Classification of Pima Indian Diabetes dataset using Decision Tree Techniques

Nilesh Verma
Atal Bihari Vajpayee Vishwavidyalaya, Bilaspur, Chhattisgarh, India

Abstract— Diabetes means blood sugar is above desired level on a sustained basis. Diabetes has become a modern day life style disease affecting millions of people around the world. The prime objective of this research work is to provide a better classification of diabetes. There are already several existing method, which have been implemented for the classification of diabetes dataset. In medical sector, the classifications systems have been widely used to exploit the patient's data and make the predictive models or build set of rules. Data mining is growing in relevance to solving real world problems and hence this can be applied to the diabetes problem as well. The study proposes to use the UCI repository dataset called PIMA Indians Diabetes dataset and decision tree algorithms like C4.5, J48, ID3 and NBs etc. The comparison study includes parameters like sensitivity, accuracy, specificity and features or nodes selected. This hybrid model enables to accurately classify the diabetes dataset and help the people providing treatment as well as those suffering from the disease.

Keywords: WEKA (Waikato Environment for Knowledge Analysis), Data Mining, Decision Tree, C4.5, J48, ID3, Artificial Neural Network, Sensitivity, Accuracy, Specificity, precision, F-measure, NBTree

I. INTRODUCTION

Data mining is a process to discover interesting knowledge, such as associations, patterns, anomalies, changes and significant structures from large amount of data stored in databases or other information repositories. In the procedure of data mining the former data is explained and future rules are calculated by data analysis. Data mining is a major advancement in the type of analytical tools. Data mining is a multi-disciplinary field which is a combination of machine learning, statistics, database technology and artificial intelligence. This technique includes a number of phases: Business understanding, Data understanding, Data preparation, Modeling, Evaluation, and Deployment. Data mining has proven to be very beneficial in the field of medical analysis as it increases diagnostic accuracy, to reduce costs of patient treatment and to save human resources. There are various data mining techniques such as Association, Classification, Clustering, Neural Network and Regression.

Classification of Pima Indians Diabetes Dataset using decision tree techniques. In medical science, diagnosis of health condition is a very challenging task. Diabetes Mellitus is one of the most important serious challenges in both developed and developing countries. Medical history data comprises of a number of tests essential to diagnose a particular disease and the diagnosis are based on the experience of the physician; a less experience physician can diagnose a problem incorrectly. Here, Decision Tree technique has been used for the classification of the diabetes diagnosis. And decision tree algorithms like C4.5, J48, ID3 and NBs etc. The comparison study includes parameters like sensitivity, accuracy, specificity and features or nodes selected. This hybrid model enables to accurately classify the diabetes dataset and help the people providing treatment as well as those suffering from the disease.

II. DATA DESCRIPTION:

The Pima Indian diabetes database collected from UCI Machine Learning Repository available at this URL- http://www.ics.uci.edu/mlearn/MLRepository.html consists of two categories namely tested positive and tested negative(Diabetes or Non-Diabetes). Pima Indian diabetes dataset has shown details on Table 2.1

The objective of this data set was diagnosis of diabetes of Pima Indians. Based on personal data, such as age, number of times pregnant, and the results of medical examinations, e.g., blood pressure, body mass index, result of glucose tolerance test, etc., it is tried to decide whether a Pima Indian individual was diabetes positive or not.

Data Set Characteristics	Multivariate
Attribute Characteristics	Integer, Real
Area	Life(Health care)
Number of Instances	768
Total Number of Attributes	9
Missing Attributes Status	No
Noisy Attributes Status	No
Input Attributes	8
Output Classes	2 (Diabetes and Non-Diabetes)

Table 2.1: Pima Indian Diabetes Data Set information

A. *Attributes Information*

Classification of Pima Indian diabetes dataset has 9th attributes. But input attributes are used only 8 attributes in our experiment seat and each 9 attributes has one class. It class show two output classes diabetes and non-diabetes. Attributes information are show details on Table2.1.1.

Attribute ID	Attribute Name	Attribute Description	Type
F1	Preg	Number of times pregnant.	
F2	Plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test.	Numeric
F3	Pres	Diastolic blood pressure (mm Hg)	Numeric
F4	Skin	Triceps skin fold thickness (mm)	Numeric
F5	Insu	Hour serum insulin (mu U/ml)	Numeric
F6	Mass	Body mass index (weight in kg/(height in m)^2)	Numeric

F7	Pedi	Diabetes pedigree function	Numeric
F8	Age	Age (years)	Numeric
F9	Class	Diabetic or Non-Diabetic (0 or 1)	Numeric

Table 2.1.1: Pima Indian Diabetes Data Set Attributes information For Each Attributes

B. *Class Distribution:*

My Pima Indian diabetes dataset has two classes set are defined as follows:

1) Diabetes.
2) Non-diabetes.

III. Performance measures

Performance of model can be evaluated various performance measures: classification accuracy, sensitivity and specificity. These measures are evaluated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) details show on Table3.1.

Actual Vs. Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 3.1: Confusion matrix

TP: Positive samples classify correctly.
FP: Positive samples classify incorrectly.
TN: Negative samples classify correctly.
FN: Negative samples classify incorrectly.

A. *Measures Description:*

Various performance measures like sensitivity, Precision, F-measure, Error Rate specificity and accuracy are calculated using this matrix: Where N is totals number of samples.

Accuracy: It is description of systematic error a measure of statistical bias as these causes a different between a results a “tree” value.

1) Sensitivity:

It is also called true positive rate the recall or probability of detection. Sensitivity measure the proportion of positive that are correctly identified as such.

2) Specificity:

It is also called the true negative rate this measure the proportion of negative that are correctly identified as such.

3) Precision:

Precision is a description of random error a measure of statistical variability.

4) F-Measure:

It is also called a F1-score, F-score.F1-score a measure of test accuracy it considers both the precision P and the recall R of the test to compute score.

5) Error:

Error rate of a classifier was defined as the percentage of the dataset incorrectly classified by the method. It is the probability of misclassification of a classifier.

Measures	Mathematical Form
Accuracy	(TP+TN) / (TP+FP+TN+FN) OR N
Sensitivity(Recall)	TP / (TP+FN)
Specificity	TN / (TN +FP)
Precision	TP / (TP +FP)
F-measure	2* (Precision*Recall) / (Precision + Recall)
Error Rate	(FP+FN) / N or 1-Accuracy

Table 3.A.1: Measures and mathematical formula.

IV. Methodology

There are various classification techniques are used to find high accuracy, sensitivity and specificity of our experiment data set and reduce errors. We use different classification techniques in this research. Those techniques with running parameters are given below:

A. *J48:*

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification and for this reason, C4.5 is often referred to as a statistical classifier.C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy.

B. *Naïve Bayes:*

The Naïve Bayes classifier provides a simple approach, with clear semantics, representing and learning probabilistic knowledge. It is termed naïve because is relies on two important simplifying assumes that the predictive attributes are conditionally independent given the class, and it assumes that no hidden or latent attributes influence the prediction process.

C. *CART:*

Classification and Regression Tree (CART) is one of commonly used Decision Tree algorithms. In this post, we will explain the steps of CART algorithm using an example data. Decision Tree is a recursive partitioning approach and CART split each of the input node into two child nodes, so CART decision tree is Binary Decision Tree. At each level of decision tree, the algorithm identify a condition - which variable and level to be used for splitting input node (data sample) into two child nodes.

D. *AD Tree:*

The alternating decision tree (AD Tree) is a successful classification technique that combines decision trees with the predictive accuracy of boosting into a set of interpretable classification rules. The original formulation of the tree induction algorithm restricted attention to binary classification problems.

V. Experimental work in WEKA

A. *Training and Testing of Diabetes dataset:*

The pima Indian diabetes data set has total 768 instances each instance divided into two parts the first one is training part and second is testing. The training part includes 576 instances and the remaining 192 instances is use testing part. In this database, there are 768 numbers of instances and 9 number of attributes it has 1 attribute is a class.

B. *Confusion Matrix: Training Time*

Attributes = 8

Instances = 576

Actual Vs. Predicted	Positive	Negative
Positive	175	23
Negative	73	305

Table 5.I.1: Confusion matrix in case of J48

Actual Vs. Predicted	Positive	Negative
Positive	141	57
Negative	58	320

Table 5.I.2: Confusion matrix in case of ID3

Actual Vs. Predicted	Positive	Negative
Positive	128	70
Negative	57	321

Table 5.I.2: Confusion matrix in case of ID3

Actual Vs. Predicted	Positive	Negative
Positive	137	61
Negative	60	318

Table 5.I.4: Confusion matrix in case of Cart

Actual Vs. Predicted	Positive	Negative
Positive	136	62
Negative	57	321

Table5.I.5: Confusion matrix in case of BFTree

Actual Vs. Predicted	Positive	Negative
Positive	123	75
Negative	24	354

Table 5.I.6: Confusion matrix in case of C4.5

Instances = 192

Actual Vs. Predicted	Positive	Negative
Positive	108	14
Negative	16	54

Table 5.I.7: Confusion matrix in case of J48

Actual Vs. Predicted	Positive	Negative
Positive	111	11
Negative	27	43

Table 5.I.8: Confusion matrix in case of ID3

Actual Vs. Predicted	Positive	Negative
Positive	109	13
Negative	17	53

Table 5.I.9: Confusion matrix in case of NBTree

Actual Vs. Predicted	Positive	Negative
Positive	109	13
Negative	27	43

Table 5.I.10: Confusion matrix in case of Cart

Actual Vs. Predicted	Positive	Negative
Positive	109	13
Negative	27	43

Table 5.I.11: Confusion matrix in case of BFTree

Actual Vs. Predicted	Positive	Negative
Positive	108	14
Negative	16	54

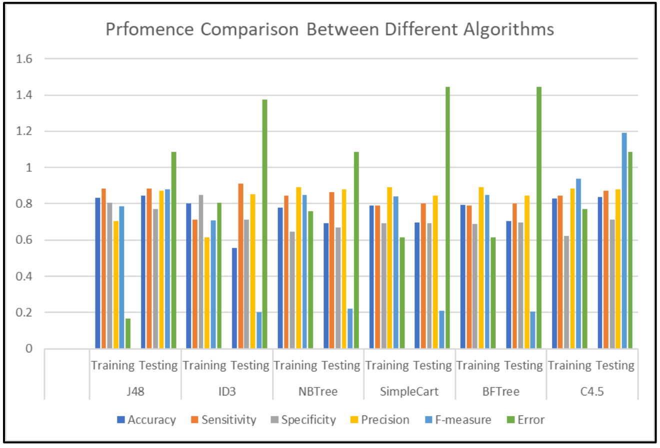
Table 5.I.12: Confusion matrix in case of C4.5

C. Confusion Matrix === Testing Time

Attributes = 8

Techniques	Stage	Accuracy	Sensitivity	Specificity	Precision	F-measure	Error
J48	Training	0.833	0.884	0.807	0.706	0.785	0.166
	Testing	0.8437	0.885	0.771	0.871	0.878	0. 1562
ID3	Training	0.8003	0.712	0.847	0.614	0.709	0.804
	Testing	0. 802	0.91	0.71	0.854	0.1996	0. 1979
NBTree	Training	0.7795	0.8437	0.646	0.893	0.849	0.757
	Testing	0.692	0.865	0.668	0.879	0.2204	0. 1562
SimpleCart	Training	0.789	0.7916	0.692	0.893	0.841	0.614
	Testing	0.695	0.801	0.694	0.845	0.210	0. 2083
BFTree	Training	0.793	0.7916	0.687	0.893	0.849	0.614
	Testing	0.705	0.801	0.696	0.845	0.206	0. 2083
C4.5	Training	0.8281	0.8437	0.621	0.885	0.937	0.771
	Testing	0.837	0.871	0.713	0.878	0. 1718	0. 1562

Table 5.I.13: Performance comparison between different algorithms with original Feature set for training set and testing set.



D. 10-Fold cross validation:

1) Confusion Matrix

Actual Vs. Predicted	Positive	Negative
Positive	432	68
Negative	99	196

Table 5.II.1: Confusion matrix in case of J48

Actual Vs. Predicted	Positive	Negative
Positive	442	58
Negative	117	150

Table 5.II.2: Confusion matrix in case of ID3

Actual Vs. Predicted	Positive	Negative
Positive	411	89
Negative	108	160

Table 5.II.3: Confusion matrix in case of NBTree

Actual Vs. Predicted	Positive	Negative

Positive	422	78
Negative	104	164

Table 5.II.4: Confusion matrix in case of NaiveBayes

Actual Vs. Predicted	Positive	Negative
Positive	406	94
Negative	108	160

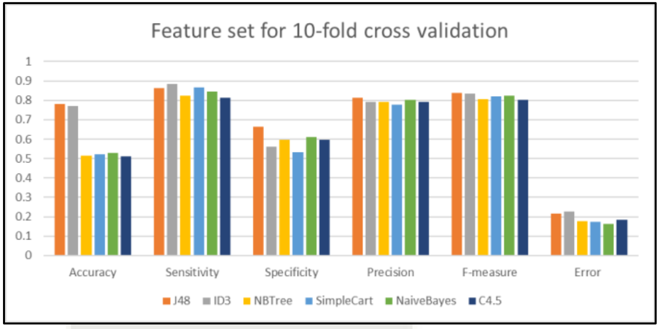
Table 5.II.5: Confusion matrix in case of C4.5

Actual Vs. Predicted	Positive	Negative
Positive	434	66
Negative	125	143

Table 5.II.6: Confusion matrix in case of CART

Techniques\Measured	Accuracy	Sensitivity	Specificity	Precision	F-measure	Error
J48	0.782	0.864	0.664	0.814	0.838	0.217
ID3	0.770	0.884	0.562	0.791	0.835	0.227
NBTree	0. 743	0.822	0.597	0.792	0.807	0. 256
SimpleCart	0. 751	0.868	0.534	0.776	0.82	0. 248
NaiveBayes	0. 763	0.844	0.612	0.802	0.823	0. 236
C4.5	0. 736	0.812	0.597	0.79	0.801	0. 263

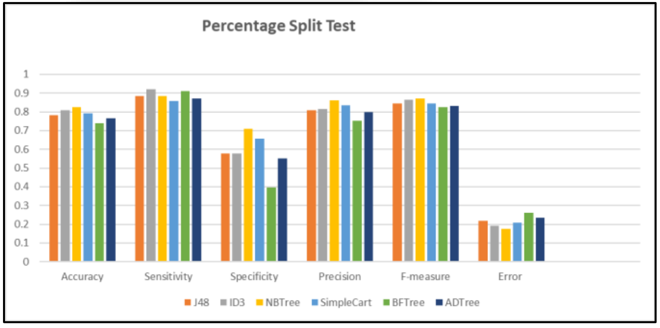
Table 5.II.7: Performance comparison between different algorithms with original Feature set for 10-fold cross validation.



E. *Percentage Split Test:*

Techniques\Measured	Accuracy	Sensitivity	Specificity	Precision	F-measure	Error
J48	0.782609	0.883116	0.578947	0.809523	0.844719	0.217391
ID3	0.808696	0.922077	0.578947	0.816091	0.865852	0.191304
NBTree	0.826087	0.883116	0.710526	0.860759	0.871794	0.173913
SimpleCart	0.791304	0.857142	0.657894	0.835443	0.846153	0.208696
BFTree	0.73913	0.90909	0.39473	0.75268	0.82352	0.26087
ADTree	0.765217	0.870129	0.552631	0.797619	0.832297	0.234783

Table 5.III.1: Performance comparison between different algorithms with original Feature set for 85:15 percentage split test option.



F. *Feature subset selection:*

Feature subset selection is an important problem in knowledge discovery, not only for the insight gained for determining relevant modeling variables, but also for the improved International Journal of Decision Science & Information Technology, understandability, scalability, and, possible accuracy of the resulting models. In the Feature selection the main goal is to find a feature subset that produces higher classification accuracy. In this research work, we have used Information gain feature selection technique.

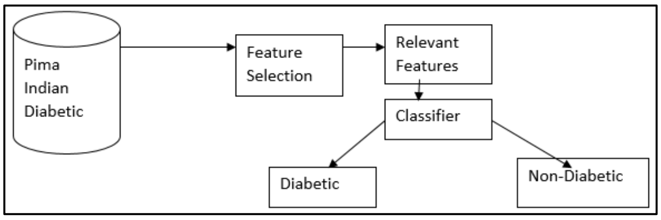


Figure 5.IV.1: Feature Selection Model

Measures after applying feature selection on Pima Indian diabetes data set used attribute evaluation (Gain Ratio Attribute Evaluation), search method (Ranking). The ranking of features from less important to high important as shown F2, F6, F8, F1, F5, F7, F4, F3. In this experiment we have eliminated the less important feature one by one and give to the best model.

1) *Confusion Matrix*

Using J48graft classifier

Actual Vs. Predicted	Positive	Negative
Positive	468	32
Negative	90	178

Table 5. IV.1: Confusion matrix in case of (F6, F8, F1, F5, F7, F4, F3)

Actual Vs. Predicted	Positive	Negative

Positive	471	29
Negative	98	170

Table 5. IV.2: Confusion matrix in case of (F2, F6, F8, F1, F5, F7, F4)

Actual Vs. Predicted	Positive	Negative
Positive	469	31
Negative	98	170

Table 5. IV.3: Confusion matrix in case (F2, F6, F8, F1, F5, F7)

Actual Vs. Predicted	Positive	Negative
Positive	468	32
Negative	126	142

Table 5. IV.4: Confusion matrix in case (F2, F6, F8, F1, F5)

Actual Vs. Predicted	Positive	Negative
Positive	461	39
Negative	123	145

Table 5. IV.5: Confusion matrix in case (F2, F6, F8, F1)

Actual Vs. Predicted	Positive	Negative
Positive	461	39
Negative	123	145

Table 5. IV.6: Confusion matrix in case (F2, F6, F8)

Actual Vs. Predicted	Positive	Negative
Positive	443	57
Negative	118	150

Table 5. IV.7: Confusion matrix in case of Features (F2, F6)

Feature Id	Accuracy	Sensitivity	Specificity	Precision	F-measure	Error
(F2,F6,F8,F1,F5,F7,F4,F3)(8)	0. 841	0.936	0.664	0.839	0.885	0. 158
(F2,F6,F8,F1,F5,F7,F4)(7)	0. 834	0.942	0.634	0.828	0.881	0. 165
(F2,F6,F8,F1,F5,F7)(6)	0. 832	0.938	0.634	0.827	0.879	0. 167
(F2,F6,F8,F1,F5)(5)	0. 794	0.936	0.53	0.788	0.856	0. 205
(F2,F6,F8,F1)(4)	0. 789	0.922	0.541	0.789	0.851	0. 210
(F2,F6,F8)(3)	0. 789	0.922	0.541	0.789	0.851	0. 210
(F2,F6)(2)	0. 772	0.886	0.56	0.79	0.835	0.227

5. IV. 8: Measures after applying feature selection on Pima Indian diabetes data set using J48graft

VI. CONCLUSION:

Diabetes is a problem with your body that causes blood sugar levels to rise higher than normal. Diabetes can cause serious health complications including blindness, blood pressure, heart disease, kidney disease and nerve damage, etc. which is hazardous to health. The PIDD obtained from UCI repository of machine learning databases on which NBs, J48, ID3, C4.5 and CART method have been applied. For the future research work, we suggest to developed an expert system of diabetes, which will provide good sensitivity, Precisions, f-measure, accuracy and this is possible to achieve only by using different Attribute selection and classification method which, could significantly decrease healthcare costs via early prediction and diagnosis of diabetes. The proposed method can also be used for other kinds of diseases but not sure that in all the medical diseases either same or greater than the existing results.

In this study, we have taken various classification methods and ensemble them to give the new hybrid model in the search of finding the better result in terms of Accuracy, Specificity Precisions, f-measure and Sensitivity. According to Table 5.IV.1, we came to the conclusion that our model has achieved the highest Accuracy of 0.841 with 8 features and with the help of Table 5.IV.1, it achieve the highest Sensitivity of 0.942 and with the help of Table 5.I.13, it achieve the highest Specificity of 0.847 and F-measure 0.937. and with help of Table 5.I.13 it achieve the highest Precision of 0.893.

REFERENCES

[1] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, San Francisco, Morgan Kauffmann Publishers, 3rd edition, (2012).

[2] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, San Francisco, Morgan Kauffmann Publishers, 2nd edition, (2012).