International Journal of Engineering & Technology, 7 (4.5) (2018) 248-253

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Evaluating the Performance of Supervised Classification Models: Decision Tree and Naïve Bayes Using KNIME

Syed Muzamil Basha[1]*, Dharmendra Singh Rajput[2*], Ravi Kumar Poluru[3], S. Bharath Bhushan[4], Shaik Abdul Khalandar Basha[5]

[1,3,5]Research Scholar, VIT University, Vellore. 632014,
[2] Associate Professor, VIT University, Vellore. 632014,
[4]Assistant Professor, CSSE Department, Sree Vidyanikethan Engineering College (SVEC), Tirupati, India,
**Corresponding author E-mail: muza.basha@gmail.com*

Abstract

The classification task is to predict the value of the target variable from the values of the input variables. If a target is provided as part of the dataset, then classification is a supervised task. It is important to analysis the performance of supervised classification models before using them in classification task. In our research we would like to propose a novel way to evaluated the performance of supervised classification models like Decision Tree and Naïve Bayes using KNIME Analytics platform. Experiments are conducted on Multi variant dataset consisting 58000 instances, 9 columns associated specially for classification, collected from UCI Machine learning repositories (http://archive.ics.uci.edu/ml/datasets/statlog+(shuttle)) and compared the performance of both the models in terms of Classification Accuracy (CA) and Error Rate. Finally, validated both the models using Metric precision, recall and F-measure. In our finding, we found that Decision tree acquires CA (99.465%) where as Naïve Bayes attain CA (90.358%). The F-measure of Decision tree is 0.984, whereas Naïve Bayes acquire 0.7045.

Keywords: *Classification Accuracy; Decision Tree; Error Rate; F-measure; KNIME Analytics platform; Naïve Bayes; Precision; Recall.*

1. Introduction

Classification is one type of machine learning problems. In which, the input data is presented to the machine learning model and the task is to predict the target corresponding to the input data. The target is a categorical variable, so the classification task is to predict the category or label of the target given the input data. Each row has specific values for the input variables and a corresponding value for the target variable. The classification task is to predict the value of the target variable from the values of the input variables. If a target is provided as part of the dataset, then classification is a supervised task. In our research we are likely to build and apply a supervised classification models namely Decision tree and Naïve Bayes. The goal in building a classifier model is to have the model perform well on training, as well as test data. There are many algorithms to build a classification model. A Decision tree is a classification model that uses a treelike structure to represent multiple decision paths. Traversing each path leads to a different way to classify an input sample. A naive Bayes model uses a probabilistic approach to classification. Baye's Theorem is used to capture the relationship between the input data and the output class. In [1] the author Guarín et al. 2015 applied an educational data mining approach to model the loss of academic status at the Universidad Nacional de Colombia. Defined two data mining models to analyze the academic and nonacademic data, this models uses two classification techniques, Naïve Bayes and a decision tree classifier, in order to attain a better understanding of the loss of academic status. This work motivated us to select the Decision tree and Naïve Bayes classification algorithms among all other algorithms. So, we made an attempt to make the readers to understand the steps, we have carried in our experiments using KNIME Analytics platform software.

A machine learning model can be represented as mathematical model to determine the relationship between its inputs and outputs. The parameters of a machine learning model are adjusted or estimated from the data using a learning algorithm. In general, building a classification model involves two phases. The first is the training phase, in which the model is constructed and its parameters adjusted using as what referred to as training data helps in creating a model. The second is the testing phase, where the learned model is applied to new data. The model is then evaluated on how it performs on the test data. In [2] the author Wei Chen et al. 2017 presented study on use of three state-of-the-art data mining techniques, namely, logistic model tree (LMT), random forest (RF), and classification and regression tree (CART) models. The Models are finally legalized and compared using receiver operating characteristics, and predictive accuracy methods. In which the RF model exhibits the highest predictive capability with a success rate of 0.837 and a prediction rate of 0.781 compared with the LMT and CART models.

Our paper is organized as follows: In Literature review section we attempts to figure out the start of art in the area of classification and its applications area, where as in Methodology section we made attempt to understand the workflow applications built on KNIME Analytics platform. In Evaluation of Model section, we discuss different ways to evaluate a supervised classification model and associated metric. The result obtained in our experiments are plotted and the explanation required to interpret them is provided in the discussion section. Finally in conclusion, we concluded

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ed our finding and gave a future direction to work with other traditional classification algorithms on different datasets.

2. Literature Review

Automatic text classification techniques are useful for classifying plaintext in medical documents. which automatically predict the cause of death from free text forensic autopsy reports by comparing various schemes for feature extraction, term weighing or feature value representation, text classification, and feature reduction. In [10] the author Ghulam Mujtaba et al. 2017 found that that unigram features obtained the highest performance compared to bigram, trigram, and hybrid-gram features. Furthermore, used text classification algorithms, support vector machine classifier outperforms random forest, Naive Bayes, k-nearest neighbor, decision tree, and ensemble-voted classifier. Supervised text classification methods are efficient when they can learn with reasonably sized labeled sets. These methods are based on comparing distributions between labeled and unlabeled instances, therefore it is important to focus on the representation and its discrimination abilities. In [11] the author Miha Pavlinek et al. 2017 presented the ST LDA method for text classification in a semi-supervised manner with representations based on topic models. This method comprises a semi-supervised text classification algorithm based on self-training and a model, which determines parameter settings for any new document collection. Conducted experiments on 11 very small initial labeled sets sampled from six publicly available document collections. ST LDA method proved to be a competitive classification method for different text collections when only a small set of labeled instances is available. To solve the issue in label propagation (LP) approaches, estimation of unknown labels of points from the original input space directly, causes unfavorable mixed signs that decrease the performance of both transductive models. In [3] the author Zhao Zhang et al. 2017 proposed a Projective Label Propagation (ProjLP) framework by label embedding, which can deliver more discriminating "deep" labels of samples to enhance representation and classification. where ProjLP delivering a linear neighborhood preserving projection classifier, by embedding deep label of each new data directly on classifier. Recently, the rapid development of electronic medical records (EMR) provides the opportunity to utilize the potential of EMR to improve the performance of Major adverse cardiac events (MACE) prediction. In [4] Zhengxing Huang et al. 2017 made a study on MACE of acute coronary syndrome (ACS), presented a new data-mining based approach for MACE prediction from a large volume of EMR data and integrated the resampling strategy into a boosting framework. Effectiveness is validated on a clinical dataset containing 2930 ACS patient samples with 268 feature types. The performance of these approach for predicting MACE remains robust and reaches 0.672 in terms of AUC. On average, this approach improves the performance of MACE prediction by 4.8%, 4.5%, 8.6% and 4.8% over the standard SVM, Adaboost, SMOTE, and the conventional GRACE risk scoring system for MACE prediction. Artificial intelligence algorithms are being applied integrally for prediction, classification or optimization of buildings energy consumption. Hybrid objective function development for energy optimization problems including qualitative and quantitative datasets in their constructs. To tackle with this issues. In [5] Saeed Banihashemi et al. 2017, employed Artificial Neural Network as a prediction and Decision Tree as a classification algorithm via cross-training ensemble equation to create the hybrid function and the model. Where as in [8] Zeyu Wang et al. 2017 conducted an in-depth review of single AI-based methods such as multiple linear regression, artificial neural networks, and support vector regression, and ensemble prediction method that, by combining multiple single AI-based prediction models improves the prediction accuracy manifold. Earthquake Early Warning System (EEWS) used to manage the critical lifeline infrastructure and essential facilities through which we can save lives. In [6] Mohammad Hossein Rafiei et al. 2017, presented a novel solution to the complex problem of earthquake prediction through adroit integration of a machine learning classification algorithm and the robust neural dynamics optimization algorithm of Adeli and Park. In software engineering, Software defect prediction (SDP) is an important task. In Which, estimating the number of defects remaining in software systems and discovering defect associations, classifying the defect-proneness of software modules plays an important role in software defect prediction. Several machine-learning methods have been applied to handle the defect-proneness of software modules as a classification problem. This type of "yes" or "no" decision is an important drawback in the decision-making process and if not precise may lead to misclassifications. To Address the issue of SDP problems are usually characterized as imbalanced learning problems, In [7] the author Diego P.P. Mesquita et al. 2016 developed a SDP method called rejoELM and its variant, IrejoELM. Both methods were built upon the weighted extreme learning machine (ELM) with reject option that makes it possible postpone the final decision of non-classified modules, the rejected ones, to another moment. IrejoELM outperforms all other methods when the F-measure is used as a performance metric. Whereas, In [13] the author Goran Mauša et al. 2017 also presented a promising ensemble strategy based on a simple convex hull approach and compared the performance of the operators for software defect prediction datasets with varying levels of data imbalance. Nasopharyngeal Carcinoma (NPC) is the most famous type of tumor in the neck and started in the nasopharynx, the area at the top of the pharynx or "throat", in which the participation of the relevant nose and tube sound including all upper respiratory tract. To understand the contextual usage of NPC. The author, In [9] Mazin Abed Mohammed et al. 2017 made an review on NPC Diagnosis. The intermittent and fluctuation of wind power has a harmful effect on power grid. To direct system operators to mitigate the harm, a combined multivariate model need to be proposed to improve wind power prediction accuracy. In [12] the author Tinghui Ouyang et al. 2017 proposed a model, In which, valid meteorological variables are used for prediction are selected by Granger causality testing approach, and reconstructed in homeomorphic phase spaces. Data mining algorithms are trained for selecting the model with high accuracy and performance of the proposed model is validated better using error metrics. In [14] the author make use of weighted fuzzy logic in initializing the exact weights to train the data in extracting sentiments from the labeled tweets. where as in [15] the author considered Time series dataset and measure the performance of predictive models. In [16] the author perform analysis on PIMA diabetes dataset and predicted the levels of diabetes based on insulin feature. where as in [17] the author used gradient ascent algorithm in finding out the exact weights of the terms used in determining the sentiment of tweet and used Boosting approach to improve the accuracy of linear classifier. In [18-20] the author explained about the clustering techniques which suits in various applications such as IoT health care and in machine learning. In [21] the author focused on clustering using K mean++ on smart card data and achieved Travel Pattern.

3. Methodology

An algorithm for constructing a decision tree model is referred to as an induction algorithm, which uses Greedy algorithms solve a subset of the problem at a time, and as a necessary approach when solving the entire problem is not feasible. This is the case with decision trees. It is not feasible to determine the best tree given a data set, so the tree has to be built in piecemeal fashion by determining the best way to split the current node at each step.

A common impurity measure used for determining the best split is the Gini Index. The lower the Gini Index the higher the purity of the split. So the decision tree will select the split that minimizes the Gini Index. Besides the Gini Index, other impurity measures

include entropy, or information gain, and misclassification rate. To evaluate the Gini Impurity for a set of items with *C* Classes, consider $I \in \{1,2,..., C\}$, and p_i be the ratio of labeled items with the class *I* in the dataset as in equation (1).

Induction Algorithm
Step 1: Start
Step 2: Starting with all samples at a single node, the root node.
Step 3: Partition the samples into subsets as pure as possible based in the input variables
Step 4: Repeat Step 3 until some stopping criterion is satisfied
Step 5: Stop

$$I_G(p) = \sum_{i=1}^{C} p_i(1 - p_i) \quad (1)$$

Different ways to stop expanding a node through induction algorithm are:

1. The algorithm can stop expanding a node when the number of samples in the node falls below a certain minimum value. At this point the number of samples is too small to make much difference in the classification results with the further splitting.
2. The induction algorithm can stop expanding a node when the improvement in impurity measure is too small to make much of a difference in classification results.
3. The algorithm can stop expanding a node when the maximum tree depth is reached. This is to control the complexity of the resulting tree.

The greedy approach used by tree induction algorithm determines the best way to split the portion of the data at a node but does not guarantee the best solution overall for the entire data set. In contrast, A Naive Bayes classification model uses a probabilistic approach to classification. In which, the relationships between the input features and the class is expressed as probabilities. So given the input features for a sample, the probability for each class is estimated. The class with the highest probability then, determines the label for the sample. Naive Bayes classifier uses Bayes' theorem.

$$p(A|B) = \frac{p(B|A) \times p(C)}{p(B)} \quad (2)$$

Naive Bayes assumes that the input features are statistically independent of one another. So for classification we want to calculate the posterior probability P(A | B) for each class C. From Bayes' theorem P(C | X) is related to the P(X | C) P(C And probability of X.

Using the information provided in the Table 2, the steps carried out in our experiments are interpreted in good manner.

Table 1: Small Scale Statistics analysis on Statlog Dataset.

	Min	Max	Mean	SD	Varience	Skeness	Kurtosis	overall
Col0	27	126	48.23829	12.23808	149.7706	2.180838	6.508792	2797821
Col1	-4821	5075	-0.01945	77.95804	6077.455	6.438146	2647.317	-1128
Col2	21	149	85.34912	8.902769	79.25929	1.096129	0.543597	4950249
Col3	-3939	3830	0.259672	36.52152	1333.821	31.68785	7698.225	15061
Col4	-188	436	34.54986	21.66014	469.1616	-1.16243	8.547202	2003892
Col5	-26739	15164	1.60819	217.5977	47348.75	-21.8618	5979.234	93275
Col6	-48	105	37.09231	13.11143	171.9095	-0.3684	1.621771	2151354
Col7	-353	270	50.88455	21.41805	458.7329	1.066204	8.416278	2951304
Col8	-356	266	13.93241	25.61402	656.0779	2.243244	8.438484	808080

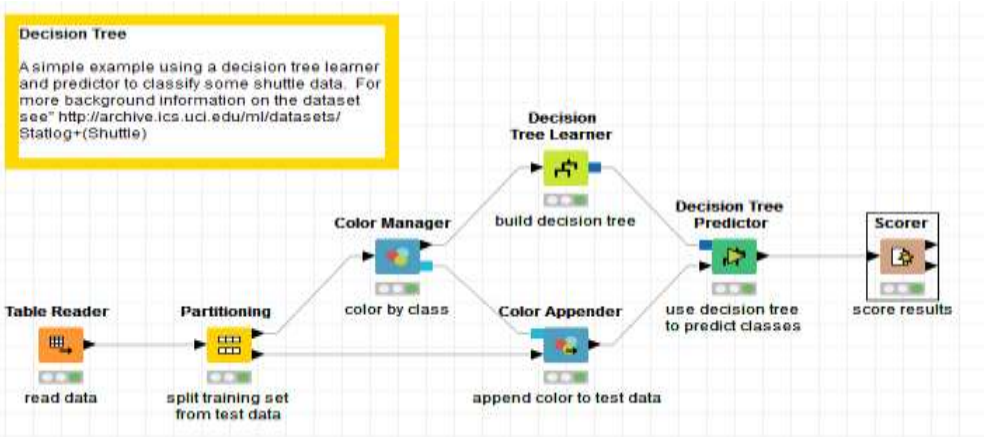


Fig. 1: Work flow of Decision Tree Model in KNIME application.

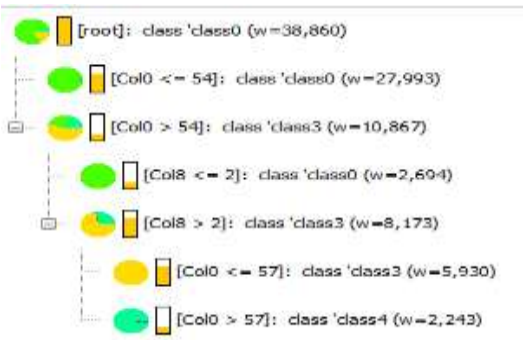


Fig. 2: Detailed View on Decision Tree constructed for prediction.

In Figure 2, the abstract view of decision tree constructed in our experiments are plotted. Similarly, the Figure 3 can be best interpreted with the help of information provided in the KNIME software.

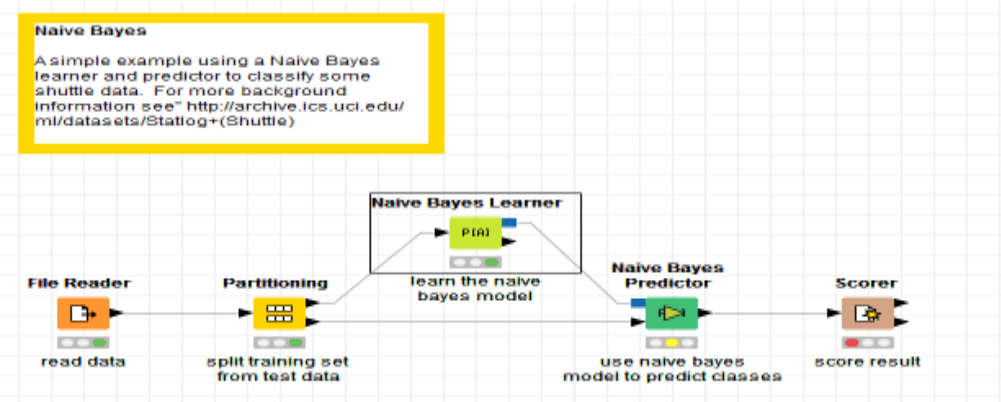


Fig. 3: Work flow of Naïve Bayes Model in KNIME application.

Table 2: Gaussian distribution per column.

Gaussian distribution for Col0 per class value							
	class0	class1	class2	class3	class4	class5	class6
Count:	15043	18	56	2938	1078	3	4
Mean:	44.04527	53.11111	54.5892	55.44724	87.58349	72	37
Std. Deviation:	6.52074	10.295	18.5067	0.52266	10.95826	14.79865	0
Rate:	79%	0%	0%	15%	6%	0%	0%
Gaussian distribution for Col1 per class value							
Count:	15043	18	56	2938	1078	3	4
Mean:	-0.99129	39.5	-48.9107	0.10313	8.34137	1242.66	-3260.75
Std. Deviation:	49.64024	21.33004	19.0666	1.52649	127.1492	640.798	1698.71
Rate:	79%	0%	0%	15%	6%	0%	0%
Gaussian distribution for Col2 per class value							
Count:	15043	18	56	2938	1078	3	4
Mean:	84.81427	78.88889	90.9642	86.4557	90.46475	90	98.75
Std. Deviation:	8.79504	2.19328	12.6404	8.94143	9.53974	15.0996	14.5
Rate:	79%	0%	0%	15%	6%	0%	0%
Gaussian distribution for Col3 per class value							
Count:	15043	18	56	2938	1078	3	4
Mean:	0.1979	0	-6.55357	0.13206	0.04267	-1	6.5
Std. Deviation:	16.32329	0	48.4035	3.49779	3.61237	1.73205	17.0782
Rate:	79%	0%	0%	15%	6%	0%	0%

4. Evaluation of Models

For the classification task, an error occurs when the model's prediction of the class label is different from the true class label. We can also define the different types of errors in classification depending on the predicted and true labels. Then the different types of errors are as follows.

1. If the true label is yes and the predicted label is yes, then this is a true positive, abbreviated as TP. This is the case where the label is correctly predicted as positive.
2. If the true label is no and the predicted label is no, then this is a true negative, abbreviated as TN. This is the case where the label is correctly predicted as negative.
3. If the true label is no and the predicted label is yes, then this is a false positive, abbreviated as FP. This is the case with the label is incorrectly predicted as positive, when it should be negative.
4. If the true label is yes and the predicted label is no, then this is a false negative abbreviated as FN. This is the case where the label is incorrectly predicted as negative, when it should be positive.

These four different types of errors are used in calculating many evaluation metrics for classifiers. The most commonly used evaluation metric is the accuracy rate, or accuracy for short. For classication, accuracy is calculated as equation (3). The accuracy rate is an intuitive way to measure the performance of a classification model as equation (5). Model performance can also be expressed in terms of error rate. Error rate is the opposite of accuracy rate listed in equation (6).

$$Classification\ rate = \frac{\#correct\ predictions}{\#Total\ predictions} \tag{3}$$

$$Error\ rate = \frac{\#Incorrect\ predictions}{\#Total\ predictions} \tag{4}$$

$$Accuracy\ rate = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Error\ rate = 1 - Accuracy\ rate \tag{6}$$

There's a limitation with accuracy and error rates when you have a class imbalance problem. That is when there are very few samples of the class of interest, and the majority are negative examples. A pair of evaluations metrics that are commonly used when there is a class imbalance are precision and recall. Precision is defined as the number of true positives divided by the sum of true positives and false positives. In other words, it is the number of true positives divided by the total number of samples predicted as being positive. Recall is defined as the number of true positives divided by the sum of true positives and false negatives. It is the number of true positives divided by the total number of samples, actually belonging to the true class. Precision is considered a measure of exactness because it calculates the percentage of samples predicted as positive, which are actually in a positive class as in equation

(7). Recall is considered a measure of completeness, because it calculates the percentage of positive samples that the model correctly identified as in equation (8). The goal for classification is to maximize both precision and recall. Both can be combined into a single metric called the F-measure as in equation (9).

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

With the F_1 measure, precision and recall are equally weighted. The F_2 measure weights recall higher than precision. And the $F_{0.5}$ measure weights precision higher than recall as in equation (10). The value for the F_1 measure ranges from zero to one, with higher values giving better classification performance.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (9)$$

$$F_1 = Evenly\ Weighted$$
$$F_2 = Weights\ recall\ more$$
$$F_{0.5} = Weights\ precision\ more \quad (10)$$

5. Results

Let us discuss how to interpret the results, obtained in our experiments conducted with the help of Table 3 and Table 4, we have constructed the Table 5 and Table 6. In which all the metric need to evaluate the classification task are listed. As Discussed in the section Evaluation of Model the classification model with highest F-measure value called the best on the dataset considered in our experiments.

Table 3: Document Term Matrix for Decision Tree

Col9\Prediction	Class 3	Class 0	Class 4	Class 1	Class 2	Class 6	Class 5
Class3	5924	41	0	0	0	0	0
Class0	0	30543	0	0	0	0	0
Class4	2	2	2185	0	0	0	0
Class1	0	19	13	0	0	0	0
Class2	2	73	40	0	0	0	0
Class6	0	9	0	0	0	0	0
Class5	2	0	5	0	0	0	0
Correct Classified: 38,652 Wrong Classified: 208 Accuracy: 99.465% Error: 0.535% Cohen's Kappa 0.985							

Table 4: Document Term Matrix for Naïve Bayes

Col9\Prediction	Class 1	Class 3	Class 0	Class 4	Class 2	Class 6	Class 5
Class1	12	0	11	9	0	0	0
Class3	48	3176	2700	0	41	0	0
Class0	36	542	29668	2	267	28	0
Class4	0	2	0	2186	1	0	0
Class2	0	0	36	9	64	6	0
Class6	0	0	5	0	0	4	0
Class5	0	1	1	1	1	0	3
Correct Classified: 35,113 Wrong Classified: 3,747 Accuracy: 90.358% Error: 9.642% Cohen's Kappa 0.705							

Table 5: Confusion Matrix for Decision Tree.

Row ID	TP	FP	TN	FN	Recall	precision	F-measure
Class3	5924	6	32889	41	0.99312	0.99898	0.99604
Class0	30543	144	8173	0	1	0.99530	0.99764
Class4	2185	58	36613	4	0.99817	0.97414	0.98601
Class1	0	0	38828	32	0	0	0
Class2	0	0	38745	115	0	0	0
Class6	0	0	38851	9	0	0	0
Class5	0	0	38853	7	0	0	0

Table 6: Confusion Matrix for Naïve Bayes.

Row ID	TP	FP	TN	FN	recall	precision	F-measure
Class3	12	84	38744	20	0.375	0.125	0.1875
Class0	3176	545	32350	2789	0.53243	0.85353	0.65579
Class4	29668	2753	5564	875	0.97135	0.91508	0.94238
Class1	2186	21	36650	3	0.99863	0.99048	0.99454
Class2	64	310	38435	51	0.55652	0.17112	0.26175
Class6	4	34	38817	5	0.44444	0.10526	0.17021
Class5	3	0	38853	4	0.42857	1	0.6

6. Conclusion

We would like to conclude that Decision tree perform better than Naïve Bayes in terms of classification accuracy using F-measure as evaluation metrics. We also addressed the problem of dataset imbalance in classification task and different ways to solve that. As a future work, we are interested to perform experiments on other traditional classification algorithm on the dataset considered in our experiments and compare the performance of them with Decision tree and Naïve Bayes to find the best classification algorithm in terms of classification accuracy.

References

[1] C. E. López Guarín, E. L. Guzmán and F. A. González,"A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining", *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, Vol.10, No.3, (2015), pp.119-125.

[2] Wei Chen, Xiaoshen Xie, Jiale Wang, Biswajeet Pradhan, Haoyuan Hong, Dieu Tien Bui, Zhao Duan, Jianquan Ma,"A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility", *CATENA*, Vol.151, (2017), pp.147-160.

[3] Zhao Zhang, Lei Wang, Lei Jia, Fanzhang Li, Li Zhang, Mingbo Zhao,"Projective label propagation by label embedding: A deep label prediction framework for representation and classification", *Knowledge-Based Systems*, Vol.119, (2017), pp.94-112.

[4] Zhengxing Huang, Tak-Ming Chan, Wei Dong,"MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records", *Journal of Biomedical Informatics*, Vol. 66, (2017), pp.161-170.

[5] Saeed Banihashemi, Grace Ding, Jack Wang,"Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption", *Energy Procedia*, Vol.110, (2017), pp.371-376.

[6] Mohammad Hossein Rafiei, Hojjat Adeli,"NEEWS: A novel earthquake early warning model using neural dynamic classification and neural dynamic optimization", *Soil Dynamics and Earthquake Engineering*, Vol.100, (2017), pp.417-427.

[7] Diego P.P. Mesquita, Lincoln S. Rocha, João Paulo P. Gomes, Ajalmar R. Rocha Neto,"Classification with reject option for software defect prediction", *Applied Soft Computing*, Vol.49, (2016), pp.1085-1093.

[8] Zeyu Wang, Ravi S. Srinivasan,"A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models", *Renewable and Sustainable Energy Reviews*, Vol.75, (2017), pp.796-808.

[9] Mazin Abed Mohammed, Mohd Khanapi AbdGhani, Raed Ibraheem Hamed, Dheyaa Ahmed Ibrahim,"Review on Nasopharyngeal Carcinoma: Concepts, methods of analysis, segmentation, classification, prediction and impact: A review of the research literature", *Journal of Computational Science*, (2017).

[10] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh,"Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study", *Journal of Forensic and Legal Medicine*, (2017).
[11] Miha Pavlinek, Vili Podgorelec,"Text classification method based on self-training and LDA topic models", *Expert Systems with Applications*, Vol.80, (2017), pp.83-93s.
[12] Tinghui Ouyang, Xiaoming Zha, Liang Qin,"A combined multivariate model for wind power prediction", *Energy Conversion and Management*, Vol.144, (2017), pp.361-373.
[13] Goran Mauša, Tihana Galinac Grbac,"Co-evolutionary multipopulation genetic programming for classification in software defect prediction: An empirical case study", *Applied Soft Computing*, Vol. 55, (2017), pp.331-351.
[14] Basha, Syed Muzamil, Yang Zhenning, Dharmendra Singh Rajput, N. Iyengar, and D. R. Caytiles,"Weighted Fuzzy Rule Based Sentiment Prediction Analysis on Tweets", *International Journal of Grid and Distributed Computing*, Vol.10,No.6, (2017), pp.41-54, DOI: 10.14257/ijgdc.2017.10.6.04.
[15] Basha, Syed Muzamil, Yang Zhenning, Dharmendra Singh Rajput, Ronnie D. Caytiles, and N. Ch SN Iyengar,"Comparative Study on Performance Analysis of Time Series Predictive Models", *International Journal of Grid and Distributed Computing*, Vol.10,No.8, (2017), pp.37-48, DOI: 10.14257/ijgdc.2017.10.8.04.
[16] Basha, Syed Muzamil, H. Balaji, N. Ch SN Iyengar, and Ronnie D. Caytiles,"A Soft Computing Approach to Provide Recommendation on PIMA Diabetes", *International Journal of Advanced Science and Technology*, Vol.106, (2017), pp.19-32, DOI: 10.14257/ijast.2017.106.03.
[17] Basha, Syed Muzamil, Dharmendra Singh Rajput, and Vishnu Vandhan,"Impact of Gradient Ascent and Boosting Algorithm in Classification", *International Journal of Intelligent Engineering and Systems (IJIES)*, Vol.11,No.1, (2018), pp.41-49. DOI: 10.22266/ijies2018.0228.05.
[18] Poluru, Ravi Kumar, and Shaik Naseera,"A Literature Review on Routing Strategy in the Internet of Things", *Journal of Engineering Science and Technology Review*, Vol.10,No.5, (2017), pp.50-60, DOI:10.25103/jestr.105.06.
[19] Bhushan, S. Bharath, and Pradeep Reddy,"A Four-Level Linear Discriminant Analysis Based Service Selection in The Cloud Environment", *International Journal of Technology*, *Vol*. 5, (2016), pp. 859-870.
[20] Bhushan, S. Bharath, and Reddy CH Pradeep,"A Network QoS Aware Service Ranking Using Hybrid AHP-PROMETHEE Method in Multi-Cloud Domain", *International Journal of Engineering Research in Africa*, Vol. 24, (2016).
[21] Gitanjali J,"Data mining from smart card data using data clustering", *International Journal of Applied Engineering Research*, Vol.11,No.1, (2016), pp.347-52.