Analysis of classification learning algorithms

Hana Rasheid Esmaeel
Department of Information & Communication, Al Nahrain University, Iraq

ABSTRACT
The paper attempts to apply data mining technique, to estimate the teacher performance of college of Information Engineering (COIE) In Al Nahrain University in Baghdad/Iraq. Five classifications algorithms were used to build data they are (ZeroR, SMO, Naive Bayesian, J48 and Random Forest). The analysis implemented using WEKA (3. 8. 2) Data mining software tool. Information was collected from within the variety of form using “Referendum”; it was stored in Excel file CSV format then regenerate to ARFF (Attribute-Relation File Format). Many criteria like (Time is taken to create models, accuracy and average error) was taken to evaluate the algorithms. Random forest and SMO Predicts higher than alternative algorithms since their accuracy is the highest and have the lowest average error compared to others, “The teacher clarification, and wanting to be useful to students”, was the strongest attribute. Further, removing the bad ranked attributes (10, 11, 12, and 14) that have a lower contact on the Dataset can increase accuracies of algorithms.

1. INTRODUCTION
“Extracting the knowledge from the massive set of information is called Data mining.” [1], The data ought to be new, not obvious, and one should be ready to use it, that ought to be helpful when deciding to find the buried patterns and interaction that should be useful in deciding. Data mining consists of five elements [1].

a) Extract, transform and load dealing information.
b) Store and manage the info in exceedingly two-dimensional info systems.
c) Provide information access to business analysts and knowledge technology professionals.
d) Analyze the info by application code.
e) Reward the info in an exceedingly helpful format like a graph or table.

The aim of this paper is to evaluate the teacher performance in (COIE) in Al Nahrain University in Baghdad/Iraq within the Referendum to the students to estimate the teacher performance by using five classifications algorithms, and then discover which attribute has the strongest effect in teacher evaluation, and which algorithm was the best.

Several works have used data mining to enhance teacher performance; Asanbe M. O. Osolisan A.O. and William W. F [2] have designed Artificial Neural Network (ANN) and Decision Tree scheme their system Was tested using data from a “Nigerian University”. Renuka Agrawal, Jyoti Singh, and Zadgoankar [3] “suggests type to estimate the performance victimization data mining” like (“association,” classification rules “Decision Tree”, “Rule Induction,” K-NN, and “Naïve Bayesian”) to search out traditions to assist them to reinforce supply the educational process”. Hemaid and El- Halees [4] in their study they use Questionnaire which has questions on the classes, they intend a form to “test teacher performance during data mining techniques like, classification, association rules to find out behavior to aid them toward enhanced the
learning procedure and improve the presentation of teachers in classroom”, to enhance the educational process and expand the contribution of teachers in the classroom”. Works published by Ahmeda, Rizaner and Ulusoyc [5] using The sequential Minimal Optimization, Naïve Bayes, J48 Decision Tree, and Multilayer Perceptron to Evaluate Student records to predict the teacher performance and investigates factors that have affected students achievements to develop the teaching system. In [6], in his study to Predict students’ performance he finds that the classification scheme is repeatedly used in educational data mining area, it includes, NeuralNetwork and Decision Tree, the two methods greatly used by the researchers for predicting students’ performance. Ms. A. Pavithra, Mr. S. Dhanaraj [7] In their study they examine the prediction accurateness of the academic performance of teaching the students using different classification algorithms like, “MLP, Naïve Bayes, Decision tree, REP tree, and J48 tree”, they concluded that “many factors will influence the student performance, and it may differ to the different locality of students”. Farid Jauhari, Ahmad Afif Supianto [8] proposes three boosting algorithms (C5.0, AdaBoost. M1, and AdaBoost. SAMME) to build the classifier for predicting student's performance. They used three scenarios of evaluation, the first scenario employs 10-fold cross-validation to compare the performance of boosting algorithms. The second scenario was accustomed to evaluate boosting algorithms below the various varieties of coaching information within the third scenario, they build models from one subject Dataset, and test using another subject Dataset. They conclude that the third scenario results indicate that they can build a prediction model using one subject to predict another. Bin Mat and N. Buniyamin [9] using neuro-fuzzy tool to classify and predict electrical engineering students graduation achievement based on mathematics competency. It's supported longitudinal progress and cross-validation model on two arithmetic subjects, semesters’ performance, and graduation achievement of electrical students. They conclude that the mixture of statistical associate analysis and machine learning will facilitate to extract data, and alter university management to assist low achievers at an early stage. They hoped that their findings can help faculty management to review mathematics curriculum with respect in the increasing range of engineering field. S. Hussain, N. A. Dahan, F. M. Ba-Alwih, and N. Ribata [10] used for classifications methods, (J48, PART, Random Forest and Bayes Network Classifiers). The high influential attributes were selected using the data mining tool Weka. They conclude that the Random Forest Classification method was the most suited algorithm for the Dataset.

Saouabi Mohamed, Abdullah Ezzati [11] proposes a data mining process for employability data using classification techniques (the Decision Tree classifier, Logistic regression, and Naïve Bayes algorithms), apply them by Rapid Miner Studio Educational Version (8.1.000), using employability Dataset. They conclude that the Decision tree classifier is more accurate than Logistic regression and Naïve Bay.

2. RESEARCH METHOD

2.1. Prepare Data

For this analysis, data were collected from college students at totally different Departments within the faculty of College of Information Engineering (COIE) at Al Nahrain University in Baghdad/Iraq for the aim of the investigation; however their skilled improvement has taken Place throughout the term. The info was collected from the college within the type of form to judge teacher performance as shown in Table 1.

a) Preparing Information. Teacher's Data was Evaluated. Attributes and value were determined.
b) Data saved in excel file in CSV (Comma Separated Values).
c) To use Weka data must be converted to Arff (Attribute-Relation File Format)
d) Using Weka (3.8.2) GUI Chooser and Explorer.
e) Using Wekapreprocessor and open Arff file.
f) Apply classification algorithms (ZeroR, SMO, Naïve Bayesian, J48 tree and Random Forest).
g) Evaluate the result and performance. Figure 1 shows the proposed system layout.

| Data Collected From (COIE) In Al Nahrain University | Attribute and Value Are Determined |
|---------------------------------------------------|-----------------------------------|
| Data Saved in CSV File Format (Excel File)        | Converted to Arff (Attribute Relation File Format) |
| WEKA (3.8.2) GUI Chooser                          | Using Preprocessor |
| Apply Classification Algorithms (ZeroR, SMO, Naïve Bayesian, J48 tree and Random Forest) | Evaluate Result and Performance |

Figure 1. The proposed system
Table 1. The Questions and Their Abbreviations

| Seq | Attribute | Description |
|-----|-----------|-------------|
| 1   | DESG.     | Title       |
| 2   | QUA.      | Degree      |
| 3   | EXP.      | Experience  |
| 4   | SC_GTDL   | The semester course content pedagogic and analysis were provided at the beginning |
| 5   | CA_OBLS   | At the beginning the teacher were clearly specified the aims and purpose of the course. |
| 6   | CW_CA     | The amount of credit allotted to the course was positively significant |
| 7   | CTA_SAAY  | The course was instructed in step with the information proclaimed on the primary day of sophistication |
| 8   | CD_HW_ASS_APP_SAT | The class discussions homework assignment applications and studies were satisfactory |
| 9   | TB_other_CR | The text book and different courses resources were enough and up to this point |
| 10  | C_WAPP_LAB_DIS | The course acceptable conversation of the laboratory applications and different studies. |
| 11  | QUIZ_ASS_PROJ_EXA_HEP | The quizzes assignment comes and exams contributed to serving to the educational. |
| 12  | ENJ_CLA_ACTIV_DUR_LEC | The lecture allows students to participate their knowledge. |
| 13  | INT_EXPE_C_END_Y | The course were met all student prospect |
| 14  | CWR_BTMS_PRO_DEVE | The course was relevant and helpful to my skilled development. |
| 15  | CHM_LA_W_PRE | The course helped Maine check up on life and world with my new perspective. |
| 16  | INS_KW_RELE_DATE | The lecturer information was relevant and up to this point. |
| 17  | INS_CP | The lecturer came ready for categories. |
| 18  | INS_TAUG_IN_ACCOR_ | The lecturer instructed in accordance with the proclaimed lesson set up. |
| 19  | INS_W_COMMI_T_THE_CO | The lecturer committed to the course and was comprehensible |
| 20  | INS_ARR_OF_T_C | The lecturer arrived of your time for categories |
| 21  | INS_HSM_C_HO | The lecturer had a sleek associate of sophistication hours |
| 22  | INS_EXP_THE_CO_A_WE_HE_T | The teacher clarification and was wanting to be useful to students |
| 23  | INS_DE_AP_APP_T_ST | The lecturer incontestable appositive approach to students |
| 24  | INS_W_OP_RES_VIE_ST_CO | The teacher was considerate of the views of student on the topic of the course. |
| 25  | INS_EN_PART_IN_CO | The lecturer inspired participation within the course. |

2.2. Data Collected

The knowledge was gathering for making ready the model, the fields that are needed for data processing was taken, this includes pre-processing or extracts vital info from it then produce correct format file of the info like inweka.arff file format (Attribute Relation File Format) [3], as shown in Figure 2.

```
@relation teacher evaluation
@attribute Name1 string
@attribute Name2 numeric
@attribute Namen?
@data
Yes, I, yes....
```

2.2. Data Collected

The knowledge was gathering for making ready the model, the fields that are needed for data processing was taken, this includes pre-processing or extracts vital info from it then produce correct format file of the info like inweka.arff file format (Attribute Relation File Format) [3], as shown in Figure 2.

```
@relation teacher evaluation
@attribute Name1 string
@attribute Name2 numeric
@attribute Namen?
@data
Yes, I, yes....
```

Figure 2. Tacher.arff file
2.3. Methodology

The classification technique was used for the forecast of teachers’ evaluation. Five classifications algorithms were used (ZeroR, SMO, Naïve Bayesian, J48 tree and Random Forest) and implemented using Weka (3.8.2) Data mining software tool.

2.4. Building Models

2.4.1. Building the Trivial model ZeroR [12, 13]

a) In Preprocess panel, click “Open file” button, choose the file named (teacher.arff) as in Figure 3.

b) Select ZeroR by Clicking “Choose” Button.

c) Invoke classifier by clicking “start” button to make a model. The analytical performance of the model characterized by the right-hand classifier output frame.

d) The Confusion Matrix for the model is bestowed at the underside part of the Classifier output window. It is seen from it that compounds are classified as (21) affirmative and (43) No.

e) The accuracy of the model is (67.187) for No and (32.813) for affirmative as in Figure 4 and Figure 5.

![Figure 3. Selected attributes of teacher.arff](image)

![Figure 4. ZeroR classifier output](image)
2.4.2. Building J48 Tree Model [12]

To get the usual illustration of the tree, the following must be done:

a) Click the correct push on the model type trees (J48) within the Result list frame and choose

b) The menu item visualizes tree size a replacement window with graphical Illustration of the tree.

c) Click with the correct push to the area during this screen, and within the popup menu choose the item appropriate screen. As in Figures 6 and 7.
2.4.3. Building Naive Bayesian Model [12, 13]

As in previous models Naive Bayesian Model was built as shown in Figure 8.

Figure 8. Navie bayesian model

2.4.4. Building Support Vector Machine Models [12-15]

The Weka software implements Sequential Minimal Optimization (SMO) algorithm for training a support vector classifier. Figure 9 show the classifier output for this model.

The accuracy of (Correctly Classified Instances) of this model is extremely high ninety eight.4375%. This truth clearly indicates that the accuracy can't be used for assessing the utility of classification models designed exploitation unbalanced datasets. For this purpose an honest selection is to use the Kappa statistic, that is =0.964 for this case. Kappa statistic is the academic degree analog of the constant of correlation. It's worth is zero for the shortage of any relation and approaches to (1) for terribly sturdy applied math relation between the category label and attributes of instances. Another helpful applied math characteristic is “ROC Area”, that the worth =0.976 means that sensible mythical monster curves may be build and therefore the cost/benefit analysis will simply be performed. As in Figure 10.
2.4.5 Building the Random Forest Models [12-15]

The classifier output used for "Random Forest" algorithm be shown in Figure 11 below the accuracy of this model is extremely high: ninety eight.4375%. This reality clearly indicates the accuracy cannot be used to assessing the worth of classification models designed victimization unbalanced datasets. For this purpose an honest alternative is to use the “Kappa statistic”, that is =0.964 for this case [6]. Its price is =0.9641 it’s terribly robust applied math relation between the category label and attributes of instances. Another helpful applied math characteristic is “ROC Area”, that the worth =1.000 means that sensible mythical creature curves are built and therefore the cost/benefit analysis will simply be performed. As in Figure 12.
Figure 11. Classifier Output for Random Forest

Figure 12. Cost/benefit Random Forest
3. RESULTS AND DISCUSSION

3.1. Attribute Ranking [16-20]

Weka explorer can evaluate the attributes of the data by using the following steps:
Select attributes→GainRatoAttribute→start→showEval→Rank Attribute

The list of attributes and their value are appearing from higher to lower, in Table 2 show that.

| Table 2. Attribute Ranking |
|-----------------------------|
| Attribute | Value |
| 22         | 0.468955 |
| 23         | 0.42375  |
| 24         | 0.404508 |
| 25         | 0.341001 |
| 26         | 0.30944  |
| 27         | 0.289066 |
| 28         | 0.269511 |
| 29         | 0.250266 |
| 30         | 0.231283 |
| 31         | 0.21283  |
| 32         | 0.19439  |
| 33         | 0.175211 |
| 34         | 0.15606  |
| 35         | 0.13672  |
| 36         | 0.11707  |
| 37         | 0.09741  |
| 38         | 0.07734  |
| 39         | 0.05747  |
| 40         | 0.03781  |
| 41         | 0.01784  |
| 42         | 0.00805  |
| 43         | 0.00293  |
| 44         | 0.00070  |
| 45         | 0.00017  |
| 46         | 0.00004  |
| 47         | 0.00001  |
| 48         | 0.00000  |

3.2. Models Comparison [21, 22]

The performances of the 5 models were evaluated primarily based on the standards as illustrated in Table 3.

a) Prediction accuracy: The share of properly classified instances is usually referred to as accuracy of a model.
b) Time is taken to create the model.
c) Error rate.

| Table 3. Comparison Analysis on The Models |
|--------------------------------------------|
| Metric                             | ZeroR | J48   | NaïveBayesian | SMO   | RandomForest |
|---------------------------------------|-------|-------|---------------|-------|--------------|
| Time To Build The Model              | 0     | 0.03  | 0.05          | 0.03  | 0.03         |
| Correctly Classified Instances       | 67.187% | 95.312% | 92.187%     | 98.437% | 98.435%      |
| In Correctly Classified Instances    | 32.821% | 4.687% | 7.812%       | 1.562% | 1.562%       |
| Kappa Statistics                     | 0     | 0.895 | 0.833        | 0.964 | 0.964        |
| Mean Absolute Error                  | 0.443 | 0.0645 | 0.0791      | 0.0156 | 0.0213       |
| Root Mean Square Error               | 0.47  | 0.210 | 0.279        | 0.125 | 0.0939       |
| Relative Absolute Error              | 100%  | 14.55 % | 17.838%     | 3.524% | 4.809 %      |
| Root Relative Square Error           | 100%  | 44.765% | 59.428%    | 26.594% |              |

3.3. Performance of the Models

Table 4 show the performance of the (5) algorithms: [11, 23-25]

TP= true positives": a variety of examples": Predicted positive that are literally positive.
FP= false positives": a variety of examples: "Expected positive that are literally negative.
TN= true negatives": a variety of Examples ": predicted negative that are literally negative.

---

Analysis of classification learning algorithms (Hana Rasheid Esmaeel)
FN=false negatives": a variety of Examples: "Expected negative that are literally positive. Weka (3.8.2) Confusion Matrix: The quantity of properly classified instances is that the total of diagonals within the matrix; all others area unit incorrectly classified.

\[
x \times y \leftarrow \text{classified as, actual } x=0 \text{ TP FN}
\]
Actual y=1 FN TP

\[
\text{TP=TP+FN / Recall}
\]
\[
\text{Precision=TP/TP+FP}
\]
\[
\text{Accuracy=TP+TN/TP+TN+FP+FN}
\]

Table 4. Performance of The Model

| Algorithm    | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area |
|--------------|---------|---------|-----------|--------|-----------|----------|
| ZeroR        | 0.672   | 0.672   | ?         | 0.672  | ?         | 0.455    |
| J48 Tree     | 0.953   | 0.047   | 0.954     | 0.953  | 0.953     | 0.951    |
| Naive Bayesian | 0.922  | 0.038   | 0.937     | 0.922  | 0.924     | 0.927    |
| SMO          | 0.984   | 0.032   | 0.985     | 0.984  | 0.984     | 0.976    |
| RandomForest | 0.984   | 0.032   | 0.985     | 0.984  | 0.984     | 1.000    |

4. CONCLUSION

From the result of comparison of the five algorithms as in Tables 4 and 5 it conclude that Algorithms SMO and Random forest predicts higher than alternative algorithms since their accuracy is that the highest and have lowest average error compared to others algorithms on functioning on performance, several attributes are tested, and found that a few of them are effective on the performance prediction. "The teacher clarification and was wanted to be useful to students", was the strongest attribute and then the result plays a vital role within the performance of academics. More a lot of removing the worst hierarchical attributes (10, 11, 12, and 14), that have a lower impact on the dataset can increase the algorithms performance accuracies.

REFERENCES

[1] Shafiq Aslam, Imran Ashraf, “Data Mining Algorithms and Their Applications in Education Data Mining”, International Journal of Advance Research in Computer Science and Management Studies, Vol 2, Issue 7, pg. 50-56, July 2014.

[2] Asanbe M.O., Oosifian A.O., William W.F. “Teachers’ Performance Evaluation in Higher Educational Institution using Data Mining Technique”, International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 10 – No.7, March 2016.

[3] Renuka Agrawall J. Jyoti Singh, A.S. Zadgoonkar, “Summative Assessment for Performance Evaluation of a Faculty Using Data Mining Techniques,” International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 10, October 2016.

[4] Randa Kh. Hemaid Alaa M. El-Halees, “Improving Teacher Performance using Data Mining”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 2, February 2015.

[5] Ahmed Mohamed Ahmeda, Ahmet Rizanelc, Ali Hakan Ulusoyc, “Using Data Mining to Predict Instructor Performance”, 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS, Vienna, Austria.29-30 August 2016.

[6] A. Mohamed Shahiria, W. Husaina, N. Abdul Rashida, “A Review on Predicting Student’s Performance using Data Mining Techniques” Procedia Computer Science 72 .414 – 422, ELSEVIER 2015. Available online at www.sciencedirect.com

[7] Ms.A.Pavithra, Mr.S.Dhanaraj, “Prediction Accuracy on Academic Performance of Students Using Different Data Mining Algorithms with Influencing Factors”, IJSRCAMS Vol 7, Issue 5, 2018.

[8] Farid Jauhari, Ahmad Afif Sapianto, “Building Student’s Performance Decision Tree Classifier Using Boosting Algorithm”, Indonesian Journal of Electrical Engineering and Computer Science (IJECS)Vol. 14, No. 3. pp. 1298-1304, June 2019.

[9] U. Bin Mat and N. Buniyamin, “Using Neuro-Fuzzy Technique to Classify and Predict Electrical Engineering Students Achievement Upon Graduation Based On Mathematics Competency,” Indonesian Journal of Electrical Engineering and Computer Science (IJECS), vol. 5, no. 3, pp. 684–690, 2017.

[10] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, “Educational data mining and analysis of students academic performance using WEKA,” Indonesian Journal of Electrical Engineering and Computer Science (IJECS), vol. 9, no. 2, pp. 447–459, 2018.

[11] Saouabi Mohamed, Abdullah Ezzati “A data mining process using classification techniques for employability prediction”, Indonesian Journal of Electrical Engineering and Computer Science (IJECS) Vol. 14, No. 2, pp. 1025-1029, May 2019.

[12] Tutorialspoint.com, “Data Mining Tutorial Simply Easy Learning,” 2-11-2014
Analysis of classification learning algorithms (Hana Rasheid Esmaeel)

[13] The WEKA Workbench Eibe Frank, Mark A. Hall, and Ian H. Witten Online Appendix for. “Data Mining: Practical Machine Learning Tools and Techniques,” Morgan Kaufmann, Fourth Edition, 2016

[14] Miss. Sharayu N. Bonde, Dr. D. K. Kirange, “Survey on Evaluation of Student’s Performance in Educational Data Mining,” Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - Part Number: CFP18BAC-ART; ISBN-978-1-5386-1974. 2018.

[15] Anil Kumar Tiwari, G. Ramakrishna, Lokesh Kumar Sharma, Sunil Kumar Kashyap, “Academic performance prediction algorithm based on fuzzy data mining,” International Journal of Artificial Intelligence (IJ-AI) Vol. 8, No. 1, pp 26 – 32, March 2019.

[16] Weka – “Data Mining Machine Learning Software”. Available at: http://www.cs.waikato.ac.nz/ml/weka/, (Accessed 20 April 2014).

[17] Yadav, S.K., Bhardwaj B.K., and Pal, S., “Data Mining Applications: A Comparative Study for Predicting Student’s Performance”, International Journal of Innovative Technology and Creative Engineering (IJITCE), 2011, pp. 13-19.

[18] Asma Lamani, Brahim Erraha, Malika Elkyal, Abdallah Sair, “Data mining techniques application for prediction in OLAP cube”, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 3, June 2019, pp. 2094–2102.

[19] Abdulsalam Sulaiman O., Babatunde Akinbowale N., & Babatunde Ronke S., “Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students’ Performance in Computer Programming”, A Multidisciplinary Journal Publication of the Faculty of Science, Adeleke University, Ede, Nigeria, 2015 Vol 2.

[20] Adillah Dayana Ahmad Dali, Nurul Aswa Omar, Aida Mustapha, “Data Mining Approach to Herbs Classification”, Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), Vol. 12, No. 2, November 2018 : 570 – 576.

[21] Kandasamy P, Palamur. “Performance analysis of classifier models to predict diabetes mellitus”, Journal of Elsevier, vol 47, pp.45-51, 2014.

[22] Parneet Kaura, Manpreet Singhb, Gurpreet Singh Josanc “Classification and Prediction based Data Mining Algorithms to Predict Slow Learners in Education Sector”, Science Direct Procedia Computer Science 57, 500-508, 2015 (ICRTC-2015).

[23] Hlaudi Daniel Masethe, Mosima Anna Masethe, “Prediction of heart disease using classification algorithms”, Proceedings of the world congress on engineering and computer science, San Francisco, USA, pp. 185-191, 2014.

[24] Aqliima Aziz, Cik Feresa Mohd Foozy, Palaniappan Shamala, Zurinah Suradi, “YouTube Spam Comment Detection Using Support Vector Machine and K–Nearest Neighbor”, Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), Vol. 12, No. 2, November 2018 : 607 – 611.

[25] Ratna Patil, et al., “A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes”, International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 5, pp. 3966-3975, 2018.