RSU Journal of Biology and Applied Sciences

ISSN:

RSUJBAS

About us

Rivers State University Journal of Biology and Applied Sciences (RSUJBAS) publication is a quarterly, open access, international journal for all academic research in science discipline. Microbiology, Botany, Zoology, Environmental Biology, Chemistry, Physics, Mathematics, Computer Science, Biochemistry, Medical Laboratory Sciences and other applied science related areas. RSUJBAS is a platform set for elites to influence, contribute and communicate to the global environment through their various academic researches. We synergistically engage our noble efforts to contribute to the knowledge development, discoveries and innovations in all fields of study. In RSUJBAS we publish research papers on current academic issues with standard scientific reviews. RSUJBAS publishes original research articles, review articles, case studies, short communications, survey report, comparative studies and many more.

Aim and scope

Rivers State University Journal of Biology and Applied Sciences aims to publish high quality papers that communicate fundamental and contemporary discoveries both theoretical and practical. Most importantly RSUJBAS seeks to establish a platform for communicating emerging trends in various discipline such as Microbiology, Botany, Zoology, Environmental Biology, Chemistry, Physics, Mathematics, Computer Science, Biochemistry, Medical Laboratory Sciences and other applied science related areas.

Description:

- Area of concentration: All science academic disciplines
- Frequency of publishing: Quarterly
- Mode of publishing: both online and print publication
- Language of publication: English
- Double Blinded Review Process
- Zero Level Plagiarism Tolerance

Why publish with us

Low Article Processing Charge (ACP) to promote the research work
Easy and Rapid review process
Instant publication upon acceptance
Dedicated editorial and review team for fast review process
RSUJBAS provides hard copies of publication every quarterly

Editorial Board

All Correspondence to
Dr. Sam Wenedu (Editor-in-Chief)
Department of Microbiology, Rivers State University
Editor.jbas@yahoo.com

Or

OLUCHI DICKSON
Publication Manager
Dicksonoluchi87@gmail.com

Consulting Editors

Guideline for Manuscripts

Manuscripts should be typewritten on an A4 sheet having B1.5= line spacing throughout the text. The margins should be 2B54cm (1 inch) in all sides and page number should be consecutively on the bottom of the page. The manuscript should be written in Times New Roman using '12' font size.

For original research paper, the manuscript should be arranged in the following order: Title page, Abstract, Keywords, Introduction, Materials and Methods, Results, Discussion, Acknowledgment, References, Tables with legends, figures with legends and supplementary materials

The title page should contain the title, the name(s) of the author(s), the name(s) and address(es) of the institution(s) where the work was carried out, including a valid e-mail address from the corresponding author along with telephone numbers. The title of the manuscript should be specific and concise but sufficiently informative.

The Abstract should not exceed 250 words and it should contain brief summary of the findings including brief introduction, methodology, results, and conclusions,

The keywords should have a minimum of five and maximum of seven words.

The introduction should provide a clear statement of the problem and indicates aim of the study citing relevant literature to support background statements.

The Materials and Methods should include the methods and methodology of the research.

The Results should be presented in the form of tables or figures. It should be presented with clarity and precision. Statements used to present results should be written in the past tense. Detailed interpretation of data should not be included in the results but should be put into the Discussion section.

The Discussion should interpret the results clearly and concisely, and should integrate the research findings of this and past studies on this topic. Highlight the significant/unique findings of the research under conclusion.

The acknowledgments of people, grants or funds should be brief.

Contents

List of Contributors

Kingsley-Opara, Ngozi
Research Scholar, Department of Computer Science, Ignatius Ajuru University of Education, Rivers State, Nigeria.
Advanced Database Management

Prof. Asagba, Prince Oghenekaro.
Visiting Scholar, Department of Computer Science, University of Port-Harcourt, Rivers State Nigeria.
Email:asagba.prince@uniport.edu.ng

Gabriel B.C., Gabriel M. N, P. O. Asagba
School of Graduate Studies Ignatius Ajuru University Of Education (IAUE), Rumuolumeni, Port Harcourt, Rivers State, Nigeria.
Department Of Computer Science.
gabrielbariyira@gmail.com, meegabz@gmail.com

WAIDOR, Tamaramiebi Keith[1]& ASAGBA, Prince Oghenekaro[2]
Department of Computer Science,
Faculty of Natural and Applied Sciences
Ignatius Ajuru University of Education, Port Harcourt
zalimaxxx@gmail.com

[2]Department of Computer Sciences,
University of Port Harcourt, Port Harcourt, Nigeria
Prince.asagba@uniport.edu.ng

Fiberesima, Alalibo Ralph
Visiting Scholar, Department of Computer Science,
University of Port-Harcourt, Rivers State Nigeria.
fiberesima.a.r@outlook.com;

Asagba, Prince Oghenekaro
Visiting Scholar, Department of Computer Science,
University of Port-Harcourt, Rivers State Nigeria.
asagba.prince@uniport.edu.ng

Kingsley-Opara, Ngozi
Research Scholar, Department of Computer Science,
Ignatius Ajuru University of Education, Rivers State, Nigeria.
Email: ngozikopara@gmail.com

Prof. Asagba, Prince Oghenekaro.
Visiting Scholar, Department of Computer Science,
University of Port-Harcourt, Rivers State Nigeria.

Comparative Analysis of Classification Algorithms in Educational Data Mining

WAIDOR, Tamaramiebi Keith
Department of Computer Science,
Faculty of Natural and Applied Sciences
Ignatius Ajuru University of Education, Port Harcourt
zalimaxxx@gmail.com

ASAGBA, Prince Oghenekaro
Department of Computer Sciences,
University of Port Harcourt, Port Harcourt, Nigeria
Prince.asagba@uniport.edu.ng

Abstract
Data mining is a field of computer science within Machine Learning (ML) and Artificial Intelligence. With the advent of the internet and increase in computing speed and storage capacity, the amount of data collected is also on the rise in different data warehouses. Hence, need to explore the humongous volume of data being generated in order to extract patterns representing knowledge for decision making arose. Data mining is fast gaining application in several facets of life, and the educational sector is not left out. This work starts with a brief overview of data mining techniques, then delved into is a comparative analysis of 3 different Classification techniques such as decision tree (j48), Bayes (NaiveBayes) and Rules (oneR) on student academic performance dataset, using statistics such as F-Measure and Percentage – Correct. The methodology adopted is the CRISP-DM due to its' advantages over other data mining methodologies as it is popularly used and provides a uniform framework for planning and managing a project. The tool employ in our data mining analysis is the Wekaito Environment for Knowledge Analysis (WEKA 3.0). Our results show that J48 performed better with 95.5556% prediction accuracy than the other algorithms. Consequently, it was chosen to build the prediction model of student academic performance.

Key words: *Data mining, Classification, NaiveBayes, Decision Tree, OneR*

1.0 Introduction

Forecasting students' academic performance has always been a thing of interest and with will always be a subject of keen interest. Any institution best interest is to produce the best of students and the ability to detect early strugglers if vital, in order to quickly offer such students the necessary help to boost their academic performance. One of such way to predict the academic performance via Educational Data Mining (EDM). EDM according to Carla Silva et al (2017) Is aimed at devising and using algorithms to improve educational results and explain educational strategies for further decision making.

On the other hand, data mining is the discovery or extraction of knowledge from a large repository of data using very clever algorithms and techniques. The Knowledge extracted are gotten through patterns from the data repository(Waidor et al., 2018).It is a field in computer science under Artificial Intelligence (AI) that has gained great popularity in recent times. It is a multi-disciplinary pool with streams such as database, statistics, information retrieval, machine learning etc. flowing into it (Hemlata et al.), as depicted is Figure 1.

Data mining or Knowledge Discovery from Database (KDD) is a seven - stage process that includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge representation as show in Figure 2.Data mining consist of four

main types of techniques which are Classification, Regression and Clustering (Hemata et al.), and Association Rule (Kodeeshwari, 2017).
This work is aimed at analysis students' result data with different classification algorithms the identify the one with the best result. This was done by employing the Waikato Environment for Data Analysis (Weka) data mining tool.

2.0 The Review of Related Literature

Data mining has found application even in the medical field, especially classification and regression as data mining techniques for predicting the diseases outbreak being permitted in the health institutions. Hakizimana L., et al., (2016) presented a survey and analysis for existing techniques on both classification and regression models techniques that have been applied for diseases outbreak prediction in datasets. The authors noted that attributes that are continuous regression model using Support Vector Machine or linear regression achieved better performance.

Considering the huge amount of data on the web, Jadranka L., et al (2000) did some analysis on techniques of information retrieval with intent to bring to fore their strong and weak points. They analysed several advanced methods for Web information mining such as syntax analysis, metadata-based searching using RDF, knowledge annotation by use of conceptual graphs (CGs), KPS: Keyword, Pattern, Sample search techniques, and techniques of obtaining descriptions by fuzzification and back-propagation. The problem of proper indexing and subjective classification were highlighted and universally known classification is recommended. The authors discovered that usage of KPS algorithm was probably more suitable for searching one site, than the whole Web, albeit, it could not mine all desired information, but useful for information extraction of textual Web pages. Furthermore, methods of fuzzification and back-propagation could aid existing classification and relying mostly on the interconnectivity of the Web pages.

3.0 Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was adopted in this research work due to its' advantages over other data mining methodologies as it is popularly used (see figure 4) and provides a uniform framework for planning and managing a project. **CRISP-DM** involves 6 **different phases which are** Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment

The dataset used for the experiment are the students' results of the faculty of Basic and Applied Sciences, University of Africa, Toru-Orua, made up of 7 different departments. The level of study are 100 and 200 level students. The attributes that make of our dataset include RegNo, Gender, Level, Dept., Previous CGPA, Current GPA, CGPA, Verdict. There are 317 instances (both for training and testing). The dataset was converted from Microsoft Excel to CSV (Comma Separated Value) before it was used. This is because the tool for our analysis WEKA does not accept files in Excel extension.

4.0 Results

In this research work, we compared the performance of 3 classification algorithms such as Decision Tree (J48), Bayes (NaiveBayes) and Rule based (RuleOne), to determine the one with the best performance in classification with respect to our dataset. WEKA's Experiment pane enables us to do a Paired T-Tester using comparison field, F_Measure. We the double –

checked using Percent_Correct. The experiment was done 0.05 level of significance. (See Table 1).
In Table 1, the symbol (v) represents Victory, the asterisk (*) symbol represents Failure and the blank space (/ /) represents the inability to determine whether it is a victory or failure. For the given dataset, both trees.J48 and baye. NaiveBayes posted a success of 0.95 at a confidence interval of 0.05% using the F-measure analysis. However, NaiveBayes algorithm posted '1' under blank space (/ /) signifying that the inability to determine when NaiveBayes is better than J48 or not, at 0.05 confidence interval. Rules. OneR posted 0.48 with an asterisk beside it – meaning it is a failure and won't be a suitable classifier algorithm with respect to our dataset.
In order to double check the validity of our claim, we used Percent-Correct as our determining parameter analysis, keeping the confidence interval still at 0.05%. J48 and NaiveBayes algorithms posted 93% while OneR posted 32% with an asterisk beside it, indicating failure.(See Table 2). By the result, it was determined that J48 and NaiveBayes were more appropriate algorithms than OneR with respect to our dataset.

In building the predictive model was built using WEKA. The option adopted was the '10 fold cross-validation'. Cross validation is a systematic way of doing Repeated Holdout, and has proved over time to produce better result. In a 10 – fold cross validation, the whole dataset is divided into 10 equal (or almost equal) parts. 9 out of the 10 parts are used as for training while the remaining 1 part is held and used for testing. The process is repeated 10 times, each time using a different portion of the dataset for the testing. At the end of the process, the average of all the results becomes the target result. We used both the J48 and NaiveBayes classification algorithm in order to do a comparative analysis.

The outputs of our results in Table 3 and Table 4.The main statistic of interest in this research work is the Correctly Classified Instances. J48 algorithm gave a prediction accuracy of 95.5556% while NaiveBayes posted a prediction accuracy of 89.2063%.

5.0 Discussion and Recommendation
You will recall that in the Paired T-Tester comparison between J48, NaiveBayes and OneR, there was no clear favourite between J48 and NaiveBayes. Using F-measure, both posted 0.95 at a confidence interval of 0.05%. Using Percentage-Correct, both posted 93.9% at a confidence interval of 0.05%. However, we could not immediately determine whether NaiveBayes was a better option than J48 in analysing the dataset. But from the result gotten in building our model, J48 gave prediction accuracy of 95.5556% which was better than Naivebayes' 89.2063%.

From our result, it is safe to conclude that J48 (Decision Tree) a better prediction of student academic performance than NaiveBayes (Bayes) data mining algorithm.
Compared to other algorithms decision trees requires less effort for data preparation during pre-processing, missing values in the data also do not affect the process of building a decision tree to any considerable extent and A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders. However, a small change in the data can cause a large change in the structure of the decision tree causing instability, and sometimes calculation can go far more complex in decision tree compared to other algorithms (Dhiraj, 2019).

As for Naivebayes algorithm, it can work very fast and can easily predict the class of a test dataset, it also allows one to solve multi-class prediction problems as it's quite useful with

them and it performs better than other models with less training data if the assumption of independence of features holds. However, If your test data set has a categorical variable of a category that wasn't present in the training data set, the Naive Bayes model will assign it zero probability and won't be able to make any predictions in this regard. The algorithm is also a lousy estimator. And another demerit is it assumption that all the features are independent. While it might sound great in theory, in real life, you'll hardly find a set of independent features.

Further work can be done to determine which of the Decision Tree algorithms (such as J48, Decision Stump, Random Tree, Random Forest and Hoeffding Tree etc.) gives a better prediction of students' academic performance**.**

References

Akpojaro, J. & Waidor, T.K. (2019). A data mining framework for improving agricultural production in Nigeria. *Computing, Information Systems, Development Informatics & Allied Research,*10*(2),* 61-68.

Hakizimana, L., Wilson, K. C. & Stephen, K. (2016).A survey and analysis on classification and regression data mining techniques for diseases outbreak prediction in datasets. *The International Journal Of Engineering And Science* , 5*(9),* 4.

Hemlata, S., Shalini, S. & Seema, G. A brief overview on data mining survey. *International Journal of Computer Technology and Electronics Engineering,*1*(3),* 114.

Keerthi Sumiran (2018). An overview of data mining techniques and their application in industrial engineering. *Asian Journal of Applied Science and Technology,* 2*(2),* 948.

Kodeeshwari, R.S. & Tamil Ilakkiya, K. (2017). Different types of data mining techniques used in agriculture - a survey. *International Journal of Advanced Engineering Research and Science,* 4*(6),* 23.

Ragavi, R., Srinithi, B. & Anitha Sofia, V. S. (2018). Data mining Issues and challenges: A Review. *International Journal of Advanced Research in Computer and Communication Engineering,* 7*(11),*119-120.

Sivanagamani, N. (2018). Review on data mining techniques. *International Journal of Computational Engineering Research,* 08*(2),*39-41.

Swati Gupta (2015). A Regression Modeling Technique on Data Mining. *International Journal of Computer Applications,* 116 *(9),* 27.

Waidor, T.K. & Akpojaro J. (2018). The Use of classification algorithm for students' academic performance forecasting. *Proceedings of the 14th iSTEAMS International Multidisciplinary Conference, AlHikmah University, Ilorin, Nigeria,* 14, 153.

Waidor, T.K. & Akpojaro, J. (2019). The Use of Classification Algorithm for Forecasting the Academic Performance of Students of Biological Sciences, University of Africa, Toru-Orua. *African Scientist,* 20*(2),* 79.

Carla Silva & José Fonseca (2017). Educational Data Mining: a literature review. *Advances in Intelligent Systems and Computing*. DOI: 10.1007/978-3-319-46568-5_9

Web Resources

CRISP-DM. https://www.datascience-pm.com/crisp-dm-2/

Data Mining – Cluster Analysis.https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm#::text=Clustering%20is%20the%20process%20of,the%20labels%20to%20the%20groups.

Data Mining Concept. https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#DMCON004

Great Learning Team (2020).*Why using CRISP-DM will make you a better Data Scientist?*https://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist/

Jason Brownlee (2016). *Logistic regression for machine learning.* https://machinelearningmastery.com/logistic-regression-for-machine-learning/

Pavel B. *Survey of Clustering Data Mining Techniques.* https://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf

Dhiraj K, (2019, May 26). *Top 5 advantages and disadvantages of Decision Tree Algorithm.* https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a

Pavan Vadapalli, (2020, December, 11). *Naive bayes classifier: pros & cons, applications & types explained.* https://www.upgrad.com/blog/naive-bayes-classifier/

APPENDICES
Tables

```
Tester:     weka.experiment.PairedCorrectedTTester "
Analysing:  F_measure
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       1/9/21 12:29 PM
```

Table 1. Output of comparative Analysis of J48, NaiveBayes and OneR using F-measure

```
Tester:     weka.experiment.PairedCorrectedTTester "
Analysing:  Percent_correct
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       1/9/21 12:30 PM
```

Table 2. Output of comparative Analysis of J48, NaïveBayes & OneR classifiers using Percentage Correct

```
=== Stratified cross-validation Using Tree (J48) ===

=== Summary ===


Correctly Classified Instances         301               95.5556 %
Incorrectly Classified Instances        14                4.4444 %
Kappa statistic                          0.9412
Mean absolute error                      0.0306
Root mean squared error                  0.1319
Relative absolute error                 10.1067 %
Root relative squared error             33.9129 %
```

```
=== Stratified cross-validation using Bayes (NaiveBayes)===

=== Summary ===


Correctly Classified Instances         281               89.2063 %
Incorrectly Classified Instances        34               10.7937 %
Kappa statistic                          0.8577
Mean absolute error                      0.0621
Root mean squared error                  0.1815
Relative absolute error                 20.5211 %
Root relative squared error             46.6833 %
Total Number of Instances              315
```

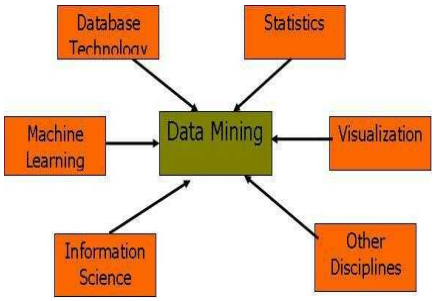

Figure 1. Multi-disciplinary pool feeding data mining (Keerthi, 2018).

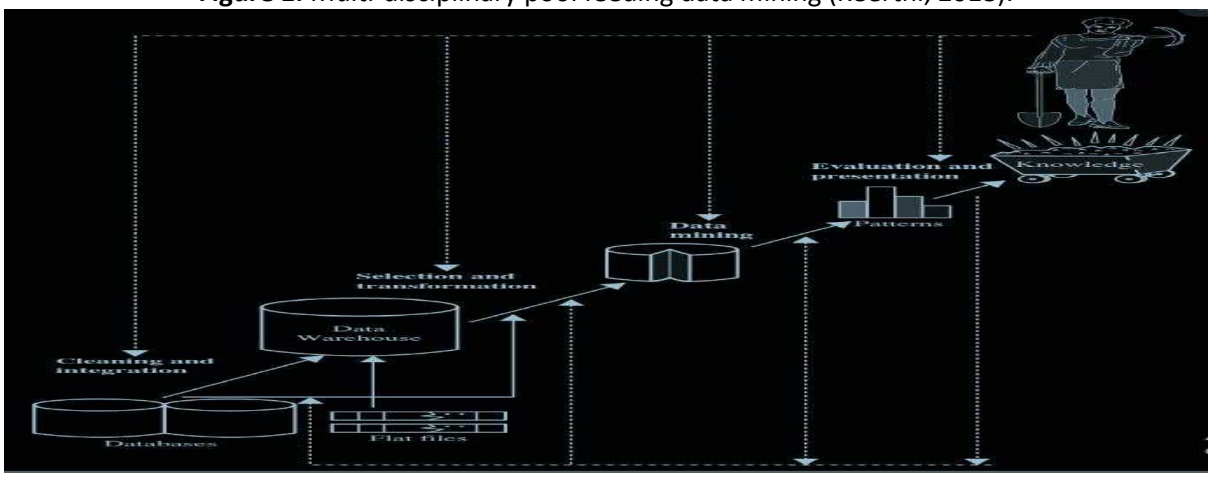

Figure 2. KDD (Akpojaro et al, 2019).

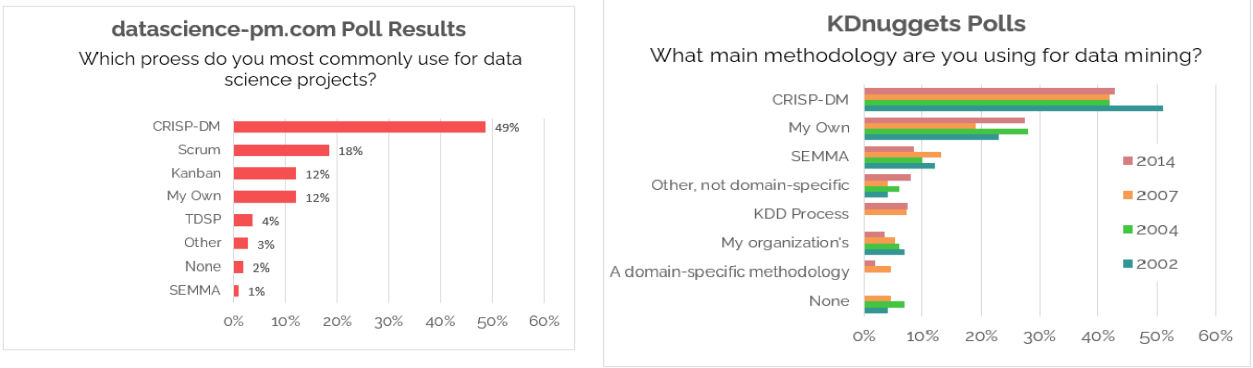

Figure 3 (A) popularity of CRISP-DM IN 2020, **(B)** Popularity of CRISP-DM over 12 year. Source: https://www.datascience-pm.com/crisp-dm-2/