International Journal of Academic Engineering Research (IJAER)
ISSN: 2643-9085
Vol. 3 Issue 12, December – 2019, Pages: 13-17

Analysis of Some Key Factors In School Results to Promote Excellence Using Big Data Classification Techniques

[1]Nzisabira Louis and [2]Abubakarsidiq Makame Rajab
[1]M&N Solutions 42, Revolution Street, P.O.Box: 1342, Bujumbura, Burundi.
[2]School of Electronic Information Communication, Huazhong University of Science and Technology, Hongshan District, Wuhan 430074, P.R. China
[1]Email:louisnzisabira@outlook.fr **and [2]Email:** mrgovery@hotmail.com

***Abstract*—**In this paper, we present one among many Big Data applications in real life. We explore classification using the J48 algorithm which implements the C4.5 decision tree for classification. We consider a study case on Burundi Schools of Excellence. We evaluate and analyze key factors in school results concerning promote excellence. In this we gain in deep comprehension how Big Data works, especially how classification can be useful for decision making in real life in daily activities including education. Big is useful in its different aspects and uses. Any institution can get its share in benefiting from science. During our case study in field, we have been given a basic grounding in Big Data Analysis and Methods to allow us to evaluate approaches. We here apply these theories and methods to decision problems and apply techniques to practical case studies in a problem-based learning.

Keywords—Big Data Analysis, Big data applications, Classification, C4.5, J48, Weka.

• INTRODUCTION

This document is a case study report. We here present a study case or problem we try to solve and presents results using useful techniques and the power of Big Data Analysis. The ability to access, analyze, and manage vast volumes of data while rapidly evolving the Information Architecture has long been a goal at many Higher Education institutions. Many have long standing data warehouses and have used analytics tools. As the competition for gifted students becomes more intense while the cost of education makes the pool of potential students more limited, many institutions are taking another look at how they are analyzing potential students and managing the experience that students have while they are enrolled. Among other needs of society and especially all educational systems, exists also the need to produce competitive well skilled students to develop the communities and its interests.

Analytics play a critical role in performing a thorough analysis of student and learning data to make an informed decision on future study offerings and their mix to cater to the potential and existing students. Big Data systems position IT to see the institution more holistically than any other areas for improved decision making. Predictive analytics or forecasting models in a Big Data environment enable institutions to make right investment decisions for higher institutional impact.
A Big Data based architecture enables the inclusion of a greater variety of data sources so that many different types of data can be analyzed. This, in turn, broadens the analytics and predictive options available and can lead to better management of the institution [1].

In this paper, apart preceding parts, summary is presented as follows: problem and its interests, some big data applications, presentation of our training sets, preprocessing, processing and classification results, interpretation of results, conclusion and some references. Some necessary definitions of some keywords will be given spontaneously.

• PROBLEM AND INTERESTS

- *Problem and context:*

In Burundi educational system, time to time pupils need to pass assessments. Some are national exams organized by the Government via the Ministry in charge of Education. Many factors influence the results obtained by pupils. And some of those factors are believed to be most determinant. We here evaluate and classify them.

In general, common people, parents and educational cadres think that place or social of origin of a candidate, attendance in class, school-regime, feeding of a candidate have much impact on results a pupil obtains at national exams [2].

- *Interests:*

Finding what is the most determinant and its impact can be helpful in designing solutions to increase rate of success to such exams. Data are about Burundi secondary school pupils/candidates to the secondary school final evaluation, here named the Exam of State or State Exam. Going at University is considered to be key to success in society [3] [4].

- *Methodology*

In this study, we consider some records from Ministry of Education on a Class of Excellence of 24 pupils. We here evaluate how 4 key factors selected among many others contribute to the results obtained by candidates. We finally evaluate if obtained model can help for further decisions.

III. SOME BIG DATA APPLICATIONS

- *Big data applications*

Different tools and algorithms are commonly used in Big Data Analysis to capture, curate, manage, and process data within a tolerable elapsed time [5].

Among many popular data mining scalable machine learning algorithms are performing clustering, classification, regression, and statistical modeling to prepare intelligent applications: association analysis, classification, clustering, statistical learning, bagging and boosting, sequential patterns, integrated mining, rough sets, link mining, and graph mining [6].

We hereby experience some tools and algorithms to understand, analyze and solve our problem. We will use classification.

- *Classification*

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance [7].The most explored technique here is classification. Our algorithm is the C4.5 as implemented by J48 in Weka 3.9.0 [8] [9].

- *C4.5, J48, Weka*

C4.5 implemented as J48 in some programs, is an algorithm used to generate a decision tree. We chose C4.5 because it is one of the most commonly used algorithms in the machine learning and data mining communities but also because it's recognized as a standard reference for many others [10]. Weka stands for Waikato Environment for Knowledge Analysis. It is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It provides a plug-in architecture for researchers to add their own techniques, with a command line and window interface that makes it easy to apply them to your own data [11].

• EXPERIMENT AND RESULTS

- *Preprocessing data.*

As we do not use the same data sources, data attributes, data tools, and algorithms all the time as all of them will not use data in the same format. Weka requires some files format. It can recognize .arff.csv, exp, .xrff, .xrff.zip, etc [13] [14]. Our code describes training data given in table above. To allow Weka to process it as one of its accepted file formats, we are using .arff format as we encoded under Notepad. This leads to the performance of data operations, such as data cleansing, data aggregation, data augmentation, data sorting, and data formatting, to provide the data in a supported format to all the data tools as well as algorithms that will be used in the data analytics. In simple terms, preprocessing is used to perform data operation to translate data into a fixed data format before providing data to algorithms or tools. The data analytics process will then be initiated with this formatted data as the input [12].

- *J48-C4.5*

We proceeded with J48 -C 0.25 -M 2 tree which is a clone of C4.5 classifier. Test option is "Use training set". So our test mode is evaluating on training set and the classifier model is full training set. This to invoke a Weka class, as we can do it using command line in java. This command tells the Simple CLI to load a class and execute it with any given parameters. E.g., the J48 classifier can be invoked on the iris dataset with the following command: java weka.classifiers.trees.J48-t c:/temp/ourflife.arff. We here use the graphical user interface

Table 1. Training Data Set Sample

TID	Origin	Meal	Regularity	School-Regime	Obtained Results
1.	City	Insufficient	No	Externa	Low

2.	City	Insufficient	No	Internal	Low
3.	City	Insufficient	Yes	Externa	Medium
4.	City	Insufficient	Yes	Internal	High
5.	City	Sufficient	No	Externa	Medium
6.	City	Sufficient	No	Internal	Medium
7.	City	Sufficient	Yes	Externa	High
8.	City	Sufficient	Yes	Internal	High
9.	Subcity	Insufficient	No	Externa	Low
10.	Subcity	Insufficient	No	Internal	Low
11.	Subcity	Insufficient	Yes	Externa	Medium
12.	Subcity	Insufficient	Yes	Internal	High
13.	Subcity	Sufficient	No	Externa	Low
14.	Subcity	Sufficient	No	Internal	Medium
15.	Subcity	Sufficient	Yes	Externa	High
16.	Subcity	Sufficient	Yes	Internal	High
17.	Village	Insufficient	No	Externa	Low
18.	Village	Insufficient	No	Internal	Medium
19.	Village	Insufficient	Yes	Externa	Medium
20.	Village	Insufficient	Yes	Internal	High
21.	Village	Sufficient	No	Externa	Medium
22.	Village	Sufficient	No	Internal	High
23.	Village	Sufficient	Yes	Externa	High
24.	Village	Sufficient	Yes	Internal	High

.

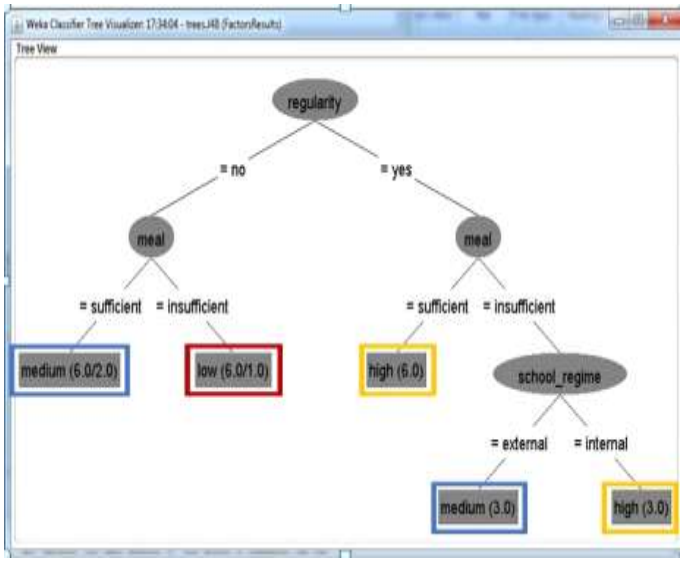
- *Processing and Classification Results*

As shown on the figure 1, we have a relation of 24 instances each having 5 attributes as defined in our code source in my_course_report.arff we provided to Weka software for classification and generating related decision tree. The results have been generated by our classifier under shown parameters on screenshots.

Generated tree has 2 leaves and a size of 9. Its description and composition of generated pruned tree are given in figure 2 here above. Decision Tree View is given below in figure 3.This figure 2 above reflects summary while using Test option is "Use training set". And we supply also a figure of test option "Cross-Validation Folds" having 10 as parameter. Results have no significant changes.

Classifier output

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

Correctly Classified Instances          21               87.5    %
Incorrectly Classified Instances         3               12.5    %
Kappa statistic                          0.8095
Mean absolute error                      0.1514
Total Number of Instances               24

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,833    0,056    0,833      0,833    0,833      0,778    0,847     0,736     low
               0,875    0,125    0,778      0,875    0,824      0,730    0,852     0,792     medium
               0,900    0,000    1,000      0,900    0,947      0,917    0,932     0,942     high
Weighted Avg.  0,875    0,056    0,884      0,875    0,878      0,820    0,884     0,840

=== Confusion Matrix ===

 a b c   <-- classified as
 5 1 0 | a = low
 1 7 0 | b = medium
 0 1 9 | c = high
```



- *Evaluation*

Results show three classes of results as impacted by other attributes. Origin of a candidate doesn't appear as it's considered to have no impact in candidate's results. As it appears in our results, impactful factors are regularity in attending classes, well or not well feeding which is meal, and school regime of a candidate; internal or external living. On a total of twenty-four instances, twenty-one instances are correctly classified while three are incorrectly classified. This is respectively 87.5% and 12.5%. In their decreasing order in importance of impact, we can see that it's respectively regularity, meal, and regime of the school. And these three classes of results (high, medium, low) formed by factors appear grouped as follows; regularity = no, meal = sufficient: medium (6.0/2.0), meal = insufficient: low (6.0/1.0) regularity = yes, meal = sufficient: high (6.0), meal = insufficient, school_regime = external: medium (3.0) and school_regime = internal: high (3.0).

Simply, the above generated model can be presented as in the Table 2 below. In other words, the proportions shown above, predictive model can be summarized by five sets as follows.

Table 2.Data Model Prediction Result

TID	Regularity	Meal	School-regime	Results	Cand.
1.	No	Insufficient		Low	6
2.	No	Sufficient		Medium	6
3.	Yes	Insufficient	External	Medium	3
4.	Yes	Insufficient	Internal	High	3
5.	Yes	Sufficient		High	6
Total of Candidates					24

We note that, the origin does not appear in the decision tree. Which means that, place of origin (city, sub city or village) of candidate is considered to have no significant influence on results a candidate obtain. On the other side, attendance and feeding are most determinant. Regularity in attendance and enough meal can contribute to achieve better which is obtaining high results.

The kappa statistic measures the agreement of prediction with the true class [15]. In our case we have 0.8095. Detailed accuracy by class, the average of precision is given in figures 4 and 5 respectively for different parameter of test option.

We do not pretend that results given by the classifier are totally perfect. However, the accuracy is 82.5% for our training set and we can be sure to correctly generate useful model of same approximate accuracy.

• CONCLUSION

Big data is able to create new profiles by using multiple data sets that effectively re-create an individual's information based on information obtained about others in the group that the individual is lumped in, or on faulty data associated with the individual in the first place[16].We here have experienced classification in big data analysis. We have obtained results. And this can be useful for making decisions. In our case study, to ameliorate results of candidate, the ideal could be having all of them attending regularly, feeding them enough and if possible having all of them as internal scholar. This can promote

excellence in results and making well skilled and successful candidates. However, as we can see. Some incorrectly classified instances occurred and this can affect future decisions. Other factors and different aspects not included here can be arose. Further researches can be done to achieve more. Human brain and its considerations still is important to achieve something better where machines can't allow to definitively give solutions.

• REFERENCES

- Robert Stackowiak, Venu Mantha, Art Licht, Oracle Enterprise Architecture White Paper – Improving Higher Education Performance with Big Data, April 2015, Redwood Shores, CA 94065, USA., p1. pp24.
- http://burundi-agnews.org/education/burundi-creation-de- 6-ecoles-secondaires-dexcellence-rentree-2016-2017/ [Accessed on December 20, 2017].
- http://www.iwacu-burundi.org/ecoles-dexcellence-une- excellente-idee-mais/ [Accessed on December 20, 2017].http://www.ibe.unesco.org/Countries/WDE/2006/SUB- SAHARAN_AFRICA/Burundi/Burundi.pdf [about objectives of educational system in Burundi, accessed on December 23, 2017].
- Jai Prakash Verma1 , Smita Agrawal1 , Bankim Patel2 and Atul Patel3, Big data analytics: Challenges and applications for text, audio, video, and social media data, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, February 2016.
- Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg, Top 10 algorithms in data mining, Survey Paper, © Springer- Verlag London Limited 2007.
- Alpaydin, Ethem.(2010),Introduction to machine learning. Massachusetts, USA: MIT Press, 2nd ed.
- Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse WEKA Manual for Version 3-9-0, April 2016, University of Waikato, Hamilton, New Zealand.
- Pete Warden, Big Data Glossary, A guide to a new generation data tools, 31, Published O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
- Peter Harrington, Machine Learning in Action, ©2012 by Manning Publications Co., Shelter Island, NY 11964, USA.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); the WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
- Vignesh Prajapati, Big Data Analytics with R and Hadoop, © 2013 Packt Publishing Ltd, Birmingham B3 2PB, UK. https://www.futurelearn.com/courses/more-data-mining-with weka/1/steps/161120 [Accessed on December 15, 2017]. https://weka.wikispaces.com/XRFF [Accessed on December 15, 2017].
- Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, David Scuse WEKA Manual for Version 3-9-0, April 2016, University of Waikato, Hamilton, New Zealand.