International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Diabetes prediction using Data mining Classification Techniques

M. Manjusree, K.A. Sateesh Kumar

Abstract—Diabetes is one of the second largest disease in the world. In the recent survey it shows that there are overall 246 million people affected with this and in that women ratio is more. By the report of WHO, this figure is going to reach to 380 million by 2025. According to the American Diabetes Association,6% of the population are not aware that there are victims of diabetes and also every 21 sec at least for an individual diabetic test result is positive. With the technology advancement in the field of medical information, data is well maintained in the databases. This paper focuses on to diagnose data to provide the solution by observing the patterns in the data using various datamining classification techniques such as Naïve basis, Logistic regression, Decision tress etc.

Index Terms: Diabetes, Data mining, WEKA, Classification algorithms

I. INTRODUCTION

Diabetes mellitus is a metabolic disorder characterized by excess glucose levels in blood and urine, hyperlipaemia, negative nitrogen balance and sometimes ketonaemia.
There are two major types of diabetes mellitus:
Type 1: This type is also called as Insulin-dependent diabetes mellitus, juvenile onset diabetes mellitus. This is the less common type of diabetes and genetic predisposition is very low. In this type, there is beta cell destruction in pancreatic islets, as result insulin production decreases. Hence flowing insulin levels are too low and patients are inclined to ketosis.
Type II: This type is also called as Noninsulin-dependent diabetes mellitus, maturity onset diabetes mellitus. In this category, the there is no loss or reduction in beta cell mass has no loss or it may be a reasonable reduction . Genetic predisposition is high. The main causes of this type are there is irregularity in gluco-receptor of beta cells due to this they react at high glucose concentration or there is, reduce in number of insulin receptors or there is extra of hyperglycaemic hormones like glucagon or obesity all these factors result in insulin deficiency.[1]

Manuscript published on 30 September 2019
* Correspondence Author
Manjusree M*, School of Computer Science and Applications, REVA University, Bangalore, India.
K.A . Sateesh Kumar, Subject Matter Expert, Merittrac Pvt. Ltd, Bangalore, India.

This paper is organized as Section I introduction about diabetes, section 2 deals with datamining introduction and a brief information about various classification algorithms, Section 3 about WEKA and info about dataset, section4 contains test results and finally conclusion and references.

II. DATA MINING

It is the collection of methods used to mine the useful information from large volumes of data such as databases and data warehouses. It occupied a vital role to deal with complicated data sets to identify the various patterns hidden in the data and also processing the data to the required format for predicting the results. in the modern era and rapid growth of the data in various fields like medical, agriculture, industries, education and weather information using traditional methods to identify the problems in the data and to retrieve information based on the needs is the crucial task. Datamining is one of solution to these kinds of the problems which contain various classifications, clustering, statistical and association rules. It allows the user to mine the data with certain constraints though the data is in the form of noisy and hidden.
The biggest challenge in the datamining process is which algorithm when we can use for what kind the data to yield better results.[2]
In data mining, classification is one of the prediction models for data analysis. It describes future trends of the data and helps to group the data into different classes.
Data classification is a twostep process
Step1: Construction of the models with predefined set of labels or classes.
Step2: Use of the model for classifying the data.
The following features can be used for comparison of classifiers.

- Accuracy: it is the ability to provide the exactness of result according to the labelled class.
- Time: the computational cost for building and predicting the model.
- Scalability: it represents the feature to build the model even for large volume of the data.
- Interpretability: the model should be simple to use and easy to understand.

Here is the brief intro on various classification algorithms.[3]

A. Logistic Regression:

It is one of the statistical models for predicting the data using logistic functions to calculate the likelihood occurrences in the input. It combines the input data linearly by analysing the properties to provide the required output.

Logistic regression also named as sigmoid function since it will take values from 0 to 1 values to calculate the probability of the instances.

B. Naive Bayes :

It is a group of algorithm works on the principle of Bayes theorem. In this it will identify the class of a sample by calculating the probability.

It is one of the fastest computational algorithm and also it works on datasets which contain many dimensions.

To predict the class label(Ci) for sample(X),it will evaluate P(X/Ci)P(Ci) and assign to the sample to that particular class iff P(X/Ci)P(Ci)> P(X/Cj)P(Cj) for 1<=j<=m,j!=I i.e it assign to the class whose P(X/Ci)P(Ci) value is maximum.

C. Stochastic gradient descent (SGD):

In the recent years Stochastic gradient descent (SGD) algorithms have occupied substantial position for data classification due to its simplicity. It is one of the iterative algorithm for solving convex optimization problems.[4]

SGD Classifier combines many binary classifiers using the scheme "one versus all" (OVA) . Consider K is the number of classes in the classification then a binary classifier is learned for each class K that differentiates class K when compare to remaining K−1classes. For each classifier the confidence score can be computed and select the classifier with highest score as output.[5]

D.Kstar:

It is also considered instance based classifier in which for any instance the class can be identified based on the similarity of the existing classes which can be computed using a similarity function. The K store uses entropy based distance function to calculated neighbourhood of the sample. Entropy distance can be identified by means of transforming the complexity of an instance into another instance. The K* classification algorithm works by summarizing the probabilities of all the instances of a class to a new instance.[6]

$$K*(yi, x) = -\ln P*(yi, x) \quad (1)$$

E .Decision table

Choosing the correct attributes in the learning process is the strategy behind decision table classifier. This is achieved by measuring the cross validation performance of different attribute subsets and best performing subset is chosen for learning.

Decision table for dataset Ds with m attributes A1, A2, ..., Am with schema S (P1, P2,..., Pn , Value of class, Support, Confidence). A row Rj = (p1j, p2j, ..., pmj, cj,supj, confj) in table S represents a constraint for the classification where ajk ($1 \le k \le m$) can be either from DOM(Aj) or a special value ANY, ci $\in$ { c1, c2, ..., cm }, minsupport $\le$ supj ≤ 1, and minimumconf $\le$ confj ≤ 1 and minsup and minconf are predetermined thresholds.

The interpretation of the rule is

if (P1 = p1) and (P2 = p2) and … and (Pm= pm) then class = cj with probability confj and having support supj, where aj $\neq$ ANY, $1 \le j \le n$.[8]

F. Random Committee:

it is classifier under meta category of weka. It randomizes different base classifies and calculates the average predictions generated by using different base classification algorithms. It always considers data with high margins and it is very useful for the data sets with large dimensions in it. [9]

G. Random Forest:

It is a collection of various tree classifiers, each classifier is designed using random vector which is sampled from the input .It produces great predication in machine learning without using hyper-parameter tuning. It builds a forest which is an ensemble of decision trees. in simple words, Random forest builds more number of multiple decision trees and merges them to get more stable and accurate predictions.[10]

H. J48:

It is decision tree type of classifier. It construct the tree by using C 4.5 Algorithm. C 4.5 expands initial tree by using divide-and-conquer algorithm. If S is set of tuples of test data If all the tuples of S belong to same class then it becomes a leaf node else select a single attribute from S and apply a test at root level of the tree if any one branch of outcome matches with it then move to child node and apply the test with other attributes and continue with it until an outcome is matched at the leaf node. Repeat this process until all the tuples of S are covered.

Selection of correct path in decision tree has many methods whereas C4.5 uses two important heuristic properties to rank the test possibilities they are information gain and default gain ratio.For a numerical attribute A the possible outcomes could be A ≤ h or A > h where h is a threshold value calculated from the training data.

Tree constructed by using training data is then pruned to avoid overfitting problem. Pruning process is started from the leaves to the root and an is error estimated at a leaf node with N cases and E errors is calculated. C4.5 calculates different error estimations by replacing branch with leaf or leaf with sub tree or branch with other branch and chooses the best combination which gives least error value.[11]

I: SMO:

The full form of SMO is Sequential Minimal Optimization. This classifier is very effective to deal with the missing values of data it replaces the nominal data with binary data. SMO is used for training the Support Vector Machine (SVM). SVM deals with large quadratic Optimization Problems to do this SMO divide the problem into various sub problems. The memory requirements for SMO for training a dataset is linear with this feature it able to handle large datasets very easily. [12]

Algorithm	Classifies by using	Advantages	Disadvantages
Logistic Regression	Influence of independent attributes on outcome attribute	Designed for classification only hence much optimised	Works well only for binary outcomes
Naïve Bayes	Bayes theorem	Small amount of training data is enough to start the classification	It is considered as bad estimator
Stochastic Gradient Descent	loss functions and penalties for classification in probability modelling	Easy for the implementation and much efficient also	Scaling the classification and needs very large sample to start the classification

K-Nearest Neighbours	Votes of k nearest neighbour's	Very efficient for large noisy data	Computing K value is difficult and has high computational cost
Decision Tree	Sequence of rules constructed by using training data to conclude the decision	Limited data is enough to estimate the decision tree	The structure of the tree change completely even for the minimal change in the real time data
Random Forest	Multiple decision trees for different sample data	Over fitting problem can be minimized	Very difficult to implement
Support Vector Machine	Classifies by separating the training data into distinct groups with clear demarcation	Performs well for data with large dimensions	Probability estimators are calculated by using expensive fivefold crass validators

Table 1: Comparison of different methods of classification algorithm[13]

Table 1 provides the brief information about for what kind of problem which can method can be used and advantages and disadvantages of each method.

III. WEKA

In the present era of artificial intelligence, data has a prominent role in taking decisions. Many machine learning techniques was designed and deployed for data analyzation. This includes data filtering, predictive modelling and visualization. Many tools and languages are designed for the same, Waikato Environment for Knowledge Analysis (WEKA) tool is one among them and it is a freeware. It is embedded with collection of data mining techniques such as classification, clustering and association rules which can be used for designing different models for machine learning. This is the comparative analysis of different classification methods on pima indian diabetes dataset. The dataset consists of 798 instances with 8 attributes.Table [2]shows the information about the attributes in the dataset.[7]

TABLE2: ATTRIBUTES OF DATASET

Attribute	Relabeled values
1. Number of times pregnant	Preg
2. Plasma glucose concentration	Plas
3. Diastolic blood pressure (mm Hg)	Pres
4. Triceps skin fold thickness (mm)	Skin
5. 2-Hour serum insulin	Insu
6. Body mass index (kg/m^2)	Mass
7. Diabetes pedigree function	Pedi
8. Age (years)	Age
9. Class Variable (0 or 1)	Class

IV.TEST RESULTS

In WEKA for the diabetes dataset when the different classification algorithms were applied on with the same parameters using a cross fold 10 time and accrucaracy and confusion matrix are provided in the table3.
The various measurements used in WEKA are:

- Kappa Statistics:

It is the measure of agreement of chance corrected among between true class label to the classifier Equation to calculate the kappa statistics is:
K=P0-Pe /1-Pe
Where P0 is the relative agreement of observed
and Pe Chance of agreement probability

- Confusion Matrix:

It also considered as contingency matrix

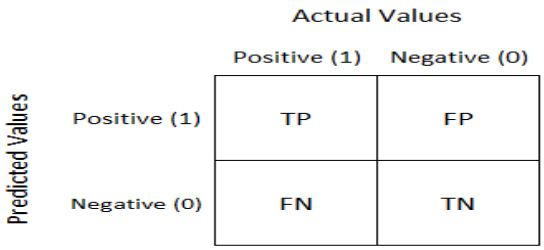

Fig 1: confusion matrix

Where TP stand s for True Positive,FP refers False Positive, FN represents False Negative,TN is the True Negative

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

Precision(P): It is the ratio between true positive to the total of false positive true positive and P = TP / (TP + EP)

- Recall (R):It is the ratio between true positive to the sum of true positive and false negative
  R=TP/(TP+FN)

- F-Score=2*precision*recall/(precision + Recall)
- ROC area: The ability to classify the records correctly with disease and without disease. The value for 0.5 to 1 can be considered as good classifier below 0.5 worst classifier.

Classification Algorithm	Time taken to build model(seconds)	Correctly Classified Instances	Confusion Matrix
Logistic Regression	0.38	77.21%	440 60 115 153
Naive Bayes	0.03	76.30%	422 78 104 164
SGD	0.16	77.99%	448 52 117 151
Kstar	0.02	69.14%	407 93 144 124
Random Committee	0.22	73.96%	420 80 120 148
Decision Table	0.16	71.22%	405 95 126 142

RandomForest	0.36	75.78%	418 82 104 164
J48	0.09	73.83%	407 93 108 160
SMO	0.09	77.34%	449 51 123 145

Table 3: Accuracy of different classification algorithms using WEKA

Table3 states that SGD algorithm is more accurate in predicting the results when compare to rest of the algorithms.

Classification Algorithm	Kappa statistic	RSE	RRE	Precision	Recall	MCC	ROC_Area
Logistic Regression	0.47	0.68	0.83	0.77	0.77	0.48	0.83
Naive Bayes	0.47	0.63	0.87	0.76	0.76	0.47	0.82
SGD	0.49	0.48	0.98	0.78	0.78	0.50	0.73
Kstar	0.29	0.72	1.04	0.68	0.69	0.29	0.71
Random Committee	0.41	0.67	0.90	0.73	0.74	0.41	0.79
Decision Table	0.35	0.76	0.90	0.71	0.71	0.35	0.77
RandomForest	0.46	0.68	0.85	0.75	0.76	0.46	0.82
J48	0.42	0.69	0.94	0.74	0.74	0.42	0.75
SMO	0.47	0.50	1.00	0.77	0.77	0.48	0.72

Table 4:Simualtion results

To study the ROC area covered by different classification algorithm following model is constructed and tested the result of the model are as shown in the figure 2.

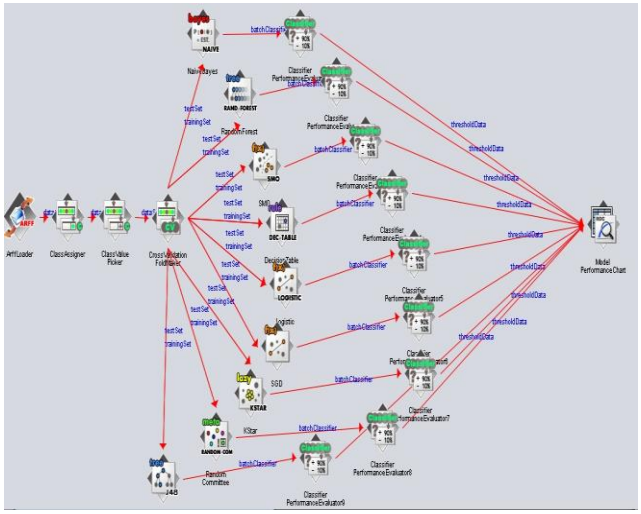

Fig 2: Flow model to construct ROC

Tested_positive:

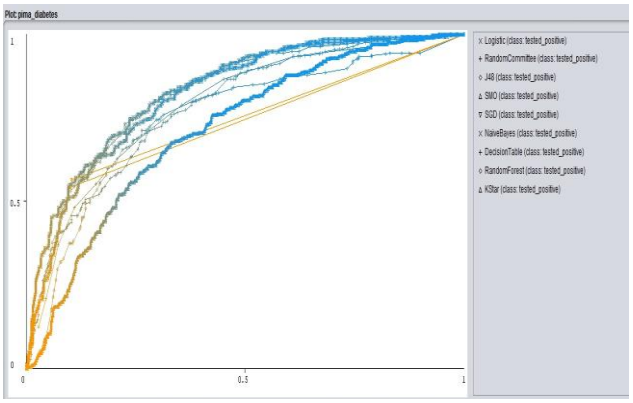

Fig 3: ROC area for positive case

Tested_Negitive

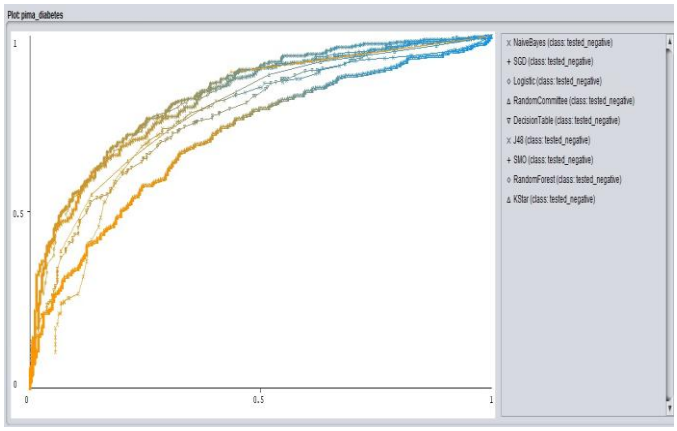

Fig 4:ROC Area for Tested _negetive

The above diagrams fig 3&4 shows the ROC area covered by various classifiers for tested+ve and tested–ve .

V CONCLUSION:

In this paper comparative study of various classifiers to predict the diabetes was done. From the results it is observed that none of the classifier gives 100% accuracy for the given samples but at least some classifiers shows the better performance over others. According to the ROC area logistic regression is best classifier. From Accuracy results, SGD is better when compare to the other classifiers. In future prediction accracy can be improved with hybrid classifiers.

REFERENCES:

1. Essentials of medical pharmacology, 6th edition, KD tripathi, Jaypee brothers medical publishers, 2008, pp-254-55
2. Y. Li, F. Advisor, T. Beaubouef, "Data Mining: Concepts Background And Methods Of Integrating Uncertainty In Data Mining", Ccsc:Sc Student E-Journal., vol. 3, pp. 2-7, 2010.
3. Han, J., Kamber, M.: Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers (2000).
4. S. Song, K. Chaudhuri and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, 2013, pp. 245-248. doi: 10.1109/GlobalSIP.2013.6736861
5. https://scikit-learn.org/stable/modules/sgd.html

6. Dayana C. Tejera Hernández," An Experimental Study of K* Algorithm", I.J. Information Engineering and Electronic Business, 2015, 2, 14-19 Published Online March 2015 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijieeb.2015.02.03
7. Iyer, Aiswarya & Jeyalatha, S & Sumbaly, Ronak. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process. 5. 1-14. 10.5121/ijdkp.2015.5101.
8. Sushilkumar Rameshpant Kalmegh "Comparative Analysis of the WEKA Classifiers Rules Conjunctiverule & Decisiontable on Indian News Dataset by Using Different Test Mode", International Journal of Engineering Science Invention (IJESI)
9. 9.http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/RandomCommittee.html
10. M. Pal (2005) Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26:1, 217-222, DOI: 10.1080/01431160412331269698
11. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A.F.M. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2008.
12. S. Singaravelan, D. Murugan and R. Mayakrishnan ," Analysis of Classification Algorithms J48 and Smo on Different Datasets", World Engineering & Applied Sciences Journal 6 (2): 119-123, 2015 ISSN 2079-220,DOI: 10.5829/idosi.weasj.2015.6.2.22162
13. .https://www.analyticsindiamag.com/7-types-classification-algorithms/

AUTHORS PROFILE

M. Manjusree has done her master degree in computer applications and published 10 papers in various conferences and journals. She has organized national conferences in REVA University. She was a key speaker in workshop organized on computer graphics using OPENGL. Her research interest includes Data Analytics and Information Security.

K.A. Sateesh Kumar has done his master degree in computer applications and published 3 papers in various conferences and journals. He was a key speaker in workshop organized on Network Simulation using NS2 and Creating project using C#.NET. His research interest includes Data Analytics, Cloud Computing and Robotic Process Automation.