International Arab Conference on Information Technology (ACIT'2016)

Decision Making System for Scientific Research using Data Mining Algorithm

Mohamed Elmohadab[1], Belaid Bouikhalene [1] and Said Safi[1]
[1]Department of Mathematics and Computer Science Sultan Moulay Slimane University, Morocco

Abstract: *Recently, the governance of information system through the use of decision aid has posed a serious challenge for the leadership of universities. In this paper, we present a decision aid system for scientific research based on a set of different decision aid algorithms, through Naive Bayes, Decision Tree, and OneR. For illustrations we treat the case of a public university.*

Keywords: Scientific Research, data mining, decision aid, business intelligence, Odoo.

1. Introduction

In the broader context, governance means the way of leading, administrating or even piloting by the decision-maker. The decision making system is a process based on best practices related to the university's strategy.
Scientific research is one of the pillars of the university which require good governance. It's highly fragmented on the research centers and laboratories, the mobility of researchers, and the organization of scientific event. The good governance requires the modernization and rationalization of management information decision-aid assist system in administrative and managerial aspects.

The building of such system required the use of algorithms of advanced analysis particular the data mining algorithms.

In this work we contribute to the implementation of management information decision support system in a public university.

The paper has five parts. First section describes our system developed by the enterprise resource planning system Odoo. Second section gives a global view of the analysis of the system using UML and Bizagui programming languages. In the third section we will see the architecture of the proposed system. The fourth section discusses the mathematical and algorithmic modelling of this system and the different classifiers used. Full results and discussions are reported in the fifth section.

2. Presentation of the management of scientific research

2.1. Classic Method of Scientific Research Management

All researchers must fill a form in order to organize or to participate in scientific events. Old method was so classic: Each applicant was asked to complete the manual paper form, after filing form we send it to administration for submission.

The department of scientific research staff filled the information in the university database based on each researcher manuscript form.

As a result, this classic method is not practical anymore, because it requires more efforts in human and material resources, besides the risk of loss of forms if their numbers are important. Due to all this reasons, it comes the need to create a web application shared between researchers and University to manage all these tasks in a simple and efficient way.

2.2. Justification for the choice of open ERP Odoo

Enterprise Resource Planning (ERP) [4] is a category of business management software it's typically a suite of integrated applications that an organization can use to collect, store, manage and interpret data from many business activities.

Odoo is a suite of open source business apps that cover all needs. Its unique value proposition is to be at the same time very easy to use and fully integrated.

The implementation of an automatic system for managing scientific research is based on an open technology such as Odoo. The reasons for this choice are:

- Open Source
- Ease of use
- Flexible
- Easily interoperable
- Multiple language.

The system is developed by Odoo ERP based on MVC architecture, Python language and Xml. Concerning database, Odoo interacts with PostgreSQL, otherwise, Odoo security has two aspects:

- Authentication: is the act of confirming the truth of a user via LDAP or through two Factor Authentications.
- Authorization: is the process of giving someone permission to do or have something in Odoo. this roles are given by administrator.

Model–view–controller (MVC) is a software architectural pattern for implementing user interfaces on computers. It divides a given software application into three interconnected parts so as to separate internal representations of information from the ways that information is presented to or received from the user

Python is a widely used high-level, general-purpose, interpreted, dynamic programming language.

Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.

PostgreSQL is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards-compliance.

3. Modeling with UML

This section discusses the modelling of the new system and different actors in reaction.

3.1.First axis

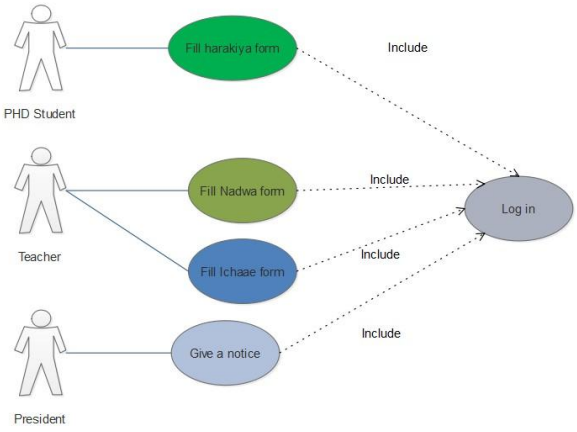

Fig. 1 Use case diagram of our system

Our use case contains the following actors: PhD Student, research professor, system administrator, and administrative staff.

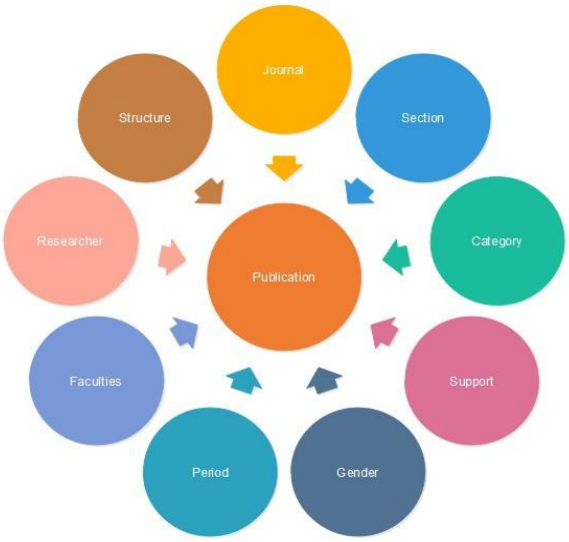

Fig. 2: Architecture of Publication

In the next figure we can discover the different actors involved in the system.

Fig. 3 Different intervener in the system

3.2.Second axis

A graphical representation of the "flow" of data through our information system, modelling its process aspects, is shown below:

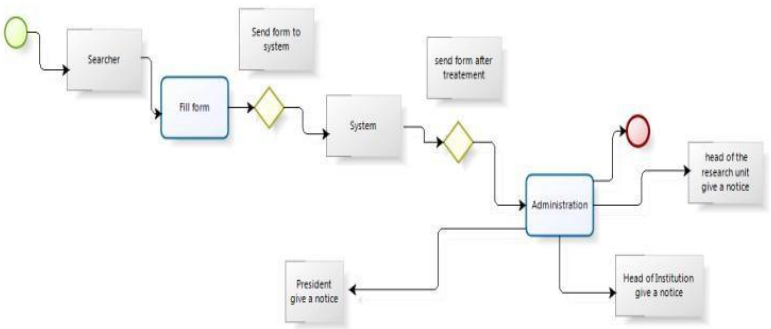

Fig. 4 Form processing process

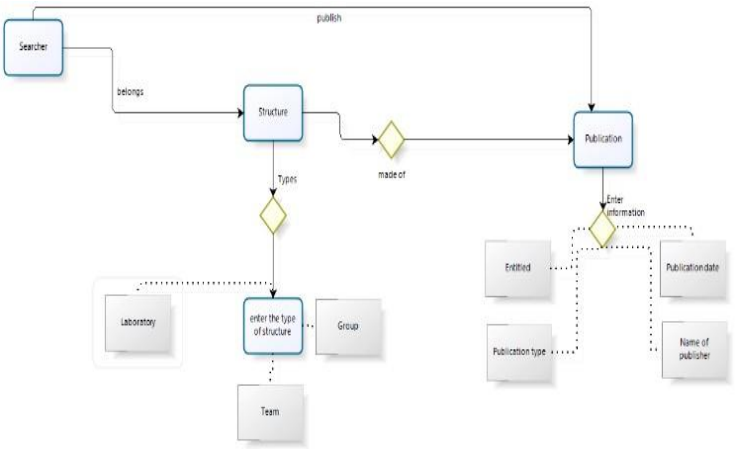

Fig. 5 research management processes

4. ARCHITECTURE OF THE SYSTEM

Thanks to the fast development of the new information and communication technologies, the university has firmly decided to facilitate the access and treatment for all process, especially in scientific research in order to help t administrative staff, PhD students and professors to provide with digital services that they need.

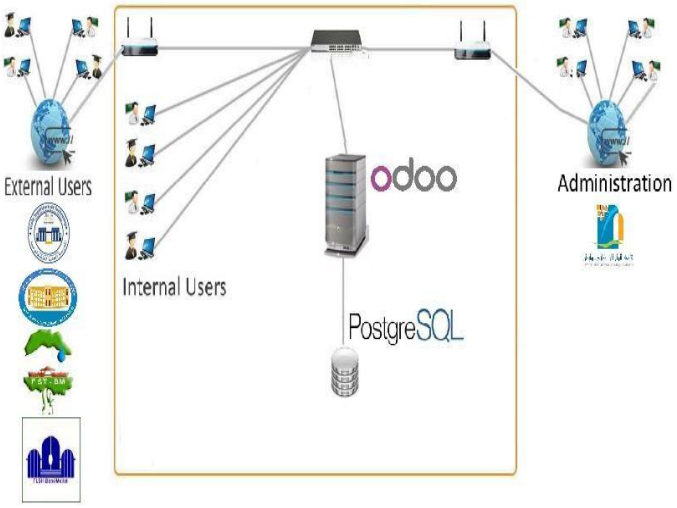

Fig.6 Architecture, technology and tools used in the system

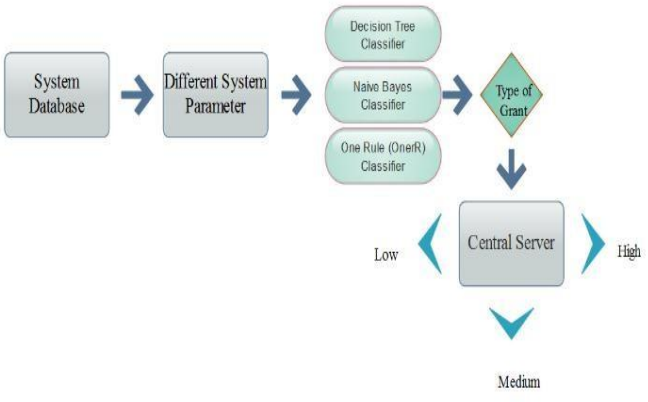

Fig. 7 Operating diagram for decision-makers

5. MATHEMATICAL AND ALGORITHMIC MODEELING OF THE SYSTEM

In our work, we choose to work with free software Weka, because it contains different classifiers in order to make decision for grants and funding. Firstly, we choose to work with Naïve Bayes classifier, decision trees classifier and OneR classifier. The three classifiers are highly efficient in evaluating a series of parameters to predict the forecast of an overall annual grant that the institution manages according to its needs.

- **Decision Tree classifier:**

The reasons for this choice are:

1. Classify correctly as much of the training sample as possible.
2. Generalize beyond the training sample so that unseen samples could be classified with as high accuracy as possible.
3. Be easy to update as more training samples become available.
4. Have as simple a structure as possible.

- **Naive Bayes classifier:**

The reasons for this choice are:

1. For each hypothesis: We associate a probability observation of one or several instances may change this probability.
2. We can talk about the most hypothesis likely, based on the conditional probabilities and Bayes rule.
3. Forecasting the future from the past, while assume independence attributes.
4. Bayesian probability is the estimation of an event knowing a preliminary hypothesis is verified (knowledge).

- **The OneR Classifier:**

The reasons for this choice are:

1. For each attribute and each value of this attribute, create a rule**.**
2. Counts how many times each class appears, and finds the most frequent class.
3. Creates a rule: attribute-value-> class.
4. Calculates the error rate of the rule.
5. Chooses the rules with the lowest error rate.

- **Mathematical Model**

Our Mathematical model can be modeled in terms of Set Theory. The proposed system is represented as follows:

S= {I, P, O}

where S= represent the proposed system,

I= {I1, I2, … , In}

Where I = inputs,

I1= input gives us number of new registered in the laboratory in the previous year.
I2= input gives us number of publications in the previous year.
I3=input gives us the events organized by laboratory us the year.
P = {P1, P2, … ,Pn}
Where P=Processes,
P1=check data in local server,
P2=store data at database,
O= {O1, O2, … , On}
Where O= output,
O1= required high grant (H).
O2= required medium grant (M).
O1= required low grant (L).

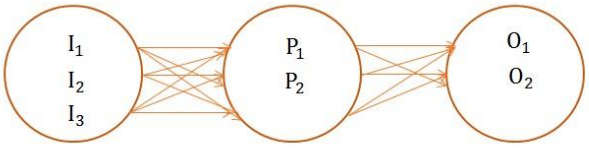

Fig. 8 Venn diagram of Proposed System

6. RESULTS AND DISCUSSIONS

6.1.Presentation

Business intelligence tools are a type of application software designed to retrieve, analyze, transform and report data for business intelligence. The tools generally read data that have been previously stored, often, though not necessarily, in a data warehouse or data mart.

To manage and measure the performance of scientific research at the university, managers or policy makers need synthetic indicators that are cleverly grouped one indicator that can offer for the leaders necessary tools to improve scientific research in SMS University.

We have identified an indicator to measure the evolution of the three procedures by the researcher in SMSU and their satisfaction and how that indicator can help to promote local scientific research.

On the other side we can measure the evolution of the number of researcher into each scientific structure as well as the number of publications and others indicators.

Among the statistics that we can evaluate:

1) Number of registered by structure.
2) Number of registered by faculty/college.
3) Number of registered per academic year.
4) The most productive structure.
5) The most active structure.
6) The most active faculty/college.
7) The number of forms of researcher mobility.
8) The number of dissemination and valorization of research.
9) The number of forms of organized scientific events.
10) Statistics on documents by type of publication.
11) Statistics on the number of new subscription/membership in each structure.
12) The evolution of publication deposits and patent filing in SMS University per academic year, per category and per author.

6.2.Statistical and graphical reports about SMS University

For simulation data, we illustrate our study by the case of Sultan Moulay Slimane University (SMSU). We use the database stored in the data center of SMS University. Firstly, we define the terminology used by the Scientific Research Committee at SMS University:

Ich'àae: This procedure is dedicated to the publication and dissemination of research works, support nature, information on the project leader, project description, etc.
Nadwa: This procedure is dedicated to the organization of scientific events, budgets, organizing committee, coordinator of the event, research entity organizing the event, etc.
Harakiya: This procedure is designed to manage the mobility of researchers and everything concerning their participation in scientific events and research internship.

Dashboards presented below were done until March 2016.

Type of activity	Number of actions
Harakiya	348
Ich'àae	250
Nadwa	70

Table1.Forms statics

Table 1 allows us to conclude that it has enough benefits of the different types of support in scientific research, especially in the Harakiya type that allows for the researcher mobility.

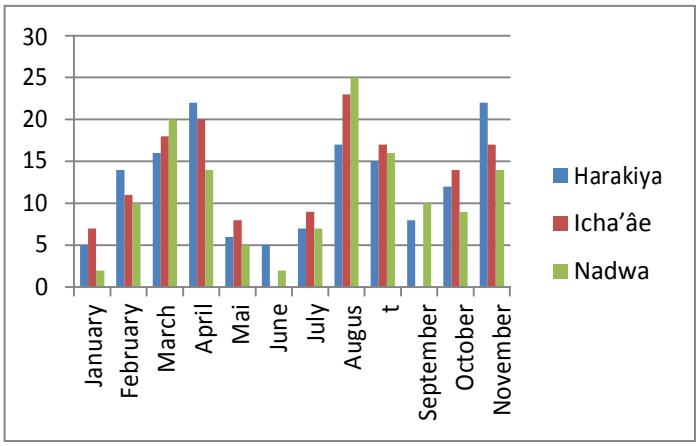

Fig .9 SRSMSU Platform, Homepage

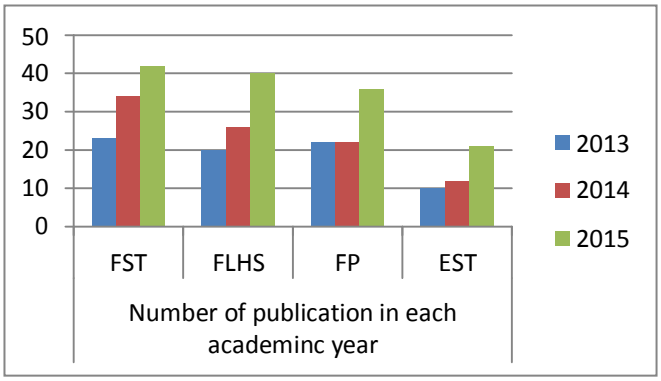

Fig .10 Numbers of publications per faculty and per year

The figure above shows that the FST, represented by its structure of research in different disciplines, has published the highest number of publications in all the university. This means that scientific research in this institution is adapted to the new Academic Policy.

Section	2012/2013	2013/2014	2014/2015
Chemistry Research	0	17	85
Computer Science Research	0	69	24
Economics Research	13	78	62
Law Research	0	86	54
Management Research	12	23	41
Mathematics Research	1	08	12
Physics Research	15	62	82
Total	41	343	360

Table 2. Number of publications by section

Table 2 shows that Chemistry structure is more productive than other structures annually, due to the dynamism of the professorial staff in all faculties of the university.

We can conclude that the number of publications in the FST Faculty is more than in others faculties (Fig. 9).

6.3.The use of decision support tools in our system

For this article we choose to work with the free software Weka [5]; it is a collection of machine learning/data software which contains tools for data pre-processing, learning algorithm and evaluation methods [6].

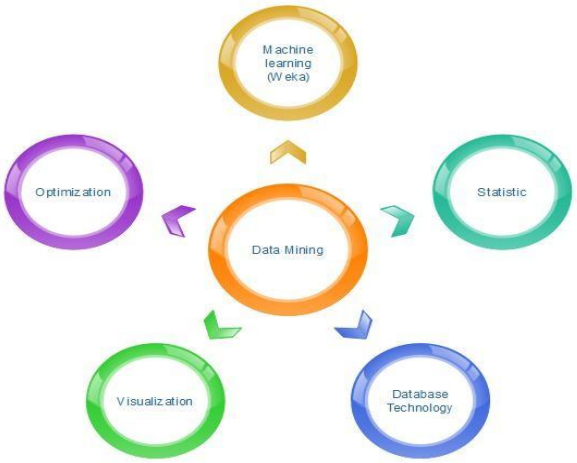

Fig .11 Data Mining axis

The table below shows our trained dataset based on previous research laboratory history:

Relation: SRSMU

No.	laboratory String	number-registred Numeric	number-publication Numeric	number-event Numeric	typ-grant Nominal
1	Mathématiques Physique Appliquées	12.0	10.0	3.0	medium
2	Sciences.Langage	13.0	16.0	5.0	high
3	Environnement.Santé	13.0	10.0	4.0	high
4	Modélisation.Ecoulements.Transferts	12.0	12.0	1.0	low
5	Recherche.Culture.Communication	13.0	8.0	2.0	low
6	Chimie.Organique.Analytique	16.0	15.0	8.0	high
7	Automatique.Conversion.Energie	14.0	10.0	3.0	medium
8	Gestion.Valorisation.Naturelles	9.0	13.0	0.0	low
9	Génie.Industriel	16.0	12.0	8.0	high
10	Physique.de.la.Matière.Nanotechnologie	16.0	13.0	3.0	medium

Fig .12 Trained Dataset

From the dataset in the table we want to take decision in order to award grants for the other laboratories.

Now we want to decide situation when the number of new registered in the laboratory in the previous year is 7, the number of publications in the previous year is 19 and the number of events organized is 5 which is not available into our trained dataset.

We will start by defining some parameters for our analysis:

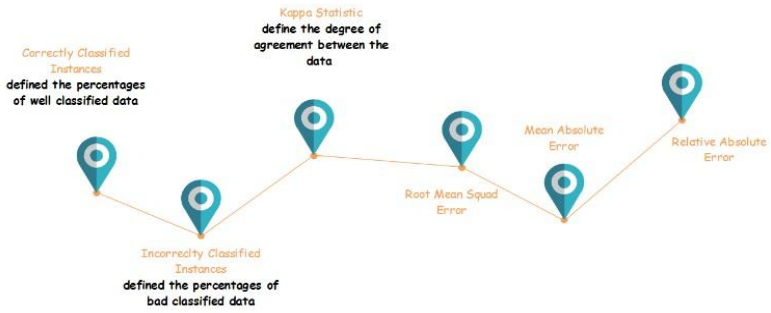

Fig .13 the keys parameters

⇨ Root Mean Squad Error

$$Root\ Mean\ Squad\ Error = \sqrt{\frac{(p_{1-a_1})^2 + \cdots + (p_{n-a_n})^2}{n}}$$

⇨ Mean Absolute Error

$$Mean\ Absolute\ Error = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$$

⇨ Relative Absolute Error:

$$= \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots + |a_n - \bar{a}|}$$

⇨ Root Relative squad Error:

$$= \sqrt{\frac{(p_{1-a_1})^2 + \cdots + (p_{n-a_n})^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}}$$

⇨ Precision: also called positive predictive value, it is the fraction of retrieved instances that are relevant.

⇨ Recall: sensitivity is the fraction of relevant instances that are retrieved.

⇨ F-Measures: A measure that combines precision and recall; it is the harmonic mean of precision and recall.

After the application of the two classifiers we got the following results:

- The application of J48 Algorithm

```
inst#,     actual, predicted, error, probability distribution
     1   2:medium      3:low       +   0       0       *1
     1   2:medium   2:medium           0      *1        0
     1   2:medium      3:low       +   0       0       *1
     1     1:high     1:high          *1       0        0
     1     1:high     1:high          *1       0        0
     1     1:high     1:high          *1       0        0
     1     1:high     1:high          *1       0        0
     1      3:low      3:low           0       0       *1
     1      3:low   2:medium       +   0      *1        0
     1      3:low      3:low           0       0       *1

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

Correctly Classified Instances           7               70      %
Incorrectly Classified Instances         3               30      %
Kappa statistic                          0.5455
Mean absolute error                      0.2
Root mean squared error                  0.4472
Relative absolute error                 41.8605 %
Root relative squared error             87.9883 %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1       high
                 0.333     0.143      0.5       0.333     0.4        0.595   medium
                 0.667     0.286      0.5       0.667     0.571      0.69    low
Weighted Avg.    0.7       0.129      0.7       0.7       0.691      0.786

=== Confusion Matrix ===

 a b c   <-- classified as
 4 0 0 | a = high
 0 1 2 | b = medium
 0 1 2 | c = low
```

- The application of Naïve Bayes Algorithm

```
inst#,     actual, predicted, error, probability distribution
     1   2:medium      3:low       +   0.306  0.006 *0.688
     1   2:medium     1:high       +  *0.939  0       0.061
     1   2:medium   2:medium           0.003 *0.977   0.02
     1     1:high     1:high          *1       0       0
     1     1:high   2:medium       +   0      *0.909   0.09
     1     1:high     1:high          *1       0       0
     1     1:high     1:high          *1       0       0
     1      3:low      3:low           0.336  0.001 *0.663
     1      3:low     1:high       +  *0.997  0.003   0
     1      3:low     1:high       +  *0.949  0       0.051

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

Correctly Classified Instances           5               50      %
Incorrectly Classified Instances         5               50      %
Kappa statistic                          0.2188
Mean absolute error                      0.3535
Root mean squared error                  0.5567
Relative absolute error                 73.9936 %
Root relative squared error            109.5249 %
Total Number of Instances               10

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.75      0.5        0.5       0.75      0.6        0.75    high
                 0.333     0.143      0.5       0.333     0.4        0.714   medium
                 0.333     0.143      0.5       0.333     0.4        0.571   low
Weighted Avg.    0.5       0.286      0.5       0.5       0.48       0.686

=== Confusion Matrix ===

 a b c   <-- classified as
 3 1 0 | a = high
 1 1 1 | b = medium
 2 0 1 | c = low
```

- The application of OneR Algorithm

```
=== Predictions on test data ===

inst#,     actual, predicted, error, probability distribution
     1   2:medium     1:high       +  *1       0       0
     1   2:medium     1:high       +  *1       0       0
     1   2:medium     1:high       +  *1       0       0
     1     1:high     1:high          *1       0       0
     1     1:high     1:high          *1       0       0
     1     1:high     1:high          *1       0       0
     1     1:high     1:high          *1       0       0
     1      3:low     1:high       +  *1       0       0
     1      3:low     1:high       +  *1       0       0
     1      3:low     1:high       +  *1       0       0

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

Correctly Classified Instances           4               40      %
Incorrectly Classified Instances         6               60      %
Kappa statistic                          0
Mean absolute error                      0.4
Root mean squared error                  0.6325
Relative absolute error                 83.7209 %
Root relative squared error            124.4342 %
Total Number of Instances               10
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.4       1         0.571      0.5     high
                 0         0          0         0         0          0.5     medium
                 0         0          0         0         0          0.5     low
Weighted Avg.    0.4       0.4        0.16      0.4       0.229      0.5

=== Confusion Matrix ===

 a b c   <-- classified as
 4 0 0 | a = high
 3 0 0 | b = medium
 3 0 0 | c = low
```

- The data for prediction:

```
@relation SRSMU

@attribute laboratory string
@attribute number-registred numeric
@attribute number-publication numeric
@attribute number-event numeric
@attribute typ-grant {high,medium,low}

@data

"Interdisciplinaire de Recherche en Sciences et Techniques",9,17,3,?
"Fins de la Révélation et de la Communication de Religieux et Culturel",5,9,1,?
"Recherche en Terme Littéraire et Linguistique",8,15,4,?
"Recherche sur la Culture et la  Communication",2,4,1,?
"Sciences du Langage",7,19,5,?
"Chimie Organique Analytique",16,15,8,?
"Automatique Conversion Energie",14,10,3,?
```

- The test set gives us the following prediction for J48 Algorithm:

```
=== Predictions on test set ===

inst#,     actual, predicted, error, probability distribution
     1          ?   2:medium      +   0      *1       0
     2          ?      3:low      +   0       0      *1
     3          ?     1:high      +  *1       0       0
     4          ?      3:low      +   0       0      *1
     5          ?     1:high      +  *1       0       0
     6          ?     1:high      +  *1       0       0
     7          ?   2:medium      +   0      *1       0
```

- The test set gives us the following prediction for Naïve Bayes Algorithm:

```
=== Predictions on test set ===

 inst#,     actual, predicted, error, probability distribution
     1          ?      3:low      +   0.107  0      *0.893
     2          ?   2:medium      +   0      *1       0
     3          ?      3:low      +   0.003  0      *0.997
     4          ?   2:medium      +   0      *1       0
     5          ?     1:high      +  *1       0       0
     6          ?     1:high      +  *1       0       0
     7          ?   2:medium      +   0      *1       0
```

- The test set gives us the following prediction for OneR Algorithm:

```
=== Predictions on test set ===

inst#,     actual, predicted, error, probability distribution
     1          ?     1:high      +  *1       0       0
     2          ?     1:high      +  *1       0       0
     3          ?     1:high      +  *1       0       0
     4          ?     1:high      +  *1       0       0
     5          ?     1:high      +  *1       0       0
     6          ?     1:high      +  *1       0       0
     7          ?     1:high      +  *1       0       0
```

	J48	Naive Bayse	OneR
Correctly Classified Instances	70%	50%	40%
Incorrectly Classified Instances	30%	50%	60%
Kapp Statistic	0.5455	0.2188	0
Mean Absolute Error	0.2	0.3535	0.4
Root Mean Squad Error	0.4472	0.5567	0.6325
Relative Absolute Error	41.8605%	73.9936%	83.7209%
Root Relative Squad Error	87.9883%	109.5249%	124.4342%

Tab3. The comparison of performance1

In order to define the algorithm [7] [8] which allows a good prediction, in this case the superior CCI(Correctly Classified Instances) is the CCI(J48)=70%. Furthermore, to have good results the error must be minimal; in this case J48 is the only algorithm which satisfies this condition. And finally we look at the value “Kapp Statistic” if it is superior, which means that we have a good quality of prediction; so our superior coefficient is that of J48 with a value equal to 0,5455.

	J48	Naive Bayes	OneR
Precision	0.7	0.5	0.16
Recall	0.7	0.5	0.4
F-Measures	0.691	0.48	0.229

Tab4. The comparison of performance2

In Table 4 we see that the best precision is for J48, and the sensitivity should be close to 1, so in this case J48 has the best Recall; and finally we look at F-Measures that combine Precision and Recall, which should also be close to 1 to define a good performance to the superior F-Measure is that of J48**.**

Therefore, according to the analysis of the parameters [9] , we conclude that the algorithm which allows a good prediction is that of J48.

Overall, the three proposed methods achieved the best performance. Specifically, the J48 classifier with the Correctly Classified Instance 70% and F-measure 69.1%.

Our example gets classified as ’High’ about the type of grant for the three classifiers.

Future Work

This project was a real opportunity to present a management system of scientific research in Moroccan public university, case of Sultan Moulay Slimane University in Beni Mellal, using the tools of artificial intelligence and also the data of scientific research imported from the Presidency of the university.

We conclude that the SRSMSU system minimizes efforts for the presidency staff and solves many problems about Nadwa, Ich'àae, Harakiya.

In the near future, we will propose a new method which will propose another new improvement of scientific research.

Acknowledgments

The authors thank very much the Vice-President responsible for Scientific Research and Cooperation and the Scientific Research Committee at SMS University for their support and helps, we thank them very much.

References

[1] Volkova V. N, Loginova A. V, Shirokova S. V, Kozlovskay E. A, «Information Support of Scientific Researches in Russia».

[2] «Morocco's strategy for the development of Research 2025» Ministry of National Education, Higher Education, Staff Training and Scientific Research, November 2009.

[3] Gašpar V, Madarász L, Andoga R, Glodová I, «On Scientific Research Information System» chez 14th IEEE International Symposium on Computational Intelligence and Informatics, Budapest, Hungary. 19–21 November, 2013 .

[4] Ganesh A, Shanil K.N, Sunitha C, Midhundas A, M«OpenERP/Odoo – An Open Source Concept to ERP Solution» chez IEEE 6th International Conference on Advanced Computing, 2016.

[5] WEKA, the University of Waikato, [En ligne]. Available: http://www.cs.waikato.ac.nz/ml/weka/..

[6] Zdravko M, Russel I ,«An Introduction to the WEKA Data Mining System».

[7] Lakshmi Devasena C, Sumathi T, Gomathi V V, and Hemalatha M, «Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set» Special Issue, vol. 1, December 2011.

[8] Ben Amor N, Salem Benferhat S, Elouedi. Z «Naive Bayes vs Decision Trees in Intrusion Detection Systems» in ACM Symposium on Applied Computing, 2004.

[9] Phanupong M, Natthakan I and Tossapon B, «Prediction of Student Dropout Using Personal Profile and Data Mining Approach».

Authors Profile

Mohamed ELMohadab is a PhD student in laboratory of Interdisciplinary Laboratory for Research in Science and Technology, Polydisciplinary faculty, University Sultan Moulay Sliman (USMS), Beni Mellal, Morocco. He has obtained a MS degree in Computer and System Engineering from Sultan Moulay Slimane University.

Belaid Bouikhalene obtained a Ph.D. degree in Mathematics in 2001 and a degree of Master in Telecommunication and Computer science in 2007 from the University of Ibn Tofel Kenitra, Morocco. He is currently a professor at University Sultan Moulay Slimane, Morocco, His research focuses on mathematics and applications, decision information systems, e-learnig, pattern recognition and artificial intelligence.

Said Safi was born in Beni Mellal, Morocco in 1971, received the B.Sc. degree in physics (option Electronics) from Cadi Ayyad University, Marrakech, Morocco in (1995), MS and Doctorate degrees from Chouaib Doukkali University and Cadi Ayyad University, Morocco, in 1997 and 2002, respectively. He has been a professor of information theory and Telecommunication systems at the National School for applied Sciences, Tangier Morocco, from 2003 to 2005. Since 2006, he is a professor of applied mathematics and programming at the Polydisciplinary Faculty, Beni Mellal Morocco.