N E W S
OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
SERIES OF GEOLOGY AND TECHNICAL SCIENCES
ISSN 2224-5278
Volume 6, Number 426 (2017), 231 – 236

UDC 004.855 - 004.891.3

Sh. Shamiluulu[1], B. Y. Amirgaliyev[2], L. Cherikbayeva[3]

[1]Department of Computer Science, Suleyman Demirel University, Kaskelen, Kazakhstan,
[2]International University of Information Technologies, Almaty, Kazakhstan,
[3]Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: shahriar.shamiluulu@sdu.edu.kz,amirgaliyev@gmail.com,lailash01@gmail.com

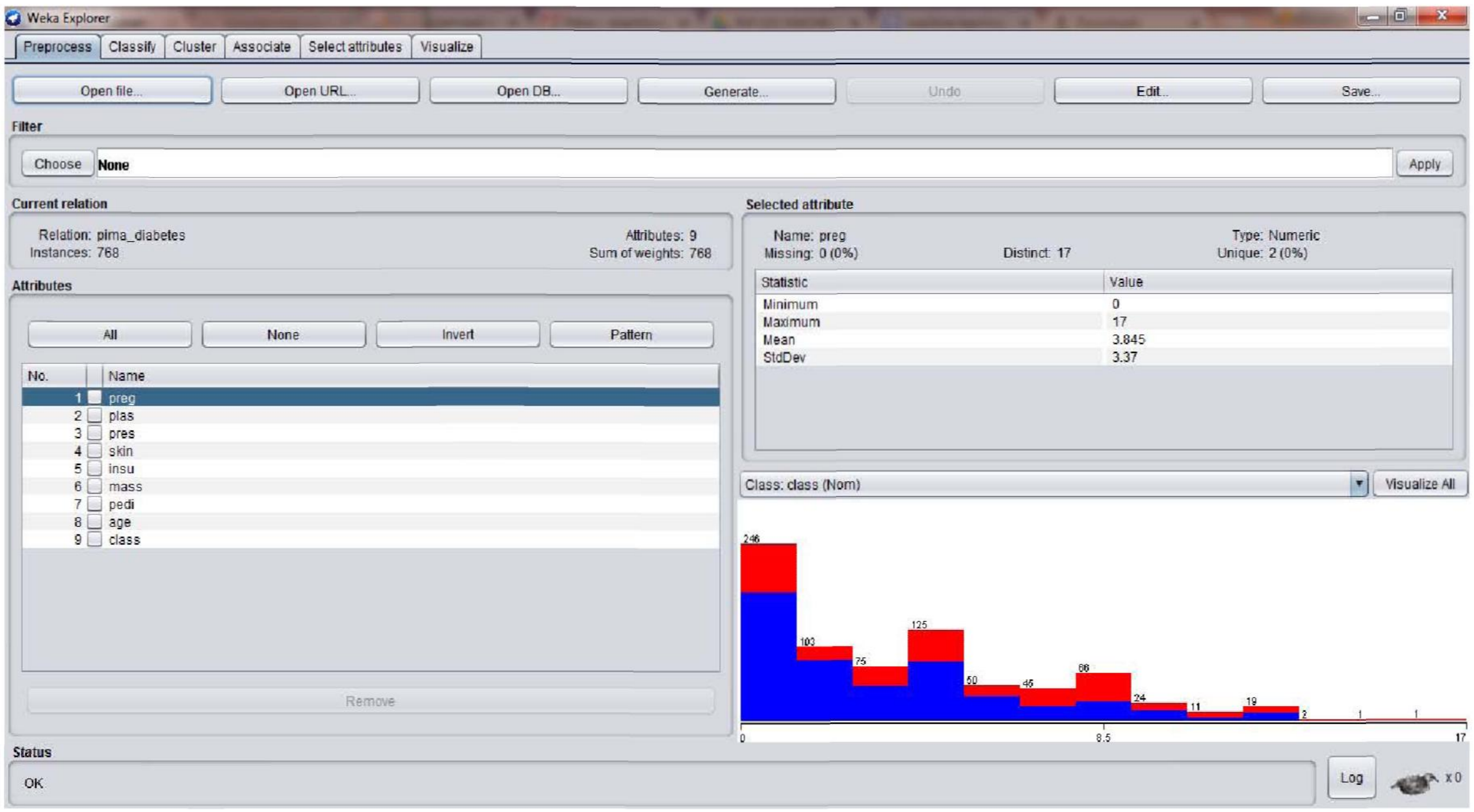
CRITICAL ANALYSIS OF SCIKIT-LEARN ML FRAMEWORK AND WEKA ML TOOLBOX OVER DIABETES PATIENTS MEDICAL DATA

Abstract. In this research study two ml tools tested i.e., scikit-learn and Waikato Environment for Knowledge Analysis, over three supervised machine learning algorithms. The comparative analysis has been performed on tree supervised machine learning algorithms i.e., decision trees, logistic regression and multi-layer perceptron neural network on medical data for patients with two types of diabetes disorders. Diabetes-Mellitus refers to the metabolic disorder that happens from malfunction in insulin secretion and action. It is characterized by hyperglycemia.The diagnosis of diabetes is very important now days using various types of techniques and thus selection of framework is important.The dataset has been obtained from UCI machine learning repository for Pima Indian Diabetes patients. During the research the comparative analysis studies have been performed which revealed that less complex algorithms can be used for disease diagnosis and possess better performance when properly configured.

Keywords: diabetes mellitus, supervised learning, performance analysis, clinical decision support systems.

Introduction. There are many machine learning frameworks and tools available for public use. This research study tries to thoroughly analysis two tools and reveal which all trade-offs.In this study, the comparative performance analysis of three supervised machine learning algorithms is studied i.e, decision trees, logistic regression and artificial neural network and advantages of using Waikato Environment for Knowledge Analysis (WEKA) machine learning toolbox [3] is shown in Figure 1. Using WEKA toolbox is relatively easy which has many build-in options as compare to scikit-learn framework where in order to use algorithms users need to write code as shown in Figure 2.

The comparative analysis is going to be performed on medical dataset. Currently the computer aided diagnosis plays an important role in the medical field. It has been shown that the benefits of introducing machine learning into medical analysis are to increase the diagnostic accuracy, to reduce costs and to reduce human resources. The algorithms are tested over the Pima Indian diabetes dataset. Pima Indian Diabetes database had been examined with several different machine learning methods in the past [5-9]. Diabetes-Mellitus refers to the metabolic disorder that happens from malfunction in insulin secretion and action. It is characterized by hyperglycemia. There are two types of diabetes disorder but generally the symptomatic and lab results are same. The diagnosis of diabetes is very important now days using various types of techniques.



Filter: None

Relation: pima_diabetes
Instances: 768
Attributes: 9
Sum of weights: 768

Name: preg
Missing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)

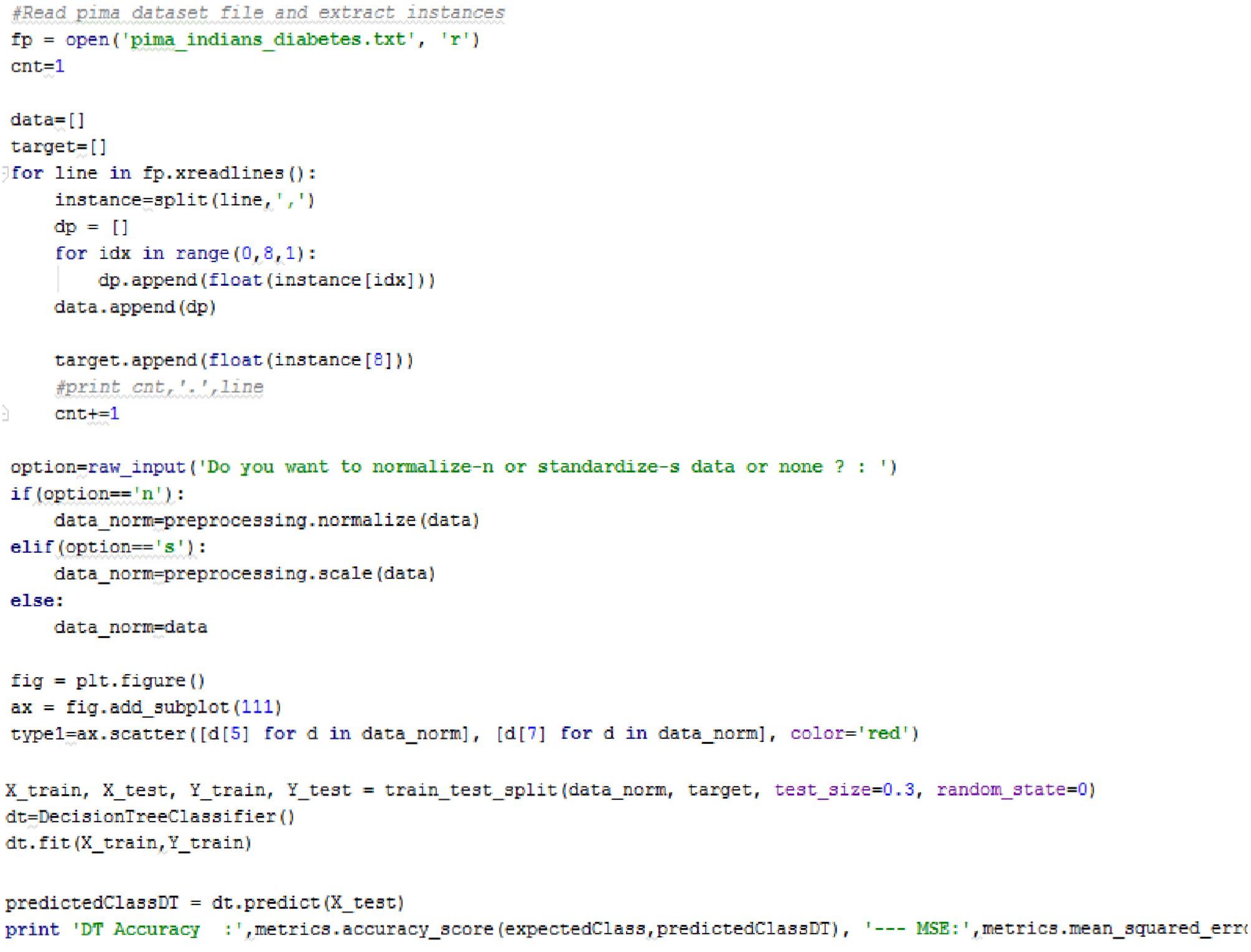
Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

No.	Name
1	☐ preg
2	☐ plas
3	☐ pres
4	☐ skin
5	☐ insu
6	☐ mass
7	☐ pedi
8	☐ age
9	☐ class

Class: class (Nom)

Status: OK

Figure 1 – Dataset visualization of WEKA toolbox



```
#Read pima dataset file and extract instances
fp = open('pima_indians_diabetes.txt', 'r')
cnt=1

data=[]
target=[]
for line in fp.xreadlines():
    instance=split(line,',')
    dp = []
    for idx in range(0,8,1):
        dp.append(float(instance[idx]))
    data.append(dp)

    target.append(float(instance[8]))
    #print cnt,'.',line
    cnt+=1

option=raw_input('Do you want to normalize-n or standardize-s data or none ? : ')
if(option=='n'):
    data_norm=preprocessing.normalize(data)
elif(option=='s'):
    data_norm=preprocessing.scale(data)
else:
    data_norm=data

fig = plt.figure()
ax = fig.add_subplot(111)
type1=ax.scatter([d[5] for d in data_norm], [d[7] for d in data_norm], color='red')

X_train, X_test, Y_train, Y_test = train_test_split(data_norm, target, test_size=0.3, random_state=0)
dt=DecisionTreeClassifier()
dt.fit(X_train,Y_train)

predictedClassDT = dt.predict(X_test)
print 'DT Accuracy  :',metrics.accuracy_score(expectedClass,predictedClassDT), '--- MSE:',metrics.mean_squared_err
```

Figure 2 – Initialization and simulation of algorithms code on scikit-learn framework

Materials and methods

A. Data Collection

The data set was obtained from the UCI Repository of Machine Learning Databases [3]. The data set was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. The patients in the Pima-Indian dataset are women at least 21 years old and living near Phoenix, Arizona, USA. The dichotomousoutcomeattribute takes the values '0' or '1', where '1' means a positive test for diabetes and '0' is a negative test for diabetes. There are 500 (65.1%) cases in class '0' and 268 (34.9%) cases in class '1'. The dataset contains eight clinical findings which are:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index
7. Diabetes pedigree function
8. Age (years)

B. Decision Trees

A decision tree (DT) is a graph that uses a branching method to illustrate every possible outcome of a decision for particular. The goal in building a tree is to identify a best splitting attribute which is being found by Entropy and Information Gain. The detailed theoretical background regarding decision trees can be found here [2].

C. Logistic Regression

The logistic regression (LRM) has wide range of implications in medical research field. The LRM model is used for the classification of the attributes, which might help to classify the outcome. The distinctive feature of the model is that the outcome variable is dichotomous. The result is not bounded to a linear form. As a result, the created model can be used to classify a newly provided data via placing them in a model for the probability P, the detailed information is provided in [1].

D. Multi-Layer Perceptron Neural Networks

Multi-layer perceptron neural networks (MLP)are processing devices, whichclosely resemble a modelof the neuronal structure of the mammalian cerebral cortex. LargeMLPs might have hundreds or thousands of processor units, whereas a mammalian brain has billions of neurons with a corresponding increase in magnitude of their overall interaction and emergent behavior. Generally, the neural network has three components or layers. The first layer is called input layer, through which it gets data inside network on our case disease related attributes. The second layer is called hidden where all operations performed. The last layer called out where network make final decision regarding patient's condition. The detailed information regarding neural networks provided in [1, 2].

E. Simulated Program

The Waikato Environment for Knowledge Analysis (WEKA), is one of the best tools in teaching machine learning without going into details first. The tool is basedon Java platform that contains a large number of algorithms for data preprocessing, feature selection, classification, clustering, and finding the associative rule [4]. WEKA uses a common data representation format, making comparisons easy. It has three operation modes i.e., GUI, Command Line, and Java API.

F. Performance Measures

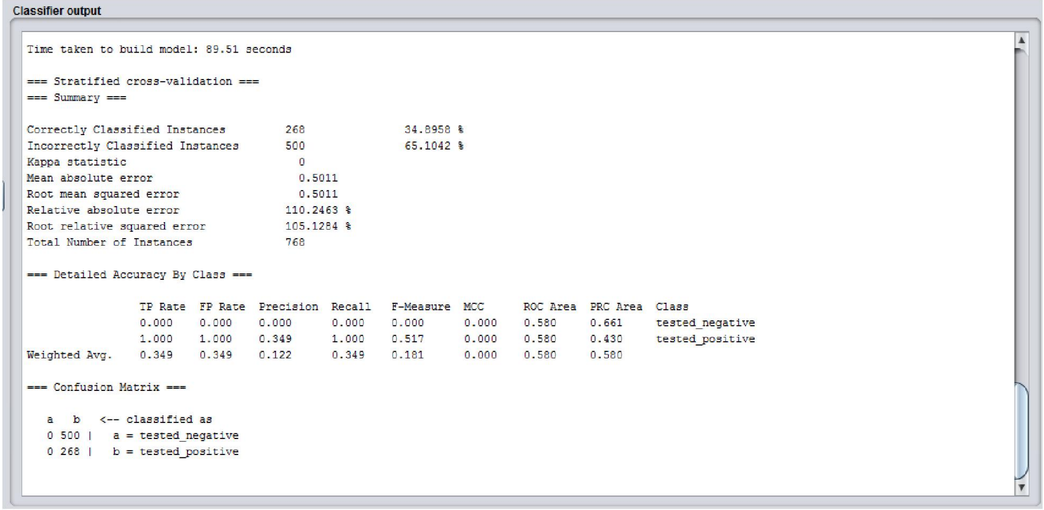
Evaluation of the classifier to measure the quality is commonly evaluated based on the data in the confusion matrix. Several standard measures have been defined for correct and incorrect classification results of the matrix. The most common practical measure to evaluate the performance is accuracy, which is defined as the proportion of the total number of instances that were classified correctly.

Recall is the mean proportion of actual positives which are correctly identified. *Precision* is the mean proportion of positives which are relevant.*F-measure* is a harmonic mean of recall and precision. *TP rate* is a measure which shows the matching states of particular instances. *FP rate* is a measure that shows the mismatching states of particular instances.

These performance metrics are calculated according to the data in the confusion matrix which are obtained by the WEKA tool.

Simulation results

In this study, the Pima dataset of patients with diabetes disorder, which containing 9 original features by using three machine learning methods i.e., DT (C4.5), LRM and MLP used for classification. The performance metrics like accuracy, recall, precision and f-measure along with error metrics forall features has been performed using 10-fold cross-validation see Figure 3. The simulations were performed by using WEKA3.8 machine learning tool. In scikit-learn users need to import special libraries to show the output results.



```
Time taken to build model: 89.51 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         268               34.8958 %
Incorrectly Classified Instances       500               65.1042 %
Kappa statistic                          0
Mean absolute error                      0.5011
Root mean squared error                  0.5011
Relative absolute error                110.2463 %
Root relative squared error            105.1284 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    0.000      0.000    0.000      0.000    0.580     0.661     tested_negative
                 1.000    1.000    0.349      1.000    0.517      0.000    0.580     0.430     tested_positive
Weighted Avg.    0.349    0.349    0.122      0.349    0.181      0.000    0.580     0.580

=== Confusion Matrix ===

   a   b   <-- classified as
   0 500 |   a = tested_negative
   0 268 |   b = tested_positive
```

Figure 3 – Simulation results in WEKA toolbox

Accuracy metrics of ML algorithms

	No of correct ins. with %	No of incorrect ins. with %	Build time
DT	567 - (74%)	201 - (26%)	0.12 sec
LRM	593 - (77%)	175 - (23%)	0.15 sec
MLP-NN	583 - (76%)	185 - (24%)	1.45 sec

According to the results provided in Table, the LRM model outperforms remaining methods with the overall accuracy of 77% even though the MLP is considered one of top classification methods it came the second one in this race. During the analysis we have identified that this problem was due to overfitting issue.

The Figure 4 contains the five different performance measure results for three machine learning algorithms. Based on the results the LRM method outperforms MLP by 1% and DT by 3% on overall. Even though the MLP considered the complex method for classification and prediction the complex nature of it, which contained one hidden layer with 5 nodes fall into problem of overfitting. In the literature survey we have find out that for three algorithms the accuracy metric was varying from partition to partition. For example, when we see the accuracy of C4.5 model, 77.08% in case of 75-25% training-testing partitions, 76.72% in case of 85-15% training-testing partitions and 75.32% in case of 90-10% training-testing partitions.

The Figure 5 shows three error metric results i.e., Kappa statistics, Mean Absolute Error (MAE) and Root mean square error (RMSE) for simulated machine learning algorithms. The Kappa statistics measures the agreement of prediction with the true class. The value which is bigger than zero indicates

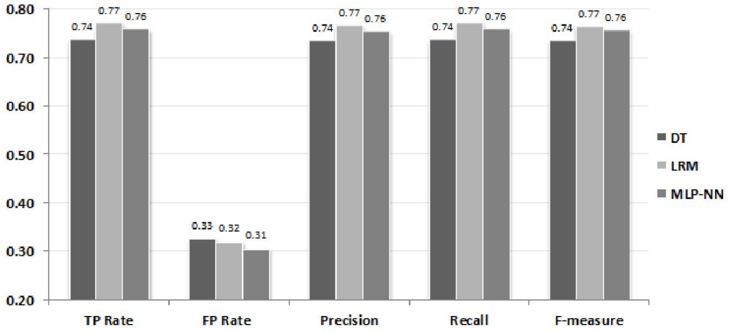


Figure 4 – Performance measure metrics of ML algorithms

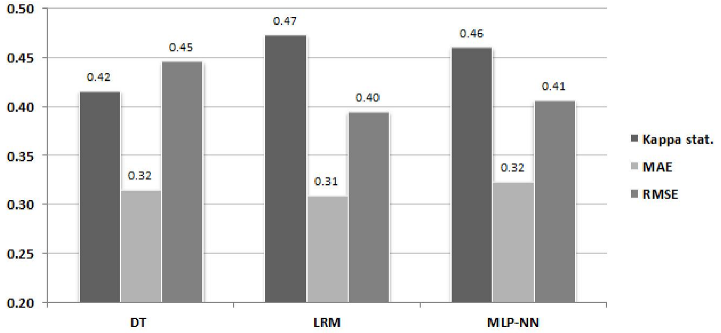


Figure 5 – Error metrics of ML algorithms

that algorithm is not performed based on chance but rather taking more logical approach. The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction and the value close to zero is better. It measures accuracy for continuous variables. The RMSE measures the average magnitude of the error and the value close to zero is better. Based on simulation results shown in Figure 2, the LRM outperforms all other methods. The scikit-learn framework is more flexible and complete as compare to WEKA but using it is more difficult.

Conclusion. In this research study two ml tools tested over two ml frameworks by using three-supervised machine learning algorithms. The algorithms applied overthe Pima Indians Diabetes (PID) medical dataset. The scikit-learn framework is more flexible and complete as compare to WEKA but using it is more difficult. The performance of LRM was the best for all performance and error metrics. The LRM method outperforms MLP by 1% and DT by 3% on overall. MLP is considered one of top classification methods it came the second one. During the analysis we have identified that this problem was due to overfitting issue. As a result,shows that, LRM methods can be a good and practical choice to classify a medical data. WEKA toolbox was more user-friendly and easy to use but less flexible.

REFERENCES

[1] Norvig P, Russell S. Artificial Intelligence: A Modern Approach, Prentice Hall. 2002.

[2] Dunham MH. Data mining: Introductory and advanced topics, Pearson Education, 2006.

[3] UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science. Available: https://archive.ics.uci.edu/ml/datasets/Diabetes (Accessed: 7 Sept. 2016).

[4] HallM., FrankE., and HolmesG., B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, 2009. - Vol. 11, pp. 10-18.

[5] Raj Anand, Vishnu Pratap Singh Kirar, Kavita Burse, " K-Fold Cross Validation and Classification Accuracy of PIMA Indian Diabetes Data Set Using Higher Order Neural Network and PCA ", IJSCE, 2013.-Vol. 2, Issue-6, pp. 436-438.

[6] Y. Angeline Christobel, P.Sivaprakasam, " A New Classwise k Nearest Neighbor (CKNN) Method for the Classification of Diabetes Dataset", IJEAT, 2013. – Vol. 2, Issue-3, pp. 396-400.

[7] Kumari V. Anuja, Chitra R. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications,2013. - Vol. 3, pp. 1797-1801.

[8] Carpenter G.A., Markuzon N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", Neural Networks, 1998. - 323-336 pp.

[9] Deng, D., Kasabov, N., "On-line pattern analysis by evolving self-organizing maps", Proc. of the 5th Biannual Conference on Aritificial Neural Networks and Expert Systems (ANNES), Dunedin, November, 2001. pp. 46-51.

[10] Farahmandian M., Lotfi Y., Maleki I. Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study. MAGNT Research Report, 2015. - Vol. 3, pp. 989-997.

Ш. Шамильуулу[1], Б. Е. Амиргалиев[2], Л. Черикбаева[3]

[1]Кафедра компьютерных наук, университет Сулеймана Демиреля, Каскелен, Казахстан,
[2]Международный университет информационных технологий (МУИТ), Алматы, Казахстан,
[3]Казахский национальный университет им. аль-Фараби, Алматы, Казахстан

КРИТИЧЕСКИЙ АНАЛИЗ МЕДИЦИНСКИХ ДАННЫХ ПАЦИЕНТОВ С ДИАБЕТОМ ПРИ ПОМОЩИ ФРЕЙМВОРК ASCIKIT-LEARNML И ПРОГРАММЫ WEKA

Аннотация. В этой работе для анализа знаний рассматриваются два инструмента: пакет Scikit-обучение и Waikato, а также три контролируемые алгоритмы машинного обучения. Сравнительный анализ был проведен на дереве с использованием алгоритмов машинного обучения, т.е., анализ был проведен на дереве решений, с использованием логистической регрессии и многослойной нейронной сети на медицинских данных для пациентов с двумя типами заболеваний сахарного диабета. В настоящее время диагностику сахарного диабета можно диагностировать с помощью различных типов технических средств, однако здесь очень важен выбор метода. Для эксперимента набор данных был получен из репозитория машинного обучения UCI для Пима Индийских больных сахарным диабетом. В ходе исследования по сравнительному анализу были проведены эксперименты, которые показали, что менее сложные алгоритмы успешно могут быть использованы для диагностики заболеваний и имеют более высокую производительность при правильной настройке.

Ключевые слова: сахарный диабет, контролируемоеобучение, анализ эффективности, клинические системы поддержки принятия решений.

Ш. Шамильуулу[1], Б. Е. Амиргалиев[2], Л. Черикбаева[3]

[1]Компьютерлік ғылымдар кафедрасы, Сулейман Демирель университеті, Каскелен, Қазақстан,
[2]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
[3]Аль-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан

SCIKIT-LEARNML ФРЕЙМВОРК ЖӘНЕ WEKA БАҒДАРЛАМАСЫ АРҚЫЛЫ ДИАБЕТ АУРУЛАРЫНЫҢ МЕДИЦИНАЛЫҚ ДЕРЕКТЕРІН КРИТИКАЛЫҚ ТАЛДАУ

Аннотация. Мақалада білімдерді талдау үшін Scikit-оқыту және Waikato атты екі құрал және бақылаулы машиналық оқытудың үш алгоритмі қарастырылады. Салыстырмалы талдау мащиналық оқытуды қолдана отырып шешімдер ағашында орындалды, яғни талдау қант диабетініңекі түрлі ауыруы үшін логистикалық регрессия және көпқабатты нейрондық желі қолдану арқылы шешімдер ағашында орындалды. Қазіргі уақытта қант диабетін әртүрлі техникалық құралдар арқылы анықтауға болады, бірақ тиімді әдісті таңдау маңызды мәселе болып қала береді. Сандық тәжірибелер үшін деректер жиыны қант диабеті ауыруларының Үнді Пима UCI машиналық оқыту репозиториясынан алынды. Салыстырмалы талдау зерттеулері барысында сандық тәжірибелер жүргізіліп, нәтижесінде аса күрделі емес алгоритмдердің жұмыс параметрлері дұрыс таңдалса, олардың жоғары өнімділік нәтижелерін беретіндігі көрсетілді.

Түйін сөздер: қант диабеті, бақылаулы оқыту, тиімділікті талдау, шешімдер қабылдауды қолдаудың клиникалық жүйелері.