.

Supervised machine learning classification algorithms namely, J48, Neural Networks/Multilayer Perception, SVM and Naïve Bayes are then used only on the attribute set obtained after implementing the chi-squared test and results obtained for each of the algorithms are compared based on the accuracy measures.

The remaining of this paper is organised as follows; Section II which discusses other related works, Section III dealing with the working methodology and the dataset considered, and finally followed by Section IV which discusses the final results obtained. Section V deliberates future works and conclusion to this paper.

II. RELATED WORK

This section of the paper is dedicated to some of the research work done in the field of medical diagnosis using machine learning and data mining techniques.

P. Chen et al[7] in their work have performed statistical testing on medical measurement index results of both patients with diabetes and without diabetes. They have further used boosting algorithms to give excellent classification of diabetes model based on the given medical data.

Sisodia et al.[1] in their study have performed machine learning experiments to find the prediction of diabetes in women patients by comparing the accuracy measures of classifier algorithms.

Nguyen et al.[3] in their study compared the various accuracy measures and also used the ensemble approach by incorporating all the algorithms into a weighted average or soft voting ensemble model where each algorithm counted towards a majority vote towards the decision outcome of whether a patient has diabetes or not.

Zou et al[8] in their paper have performed machine learning experiments. For best feature selection they have used principal component analysis and extracted principal component factors affecting patients.

Gangil et al.[4] in their research performs the selection of optimal features, by the analysis of features in the diabetic dataset based on their correlation values. Based on the best attribute selection the classification techniques are improved thus providing superior classification for diabetes prediction.

Pethalakshmi et al[5]. in their paper also discuss and compare two classifier algorithms namely Naïve Bayes and J48 and study how each of their accuracy rate, time and error rate differs to give accurate predictions of diabetes disease prediction.

Velmurugan et al[6] have also performed a similar research and have also included clustering algorithms in addition to the classifiers to perform experiments and analyse the diabetes datasets, to find highly affected patients.

III. WORKING METHODOLOGY

This paper is mainly aimed at identifying classification algorithms which will most accurately predict the presence of diabetes in patients for early diagnosis and treatment. The following fig.1, shows the model diagram for the proposed study.

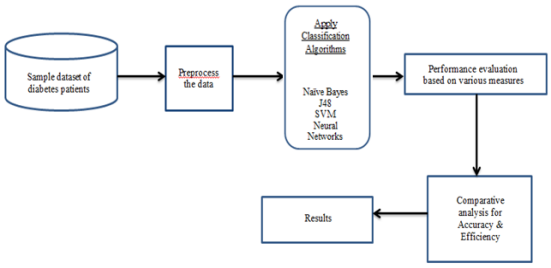

Fig 1: Model Diagram

From the above model diagram we can see that the following steps are followed in this study.

Step 1: The training set used is the PIDD dataset which is a pre-loaded dataset on the WEKA tool. The Pima Indians are a set of Native American people living in Arizona territory with modified genetic tendencies that allow them to survive normally on a diet with very less carbohydrates, for years. But as seen recently, due to a sudden change from the healthier traditional agricultural crops towards pre-processed foods, and also reduced physical activity, they have now evolved to having the highest occurrence of type 2 diabetes and hence, they have been subjected to many a diabetes study. This dataset has been originally obtained from the National Institute of Diabetes and Digestive and Kidney Diseases, and is now an existing training dataset that can be accessed through WEKA. Various limitations were placed while choosing the instances from a much larger dataset. In particular, all featured patients here are females aged between 21 to 81 years and from the Pima Indian heritage. The presently considered database has an instance of 768 women with 8 characteristics as shown in Table 1, along with the abbreviations and type of the attributes,

Attributes	Abbrv	Type
Frequency of pregnency	preg	cont
Glucose levels in the plasma	plas	cont
Bottom reading of blood pressure (mm Hg)	pres	cont
Thickness of triceps skin fold in (mm)	skin	cont
Serum insulin for 2-hours (mu U/ml)	insu	cont
BMI (weight in kg/(height in m)^2)	mass	cont
Pedigree function of diabetes	pedi	cont
Age in years	age	num
The final column of the dataset specifies if the person has been diagnosed with diabetes or not , represented by (1) or (0) respectively	class	disc

Table 1: Attributes of PIDD dataset

Step 2: The WEKA device is utilized for information pre-handling. Information is utilized in .arff design by WEKA. WEKA (Waikato Environment for Knowledge

Analysis) is well-acknowledged open source programming which is an unreservedly accessible suite of AI programming written in Java and created by the University of Waikato, New Zealand. Since datasets utilized includes an excessive number of information passages, information mining is a basic advance towards information examination. As a first step the dataset must be pre-processed as the data obtained might be incomplete, noisy and dirty. The data might lack attributes, have errors and outliers and could be inconsistent. Hence, to solve this problem, data is pre-processed as quality decisions can only be based on quality data. Fig 2, shows the WEKA GUI and the visualization of the pre-processed data.

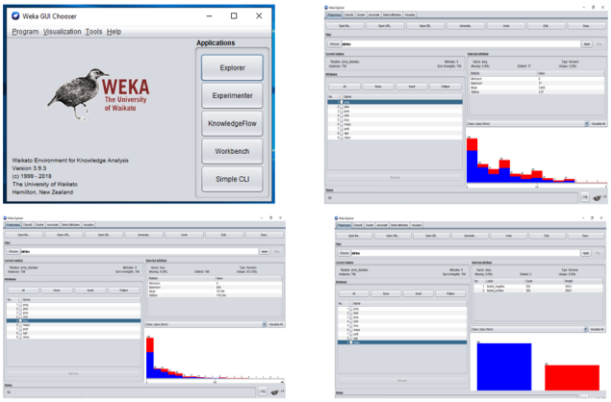

Fig 2: WEKA GUI and sample visualisation of pre-processed data

Step 3: As found in any information mining process, achievement of information mining dependably depends very on information that has been chosen for activity. For any order issue, just those properties with high discriminative forces ought to be chosen. The Chi-square factual test can be utilized to choose the most significant properties that will choose the objective classes in a given characterization issue. Highlight determination is a noteworthy procedure in information characterization after information pre preparing. The primary test is here to expel the undesirable highlights from the dataset that were insignificant to the undertaking performed.

$$\chi^2 = \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency}$$

where –

Observed frequency = No. of observations of class
Expected frequency = No. of expected observations of class if there was no relationship between the feature and the target.

The process of automatically selecting only those features of the given data that contributes the most to the prediction variable or output in which we are interested in, is known as Feature Selection. A Chi-square test is intended to examine unmitigated information, which is information that has been checked and isolated into classifications Attribute selection or selecting of attributes is done to narrow down the potential list to obtain only those attributes that might be most influential in the given prediction model.
Weka's machine learning capabilities are taken advantage of to help find those patterns that are not always obvious in visualizations. Using the 'chisquareattributeeval' attribute evaluator in Weka, we can see the attributes ranked along with merit weightage given to each attribute.
This rank and merit analysis helps us understand about the attributes that are most powerful in helping us predict the occurance of diabetes in the clan of Pima Indian women.

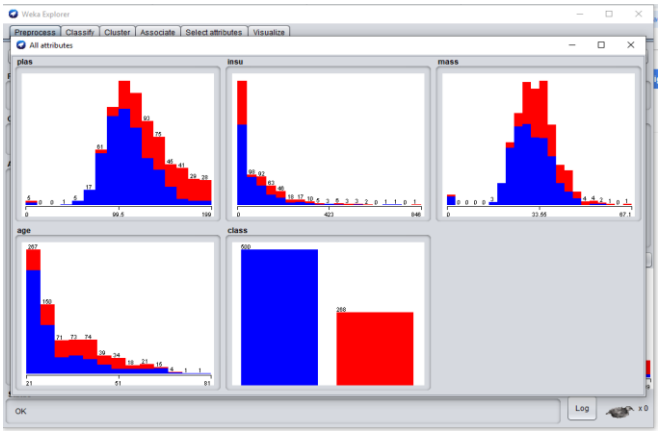

Fig 3. Top ranked attribute selection using ChiSquare Test on WEKA

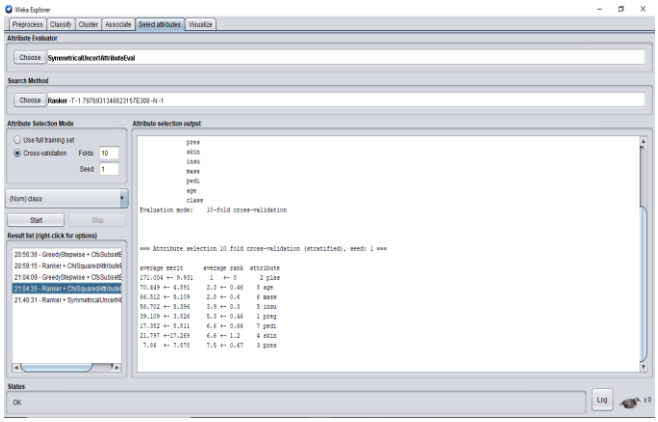

Fig 4. Visualization of top ranked attributes

Thus, on running the chi-squared test on the diabetes dataset we can see that the top ranked attributes that maybe a high contributor towards diagnosing the onset of diabetes in Pima Indian women are,

- Glucose focus in plasma (following 2 hours of oral glucose resistance test) – plas
- Age in years – age
- BMI (weight in kg/(tallness in m)^2) – mass
- Serum insulin in 2 hours (mu U/ml) – insu
- The final column of the dataset denotes if the person does actually have diabetes or not , which is represented with a (1) or (0) – class

Step 4: This step applies the various supervised classificaion algorithms namely, Naïve Bayes, J48, Neural Networks/Multilayer Perception and Support Vector Machine (SVM), on the training dataset that is obtained after following the first to third steps.
Naive Bayes - This arrangement method depends on the Bayes' hypothesis. This classifier accept that the nearness of a specific component in a class is inconsequential to the

nearness of some other element**.**

J48 - This is an open source Java execution of basic C4.5 choice tree calculation. J48 is a straightforward technique to manufacture a choice tree from the preparation information by beginning at the top, with the entire preparing dataset.

Neural Networks/Multilayer Perception - Multi Layer Perception can be characterized as Neural Network and Artificial insight without capability. Neural systems, have a wonderful capacity to get significance from entangled or uncertain information, and can be utilized to concentrate examples and distinguish patterns that are too mind boggling to be in any way seen by either people or other PC procedures.

SVM – This is an arrangement and relapse expectation device that utilizations AI hypothesis to amplify prescient exactness while consequently maintaining a strategic distance from over-fit to the information**.**

The Table 2, below shows the Confusion Matrix obtained for each of the classification algorithms. This matrix describes the performace of the classification model and allows for the visualization of performace of the algorithm.

Machine Learning Algorithms	Negatively Tested (A)	Positively Tested (B)
Naïve Bayes	435	65
	124	144
J48	413	87
	110	158
Neural Networks	431	69
	110	158
SVM	495	5
	267	1

Table 2. Confusion Matrix for each algorithm

Step 5: This is the final step of comparing and anlysig the accuracy measures and performance on the training dataset by the machine learning algorithms mentioned above. Experiments on all of the above mentioned classification algorithms are carried out by the internal 10-folds cross-validation method.

Accuracy Measures	Denotations	Formulae
Accuracy	Accuracy of the algorithm in predicting instances	A = (True Pos +True Neg) / (Total no of samples)
Precision	Correctness of the classifiers	P = True Pos / (True Pos + False Pos)
Recall	Measures the correctness or sensivity of classifiers	R =True Pos / (True Pos + False Neg)
F-Measure	Weightedd average of Precision and Recall	F=2*(P*R)/(P+R)
ROC	Receiver Operating Characteristic curves which will compare the tests	

Table 3: Table of Accuracy measures

As shown in Table 3, Accuracy/Exactness, Precision, Recall,F-Measure and Receiver Operating Curve(ROC) measures are utilized for the grouping of the outcomes, where, True Positive is meant as TP, True Negative is indicated as TN, False positive is meant as FP and False Negative is signified as FN.

IV. RESULTS AND DISCUSSION

On running experiments on the training set obtained, the training models were compared based on the performace of each of the classifier algorithms. The comparison of the previously mentioned accuracy measures that were finally obtained are as given below in Table 4, and Table 5 depicts the instances classified correctly and incorrectly by each algorithm.

Machine Learning Classifiers	Accuracy %	Precision	Recall	F-Measure
Naïve Bayes	75.39	0.747	0.754	0.746
J48	74.34	0.739	0.743	0.741
Neural Networks	76.69	0.762	0.767	0.762
SVM	64.58	0.481	0.64	0.51

Table 4. Comparison of algorithms on various measures of accuracy

Machine Learning Classifiers	Instances Correctly Classified	Instances Incorrectly Classified
Naïve Bayes	579	189
J48	571	197
Neural Networks	589	179
SVM	496	272

Table 5. Instance classification by each of the algorithms

Thus, from the above two tables 4 and 5, we can see that the Multilayer Perception/Neural Networks algorithm out performs all of the others in accurately predicting the early onset of diabetes in Pima Indian women. It is also noticed that this algorithm has correctly classified 589 instances out of the training set with 786 instances. It can also be observed that the SVM algorithm has the least accurate performance, for the given training dataset.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

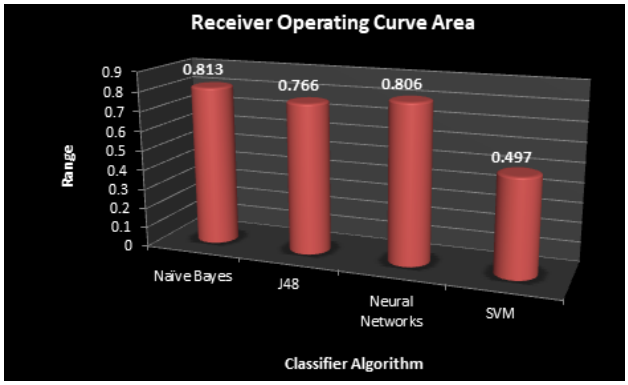

Fig 5. ROC Area for each classifier algorithm

The above fig 5, gives the receiver operating characteristic curve for each of the classifier algorithms. The ROC curve helps us understand the performance measures of the classification programs. These curves are used to choose the most appropriate cut off for any test. Here sensitivity predicts the exact right values and specificity depicts the exact negative values or the persons not affected. Thus there is higher overall accuracy in the tests being performed.

V. CONCLUSION AND FUTURE SCOPE

To conclude, various supervised classifier machine learning algorithms were applied onto the training set that was obtained by eliminating attributes that did not have much context towards predicting diabetes. This was done using the chi-squared test and only that attributes which were ranked highest and was given more weightage and more likely to predict the onset of diabetes was considered. It was seen that on this training set the Neural Networks algorithm provided the most accurate results.

For future work, the same method could be considered and many other machine learning classifiers algorithms could be considered to compare the most accurate one. This method can also be implemented on various other disease and medical datasets. In this study only a small sample dataset of 786 instances was considered, which is a drawback, but the same method could be applied for much larger datasets which would immensely help the scope of disease prediction and hence, provided much needed early detection, diagnosis and timely help to keep health issues under control and maybe find a way to eliminate them altogether in the future, through various research methods.

REFERENCES

[2] Abiraami TT, Sumathi A, "Analysis of Classification Algorithms for Diabetic Heart Disease", International Journal of Pure and Applied Mathematics, Vol. 118, No.20, 2018, 1925-1934.

[3] Deepti Sisodia, Dilip Singh Sisodia,,"Prediction of Diabetes using Classification Algorithms ", ICCIDS 2018 , Procedia Computer Science 132 (2018) 1578–1585, Science Direct.

[4] Ni, Nguyen, Garibay, Ivan, Akula, Ramya. (2019), "Supervised Machine Learning based Ensemble Model for Accurate *Prediction of Type 2 Diabetes*", IEEE Southeastcon, April 2019

[5] Sneha, N, Gangil, Tarun, *"Analysis of diabetes mellitus for early prediction using optimal features selection",* Journal of Big Data, Vol. 13.

[6] J. Anitha, A. Pethalakshmi, "Comparison of Classification Algorithms in Diabetic Dataset", International Journal of Information Technology (IJIT), Vol. 3, Issue 3, May-June 2017.

[7] K. Saravananathan, T. Velmurugan, "*Analyzing Diabetic Datausing Classification Algorithms in Data Mining*", International Journal of Science and Technology, Vol. 9 (43), November 2016.

[8] P. Chen, C. Pan, "Diabetes Classification Model based on Boosting Algorithms", BMC Informatics, Vol. 19, PMCID:PMC 58722396, March 2018.

[8] Quan Zou, Kaiyang Qu,Yamei Luo,Dehui Yin, Ying Ju, Hua Tang, "Predicting Diabetes Mellitus with Machine Learning Techniques", Front Genet, Vol. 9, PMCID:PMC6232260, November 2018.

[9] Dilip Kumar Choubey, Sanchita Paul, Santosh Kumar, "*Classification of Pima Indian Diabetes dataset using Naïve Bayes with genetic algorithm as an attribute selection*", Communication and Computing Systems, ISBN 978-1-138-02952-1, 2017.

[10] Aized Amin Soofi, "Classification *Techniques in Machine Learning: Application and Issues*",Journal of Basic & Applied Sciences, Vol 13, 459-465, 2017.

[11] Rahul Joshi, Minyechil Alehegn, "*Analysis and Prdiction of Diabetes disease using machine learning algorithm: Ensemble Approach*", International Research Journal of Engineering and Technology (IRJET), Vol 4, Issue 10, October 2017.

[12] Priya. B. Patel, Paryh. P. Shah, Hmanshu D. Patel, "*Analyze Data Mining Algorithms for Prediction of Diabetes*", International Journal of Engineering Development and Research (IJDER), Vol. 5, Issue 3, ISSN: 2321-9939.

[13]https://machinelearningmastery.com/use-machine-learning-algorithms-weka/, Accessed Marchl 03, 2019.

[14]https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes, Accessed April 26, 2019.

[15] Centers for Disease Control and Prevention. National diabetes statistics report, 2017. Centers for Disease Control and Prevention website. www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf (PDF, 1.3 MB) . Updated July, 18 2017. Accessed April 23, 2019.