Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach

Article · October 2014

DOI: 10.12777/ijse.7.2.155-160

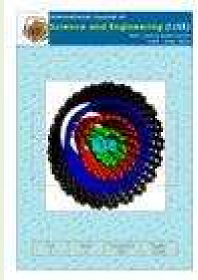
2 authors:

Kawsar Ahmed
Mawlana Bhashani Science and Technology …

Tasnuba Jesmin

International Journal of Science and Engineering (IJSE)

Home page: http://ejournal.undip.ac.id/index.php/ijse



Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach

Kawsar Ahmed[1], Tasnuba Jesmin[2]

[1]Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

[2]Department of Software Engineering, Daffodil International University, 102-Sukrabad, Mirpur Road, Dhanmondi, Dhaka-1207, Bangladesh

Email: [1]kawsar_it08050@yahoo.com; [2]tasnuba_it08005@yhaoo.com

Abstract - The goal of this paper discusses about different types of data mining classification algorithms accuracies that are widely used to extract significant knowledge from huge amounts of data. Here illustrate 20 classifications of supervised data mining algorithms base on type-2 diabetes disease dataset perspective to Bangladeshi populations. In this paper we compare 20 classification algorithms by measuring accuracies, speed and robustness of those algorithms using WEKA toolkit version 3.6.5. Accuracies of classification algorithms are measured in 3 cases like Total Training data set, 10 fold Cross Validation and Percentage Split (66% taken). Speed (CPU Execution Time) and error rate also measured as like as accuracy. Firstly checked top perform algorithms that have best outcome for different cases and then ranked top outcomes algorithms. Finally ranked best 5 algorithms among 20 algorithms based on their accuracies.

Keywords *— Accuracy, Classification Algorithms, Confusion Matrix, Data Mining, Error Rate, Type-2 Diabetes in Bangladesh, WEKA toolkit*

Submission: Sept 17, 2014 Corrected : October 12, 2014 Accepted: October 15, 2014

Doi: 10.12777/ijse.7.2.150-154

How to cite this article: Ahmed, K., and Jesmin, T. (2014). Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach, *International Journal of Science and Engineering*, *7(2),155-160*. Doi: 10.12777/ijse.7.2.150-154

I. INTRODUCTION

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data. Of course, there will be problems. Many patterns will be banal and uninteresting. Others will be spurious, contingent on accidental coincidences in the particular dataset used. Data Mining is used to extract information from the raw data in databases—information that is expressed in a comprehensible form and can be used for a variety of purposes like as Type-2 Diabetes patients classified.

Now-a-days the incidence of diabetes has soared worldwide and is expected to keep growing, with the greatest increase seen in metabolic forms of diabetes, notably type 2. Diabetes is one of fatal, metabolic and costly disease that increases blood sugar level. It is not only a disease but also responsible of occurring different kinds of diseases like heart attack, blindness, kidney diseases etc. If diabetes goes out of controlled then it increases blood glucose level more than 200mgI/dL which leads to micro and macro vascular disease complications (K. Ahmed et. al., 2012).

The estimated number of people with diabetes has jumped from 30 million in 1985 to 150 million in 2000 and then to 246 million in 2007, according to the International Diabetes Federation. It expects this number to hit 380 million by 2025. According to World Health Organization there are more than one million people in this world who are suffering from diabetes. The prevalence of Type 2 Diabetes is increasing at an alarming rate in a developing country like Bangladesh in recent years (Unwin N. et. al., 2009).

Now patient and non-patient information of type-2 diabetes perspective to Bangladesh used to find out classification algorithm's accuracy and error rate (ER). This has done using Weka version 3.6.5, a comprehensive software resource, written in the Java language, has been created to illustrate the ideas called the Waikato

Environment for Knowledge Analysis (Weka), which is available as source code on the World Wide Web at http://www.cs.waikato.ac.nz/ml/weka.

The main goal of this paper is to compare different classification algorithms (taken 20 classification algorithms in different types) accuracies not only for any single cases but also for every cases as well as select top 5 algorithms what is averagely good for every cases. Finally those will be shown in graphs.

Table 1.Parameters of Diabetics Data sets

No.	Parameters	Descriptions
1	Age	Taken as numeric value (years)
2	Relatives?	Taken three types input like 1. No, 2. Grandparent, Uncle, Aunty, 3. Parents, Brother, Sister
3	Sugar?	Taken two types input like 1. Yes, 2. No If No. taken another 2 types input, before 1. Yes, 2. No
4	Vegetables eat?	Taken two types input like 1. Yes, 2. No
5	Physical Activity?	Taken two types input like 1. Yes, 2. No
6	BMI	Taken as numeric value (Weight as Kg./ (Height as Meter)^2)
7	Red Meat?	Taken two types input like 1. Yes, 2. No
8	Waist	Taken as numeric value (Cm)

II. Background

A widely recognized formal definition of data mining can be defined as “Data mining is the non- trivial extraction of implicit previously unknown and potentially useful information about data” (Frawley and Piatetsky-Shapiro, 1996). Data mining is often defined as finding hidden information in a database. Data mining has some fields to analysis of data such as classification, correlation, clustering, association rule etc. Now-a-days many organizations have been used data mining intensively and extensively. In-healthcare, data mining is becoming increasingly popular (H. C. Koh and G. Tan., 2011). Data mining provides the methodology and technology to identify the useful information of data for decision making.

Classification, major part of data mining can be classified into 2 sectors (One is supervised and another is unsupervised). There present around 60 algorithms for classification. But all are not enough good according to need. Classification algorithms have 3 basic criteria like accuracy, error rate and execution time for choice. For different kinds of data different classification algorithms are used. Here we use type-2 diabetes patient information to classify and analysis there performance that described in section 5 briefly.

Accuracy means to percentage of correctly classified. The accuracy is calculated based on addition of true positive and true negative followed by the division of all possibilities. This can briefly describe using Table-4. Sensitivity and Specification will also describe using Table-4 in section 3. Accuracy is measured in 3 ways like total training data, 10 fold cross validation and percentage split.

Tenfold cross-validation is the standard way of measuring the error rate of a learning scheme on a particular dataset; for reliable results, 10 times 10-fold cross-validation.10-fold cross-validation has become the standard method in practical terms. Different 10-fold cross-validation experiments with the same learning method and dataset often produce different results, because of the effect of random variation in choosing the folds themselves. Percentage split holds out a certain percentage of the data for testing. Splits a dataset according to the given percentage into a train and a test file, here use 66% split.

Mean absolute error are mainly consider here as Error rate. Here we consider below 2 seconds of execution time. Table-2 shows accuracy of different sectors and theirs average.

This paper mainly discusses about classification algorithm’s accuracy with execution time and error rate using Weka toolkit. Here also discusses accuracy by dividing 3 subsectors that are briefly described in below sections. Section 2 describes fundamental parts of classification algorithm and section 4 and section 5 describe about different types of classification algorithms and their performances respectively.

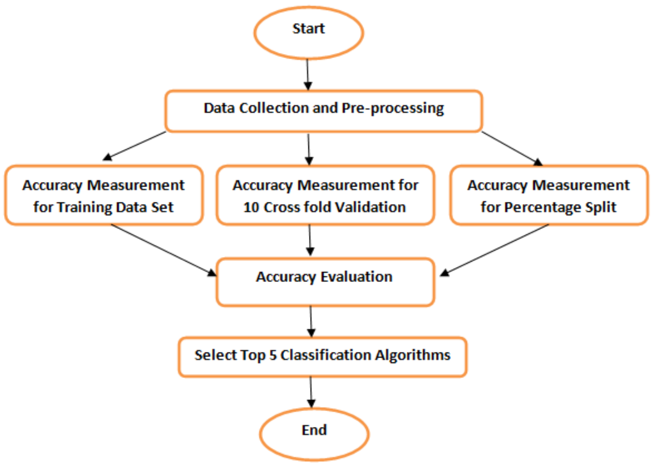


Figure 1.Graphical Representation of Working Process

III. Working Process

In this section mainly describes about whole working process that is shown in Fig. 1. This paper performance analysis process is sub-sectional. First subsection discusses about data collection and pre-processing. Next discusses performance of classification algorithm.

Data collection and Pre-processed:

400 patients’ data (200 diabetes patients and 200 non-diabetes patients) is collected from different diagnostic centre. There are 200 male and 200 female patients whose age between 20 to 80 years old. From the previous studies 13 risk factors were considered for type 2 diabetes assessment in Bangladeshi population, which includes- age, gender, hereditary, previous health examination, use of anti-hypersensitive drugs, smoking,

food habit, physical activity, BMI (Body Mass Index), waist circumference, mental trauma, uptake of red meat, hypertension, heart disease. Mostly associated attributes of Diabetes prediction data with risk factors parameters and their description are shown in Table-1. Those data mainly collected perspective to Bangladesh from (K. Ahmed et. al., 2012).

In this paper weka version 3.6.5 is used to test accuracy of different classification algorithms. Sometimes data maybe missing and need some specific formats for Weka. So there need to pre-processing data. Weka support Arff (attribute-relation file format), CSV, and JDBC database format data. So data will be saved according to above format. Then run weka. Here are used both ARFF and CSV format for testing accuracy.

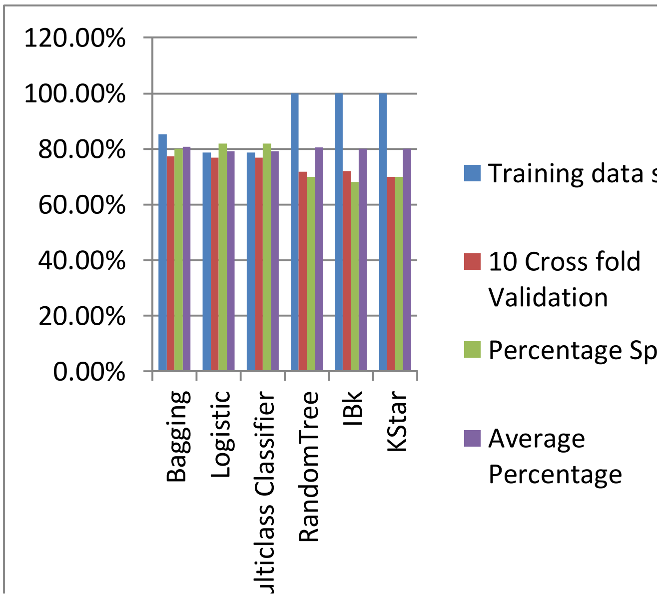


Figure 2.Graphical Representations of Highest Accuracies Algorithms for Different Cases

Accuracy Measurement:

For accuracy measurement here considers 20 classification algorithms. Brief description of those algorithms is given in section 4. Firstly run weka version 3.6.5, select diabetics data file and measure accuracy for 3 sectors and average (through rows) those accuracies that are shown in Table-2, then we took highest accuracies algorithm that are shown in Table-3. Accuracy mainly calculated using Confusion Matrix (CM). It can be represented by Table-4. It is known that accuracy means the ratio of total number of correct classification attributes and total number of using attributes. So from confusion matrix (Table-4) classification accuracy can be represented as below equation

Accuracy = (TP+TNP)/ (TP+FP+FNP+TNP) ……… (1)

Sensitivity = TP / (TP+FNP) ……… (2)

Specificity= TNP / (TNP+FP) ……… (3)

Result Evaluation:

Now find the averages of accuracies through columns (for particulars cases) in Table-3. And mark of those accuracies that is more or around to average. Here highest marking algorithm is ranked. Then choses best 5 algorithms. In Table-3 last column shows the ranking of best 5 classification algorithms.

IV. CLASSIFICATION ALGORITHMS

In this section will be discussed about 20 classification algorithms that are used for accuracy prediction. The classification algorithms can be sub sectioned that are briefly described into below 4.1 to 4.7 sub-sections.

Bayesian Networks Classifiers:

The Naïve Bayes classifier (Standard Probabilistic Classifier) that can only represent simple distributions produces probability estimates rather than predictions. This estimates the probability that a given instance belongs to that class and allows predictions to be ranked, and their expected cost to be minimized. Bayesian networks provide a good way of using them at prediction time as well as complex data (Ian H. Witten and Eibe Frank, 2005).

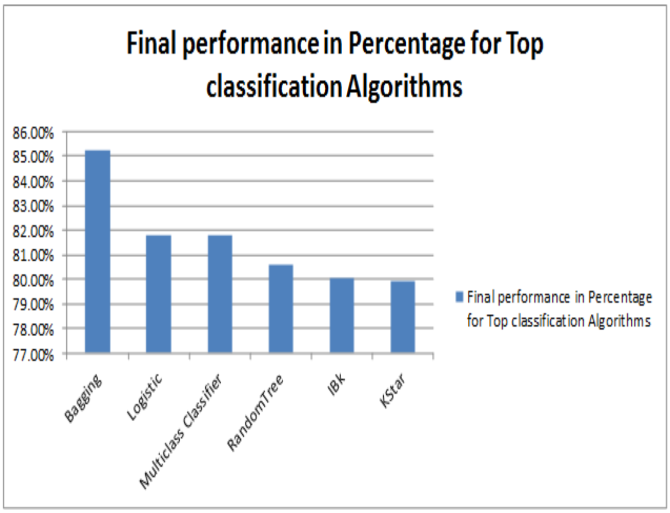


Figure 3.Graphical Representations of Final Accuracies of Top Algorithms

Trees Classifiers:

Here uses 4 tree bases classification algorithms like BFTree, FT, RandomTree, J48 (implements C4.5 revision 8) etc.

BFTree is a classification algorithm that builds a decision tree using a best-first expansion of nodes rather than the depth-first expansion used by standard decision tree learners (such as C4.5). Pre- and postpruning options are available that are based on finding the best number of expansions to use via cross-validation on the training data. While fully grown trees are the same for best-first and depth-first algorithms, the pruning mechanism used by BFTree will yield a different pruned tree structure than that produced by depth-first methods (Gama, J., 2004). Another tree base classification algorithm is FT that builds a functional tree with oblique splits and linear functions at the leaves. FT algorithm uses standard C4.5 pruning rather than minimal cost-complexity pruning. Trees built by RandomTree test a given number of random features at each node, performing no pruning. RandomForest constructs random forests by bagging ensembles of random trees. J48 is a classification algorithm of C4.5 decision tree learner. The algorithm, the classifier it builds, and a procedure for outputting the classifier is all part of that instantiation of the J48 class. It includes references to instances of other classes that do most of the work (Ian H. Witten et. al., 2011).

Rules Classifiers:

Here uses 4 rule bases classification algorithms like DecisionTable, JRip, OneR, ZeroR etc. DecisionTable builds a simple decision table majority classifier that evaluates feature subsets using best-first search. An option uses the nearest-neighbor method to determine the class for each instance that is not covered by a decision table entry, instead of the table's global majority, based on the same set of features. OneR (G. Holmes et. al., 1996) is the 1R classifier with one parameter: the minimum bucket size for discretization. The information gain (nominal class) or variance reduction (numeric class) of each antecedent is computed, and rules are pruned using reduced-error pruning. ZeroR is even simpler: it predicts the test data's majority class (if nominal) or average value (if numeric) (Ian H. Witten and Eibe Frank, 2005). JRip (ripper algorithm for fast, effective rule induction) implements RIPPER including heuristic global optimization of the rule set (Cohen, W. W., 1995).

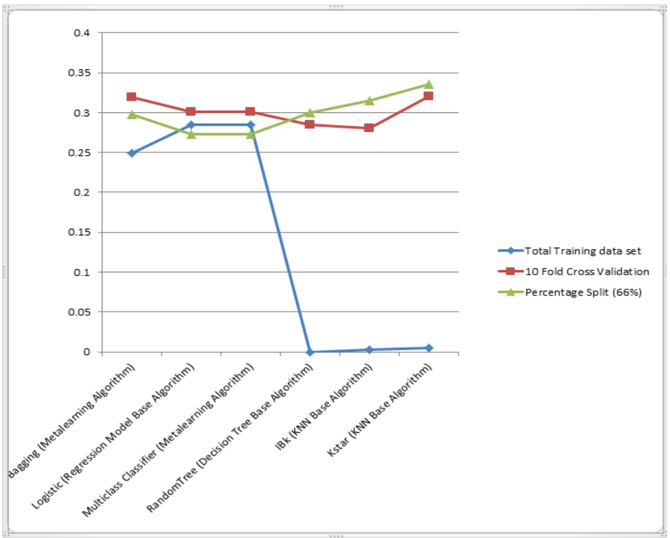


Figure 4.Graphical Representations of Error Rates of Top Algorithms in Different Cases

Functions Classifiers:

Here 3 types of function base classification algorithms are mainly described like support vector machine base (SMO), regression model base (Logistic) and neural network base (Multilayer Perceptron).

SMO implements the sequential minimal optimization algorithm for training a support vector classifier using polynomial or Gaussian kernels (Platt, J. 1998 and S. Keerthi et. al., 2001). Missing values are replaced globally, nominal attributes are transformed into binary ones, and attributes are normalized by default. Logistic is an alternative implementation for building and using a multinomial logistic regression model with a ridge estimator to guard against over fitting by penalizing large coefficients (le Cessie S. et. al., 1992). Multilayer Perceptron is a neural network that trains using back propagation. it differs from the other schemes because it has its own user interface. This network has three layers: an input layer on the left with one rectangular box for each attribute; a hidden layer next to it to which all the input nodes are connected; and an output layer at the right. The labels at the far right show the classes that the output nodes represent. Output nodes for numeric classes are automatically converted to unthresholded linear units.

Lazy Classifiers:

Lazy learners store the training instances and do no real work until classification time. IB1 is a basic instance-based learner which finds the training instance closest in Euclidean distance to the given test instance and predicts the same class as this training instance. If several instances qualify as the closest, the first one found is used. IBk is a k-nearest-neighbor classifier that uses the same distance metric. The number of training instances kept by the classifier can be restricted by setting the window size option. As new training instances are added, the oldest ones are removed to maintain the number of training instances at this size. KStar is a nearest neighbor method with a generalized distance function based on transformations (Ian H. Witten and Eibe Frank, 2005).

Miscellaneous classifiers:

VFI (voting feature intervals) constructs intervals around each class by discretizing numeric attributes and using point intervals for nominal ones, records class counts for each interval on each attribute, and classifies test instances by voting (Demiroz G. and A. Guvenir, 1997). A simple attribute weighting scheme assigns higher weight to more confident intervals, where confidence is a function of entropy. VFI is faster than Naïve Bayes but slower than hyperpipes. Neither method can handle missing values.

Metalearning Classifiers:

Metalearning algorithms take classifiers and turn them into more powerful learners. One parameter specifies the base classifier; others specify the number of iterations for schemes such as bagging and boosting. Bagging bags a classifier to reduce variance. In the case of classification, predictions are generated by averaging probability estimates, not by voting. One parameter is the size of the bags as a percentage of the training set. Another is whether to calculate the out-of-bag error, which gives the average error of the ensemble members (L. Breiman, 1996). AdaBoostM1 can be accelerated by specifying a threshold for weight pruning and resamples if the base classifier cannot handle weighted instances.

Classification Via Clustering and Classification Via Regression perform classification using a clusterer and a regression method respectively. Another metalearning algorithm is Multi Class Classifier that uses two-class classifier for multiclass datasets. It handles multiclass problems with two-class classifiers using any of these methods (Ian H. Witten et. al., 2011):

1. One versus all the rest.
2. Pairwise classification using voting to predict.
3. Exhaustive error-correcting codes.
4. Randomly selected error-correcting codes.

V. Performance Analysis

Here is seen that in training dataset 3 algorithms are parallel high in case of accuracy. Similarly in 10 fold cross validation 1 algorithm, in percentage split 3 algorithms and in average 1 algorithm carry high accuracy. Now using those algorithms Table -3 is drawn. And finally rank most five algorithms for diabetes prediction data. This work is also checked for speed and error rate. Random Tree, IBk, KStar algorithms have best accuracies for total training dataset. On the other hand Bagging, SMO algorithms are best for 10 fold cross validation and Logistic, Multiclass Classifier, Bagging algorithms are best for percentage split respectively. But Bagging is best for all cases where Logistic and Multiclass Classifier are second top ranker classification algorithms.

Table 2.Accuracies of Different Classification Algorithms in 3 Cases and their Averages [1]

Classification Algorithms	Training data set	10- fold Cross Validation	Percentage Split	Average Accuracy
NavieBayes	77.5385 %	75.3846 %	81.8182 %	78.2471%
Logistic	78.7692 %	76.9231 %	81.8182 %	79.1702%
MultilayerPerceptron	92.3077 %	73.5385 %	72.7273 %	79.5245%
SMO	78.1538 %	78.1538 %	79.0909 %	78.4662%
KStar	100 %	69.8462 %	70 %	79.9487%
AdaBoostM1	79.0769 %	75.6923 %	80.9091 %	78.5594%
Bagging	85.2308 %	77.2308 %	80 %	80.8205%
ClassificationViaClustering	69.5385 %	65.2308 %	70.9091 %	68.5595%
ClassificationViaRegression	78.1538 %	76.9231 %	79.0909 %	78.0559%
MultiClassClassifier	78.7692 %	76.9231 %	81.8182 %	79.1702%
VFI	78.4615 %	76.6154 %	79.0909 %	78.0559%
OneR	78.1538 %	78.1538 %	79.0909 %	78.4662%
ZeroR	53.8462 %	53.8462 %	53.6364 %	53.7763%
BFTree	78.1538 %	75.0769 %	78.1818 %	77.1375%
FT	85.5385 %	72.6154 %	79.0909 %	79.0816%
RandomTree	100 %	71.6923 %	70 %	80.5641%
DecisionTable	78.1538 %	76.3077 %	78.1818 %	77.5478%
J48	88.3077 %	74.7692 %	76.3636 %	79.8135%
IBk	100%	72%	68.1818 %	80.0606%
JRip	78.1538 %	76.9231 %	78.1818 %	77.7529%

[1]In table 2 bold color accuracies percentage represent highest of individual cases

Table 3.Accuracies of Top Classification Algorithms in 3 Cases and their Averages[2]

Algorithm	Training data set	10- fold Cross Validation	Percentage Split	Average Accuracy	Ranking
NavieBayes	77.5385 %	75.3846 %	81.8182 %	78.2471%	No Rank
Logistic	78.7692 %	76.9231 %	81.8182 %	79.1702%	2
Bagging	85.2308 %	77.2308 %	80 %	80.8205%	1
Multiclass Classifier	78.7692 %	76.9231 %	81.8182 %	79.1702%	2
RandomTree	100 %	71.6923 %	70 %	80.5641%	3
IBk	100%	72%	68.1818 %	80.0606%	4
SMO	78.1538 %	78.1538 %	79.0909 %	78.4662%	No Rank
KStar	100 %	69.8462 %	70 %	79.9487%	5
Average All	87.3077%	74.7692%	76.5909%	79.5560%	

[2]In table 3 bold color accuracies percentage represent highest of individual cases that is marked base on their averages through row and column respectively and finally ranked of maximum of classification algorithms base on maximum number of bold color

VI. Compare To Existing Work

Before doing this work some papers (XindongWu, et. al., 2008; Smitha T. and V. Sundaram, 2012; Trilok Chand Sharma, Manoj Jain. 2013; Pardeep Kumar et. al., 2012; Gopala et. al., 2013; Araken M Santoset. et. al., 2011; Manpreet Singh et. al., 2013, Gama J., 2004 and V. Karthikeyani et. al., 2012) have been read. Some papers discuss about only accuracies, some discuss about accuracies only for 10 fold cross validation case, another discuss about diabetics data but this paper discusses about accuracies of Diabetics patients' data perspective to Bangladesh in 3 cases (total training data set, percentage split and 10 fold cross validation) and shows top ranking algorithms for all cases shown in Fig. 2 as well as finally select top 5 classifier algorithms that are best for all cases show in Fig. 3. Fig. 4 shows error rate of top 5 ranker algorithms.

VII. Conclusions

The conclusion that can be drawn from this research is the product of polyester from PFAD has physical properties that close to the commercial polyester has a good quality of acid value and can be classified in low molecular weight of polyester which is more suitable for the application of modified polyester. The synthesis of polyester reaction is a reversible reaction in which the acquisition of the product depends on the concentration of catalyst. This paper only uses 20 classification algorithms for classify diabetes patient data perspective to Bangladesh. Lastly find top 5 algorithms for 3 cases like total training data set, percentage split and 10 fold cross validation. The most top ranker classification algorithm is Bagging (Accuracy 85.2308 %). Second top ranker classification algorithms are Logistic and Multiclass Classifier whose accuracies are 81.8182 %. The algorithms are ranked according to training data set, percentage split and 10 fold cross validation and their average accuracies using WEKA toolkit version 3.6.5.

Table 4.Confusion Matrix[3]

Classification → / Parameters ↓	Classified As Patient	Classified As Not Patient
Diabetics Patient	TP	FNP
Diabetics Not Patient	FP	TNP

[3]Hints: P=Patient, NP = Not Patient, T=Correct Classification, F=Wrong Classification

VIII. Future Work

This only discusses about accuracies of different classification algorithms using WEKA toolkit. In future we will try to create hybrid algorithm or new algorithm that will be able to provide best classification result for every case like total training data set, percentage split and 10 fold cross validation.

Acknowledgement

The authors are grateful to the participants who contributed to this research.

REFERENCES

Araken M Santos, Anne M P Canuto, Antonino Feitosa Neto. 2011. A Comparative Analysis of Classification Methods to Multi-label Tasks in Different Application Domains. International Journal of Computer Information Systems and Industrial Management Applications, 3(1):218-227.

Breiman, L. 1996. Bagging predictors. Machine Learning, 24(2):123–140.

Cohen, W. W. (1995). Fast effective rule induction. Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA. San Francisco: Morgan Kaufmann, pp. 115–123.

Demiroz, G., and A. Guvenir. 1997. Classification by voting feature intervals. Proceedings of the Ninth European Conference on Machine Learning, Prague, Berlin: Springer-Verlag, pp. 85–92.

Frawley and Piatetsky-Shapiro. 1996. Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A.

Gama, J. 2004. Functional trees, Machine Learning, 55(3):219–250.

Gopala Krishna Murthy Nookala, Bharath Kumar Pottumuthu, Nagaraju Orsu, Suresh B. Mudunuri. 2013. Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification. International Journal of Advanced Research in Artificial Intelligence, 2(5):49-55.

Hian Chye Koh and Gerald Tan. 2011. Data Mining Applications in Healthcare. Journal of Healthcare Information Management, 19 (2): 64-72.

Holmes, G., and C. G. Nevill-Manning. 1995. Feature selection via the discovery of simple classification rules. Proceedings of the International Symposium on Intelligent Data Analysis, pp:75–79.

Ian H. Witten and Eibe Frank. 2005. Data Mining Practical Machine Learning Tools and Techniques. 2nd Edition, Series Editor: Jim Gray, Microsoft Research, Elsevier.

Ian H. Witten, Eibe Frank and Mark A. Hall. 2011. Data Mining Practical Machine Learning Tools and Techniques. 3rd Edition, Elsevier.

K. Ahmed, T. Jesmin, U. Fatima, Md. M., Abdullah-al-E., Md. Z. Rahman. 2012. Intelligent and Effective Diabetes Prediction System Using Data Mining Approach. ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY, 5(1):215-21.

Keerthi, S. S., S. K. Shevade, C. Bhattacharyya, and K. R. K.Murthy. 2001. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation, 1 (3): 637–649.

le Cessie, S., and J. C. van Houwelingen. 1992. Ridge estimators in logistic regression. Applied Statistics, 41(1):191–201.

Manpreet Singh, Sonam Sharma, Avinash Kaur. 2013. Performance Analysis of Decision Trees. International Journal of Computer Applications, 71(19):10-14.

Pardeep Kumar, Nitin, Vivek Kumar Sehgal, Durg Singh Chauhan. 2012. A BENCHMARK TO SELECT DATA MINING BASED CLASSIFICATION ALGORITHMS FOR BUSINESS INTELLIGENCE AND DECISION SUPPORT SYSTEMS. International Journal of Data Mining & Knowledge Management Process (IJDKP), 2(5):25-42.

Trilok Chand Sharma, Manoj Jain. 2013. WEKA Approach for Comparative Study of Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering, 2 (4):1931-1995.

Unwin N, Whiting D, Gan D, Jacqmain O, Ghyoot G, editors. 2009. IDF Diabetes Atlas, 4th ed. Brussels: International Diabetes Federation.

Platt, J. 1998. Fast training of support vector machines using sequential minimal optimization. Advances inkernel methods: Support vector learning. Cambridge, MA: MIT Press.

V.Karthikeyani, I.Parvin Begum, K.Tajudin and I.Shahina Begam. 2012. Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction. International Journal of Computer Applications, 60(12):26-31.

XindongWu, et. al. 2008. Top 10 algorithms in data mining. Knowledge Information System, 14(1):1–37.

Smitha T., V. Sundaram. 2012. COMPARATIVE STUDY OF DATA MINING ALGORITHMS FOR HIGH DIMENSIONAL DATA ANALYSIS. International Journal of Advances in Engineering & Technology, 4(2):173-178.