Indonesian Journal of Electrical Engineering and Computer Science
Vol. 14, No. 2, May 2019, pp. 1025~1029
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v14.i2.pp1025-1029

A data mining process using classification techniques for employability prediction

Saouabi Mohamed, Abdellah Ezzati
University Hassan the 1st, FST, LAVETE Laboratory, Morocco

Article Info

Article history:

Received Sep 23, 2018
Revised Nov 11, 2018
Accepted Dec 15, 2018

Keywords:

Classification
Data mining
Data mining process
Employability
Rapid Miner

ABSTRACT

The use of the data mining has become wider today; it can be applied in several fields like marketing, customer relationship management, medicine, engineering, etc. It can be used also in employability, the use of data mining in this field will give opportunities and solution for decision makers in this field in order to improve the employability and propose solutions. In this paper, we propose a data mining process for employability data using classification techniques, presenting in details all the phases in the process and what should be done in every phase. We used Rapid Miner Studio Educational Version 8.1.000, using an employability dataset.

Corresponding Author:

Saouabi Mohamed,
Department of Electrical and Computer Engineering,
FST, LAVETE Laboratory, Settat, Morocco,
University Hassan the 1st,
BP : 577, Route de Casa, Settat, Morocco.
Email: mohamed.saouabi@gmail.com

1. INTRODUCTION

Data mining is about using data analysis tools to find out new unknown knowledge, hidden relationships between the large data set we have on hands. These tools can handle mathematical algorithms, statistical models and machine learning methods. Data mining it's not just about collection and managing the data, it includes data analysis and prediction.

In this paper, we'll present the whole data mining process in details, and all the phases in this process. Reasearchers may find it quite difficult to come up with a data mining process and following it in order to resolve a solution. We proposed a data mining process and we took as example employability data to apply on the process. For that, we used Rapid Miner Studio Educational Version 8.1.000, using an employability dataset.

2. RESEARCH METHOD

In this phase, we present the data mining tool we used explaining whi we choose this tool. Here we present a brief overview about rapid miner and few other tools dedicated for data mining in order to choose which tool to work with.

Kalpana Rangra and K.L. Bansal [1] presented a comparative study of data mining tools, in order to present the advantages and the limitations for each, and the result of this comparative have shown that the choice depends on the nature of the experiment needed. For example, Weka is for people who are highly skilled, because it is very robust with built-in features and offers additional functionalities. Rapid Miner and Orange are dedicated for advanced users, particularly in the hard sciences, because it requires additional programming skills, and the limited visualization support. Rapid Miner is the only tool which is independent of language

limitation and has statistical and predictive analysis capabilities, so it can be easily used and implemented on any system, also it integrates maximum algorithms of any other mentioned tools and more important, it can be used for big data, as shown in Table 1.

Table 1. Data Mining Tools Description

Tool Name	Type	Advantages	Limitations
Rapid Miner	Statistical analysis, data mining, predictive analytics	Visualization, Statistical, Attribute Selection, Outlier detection, parameter optimization	Requires prominent knowledge of database handling
Orange	Machine learning, Data mining, Data visualization	Better debugger, Shortest scripts, poor statistics, suitable for novoice Experts	Big installation, Limited reporting capabilities
R	Statistical Computing	Purely statistical	Less specialized for data mining, requires knowledge of array language
Weka	Machine Learning	Ease of use, can be extended in RM	Poor documentation, weak classical statistics, poor parameter optimization, weak csv reader

3. RESULTS AND DISCUSSION

In this part, we presented the data mining process we propose. Data mining is an iterative process that typically involves the described phases bellow. Figure 1 shows the data minig general process.

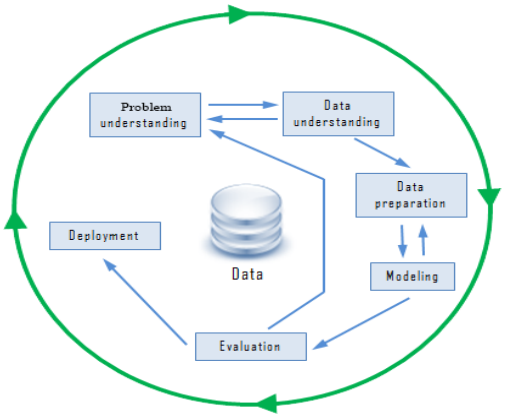

Figure 1. Data mining general process

Here presented below the description of the different phases of the data mining process.

3.1. Problem understanding

A data mining project starts with the understanding of the problem. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required of course.

3.2. Data exploration or Data understanding

Collect the data: We collected the data used in our experiment from a survey of employability conducted by Hassan the 1st University in 2016 in partnership with the National Evaluation Office (NEO) under the Higher Council for Education.

Describe the data: We need to understand the meaning of the data we have. So, we explore this data. And then, we identify quality problems of the data so we can fix it in the next phase, which is the preparation phase.

Data preparation: In this phase, we collect, cleanse, and format the data because some of the mining techniques accept data only in a certain format. We can also create new derived attributes if needed, for example, an average value.

Data selection: Here, we select the data we need in order to answer the problem in hand, so we explore

the data and we select only the data we want to use in our classification problem.

Remove columns: We remove all the irrelevant columns that have nothing to do with our classification problem, such us name, phone, etc.

Replace missing values: Missing values lead to inaccurate model generation, which mean false results which we cannot rely on, so we remade to this problem through imputation techniques, or we remove the instances completely.

Reorder attributes: We re-order the attributes of the data set, just for better organization of the data. Now after these steps, the final data contains 1208 instances of 13 attributes; we'll give some graph examples visualizing the data.

This graph represents the number of graduates graduated from different universities, as we can see FST has more graduates. Figure 2 is shown in Number of graduates graduated from different universities. This graph illustrates the number of graduates grouped by grade is shown in Figure 3.

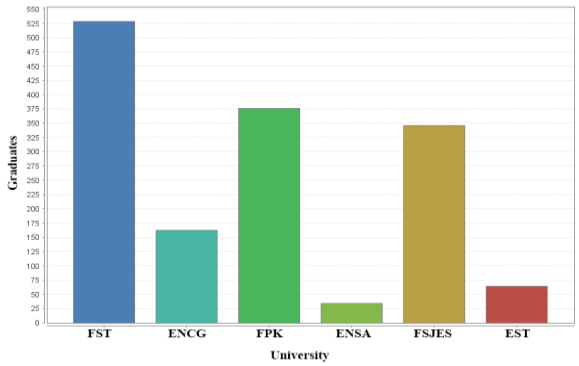

Figure 2. Number of graduates graduated from different universities

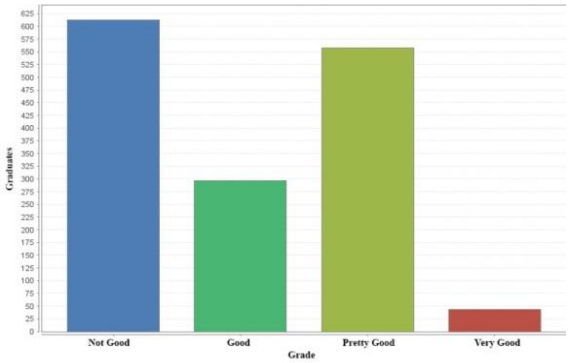

Figure 3. Number of graduates grouped by grade

Figure 4 shows the graduates' employability of every university, we can see the universities whom have more graduates working, and not working.

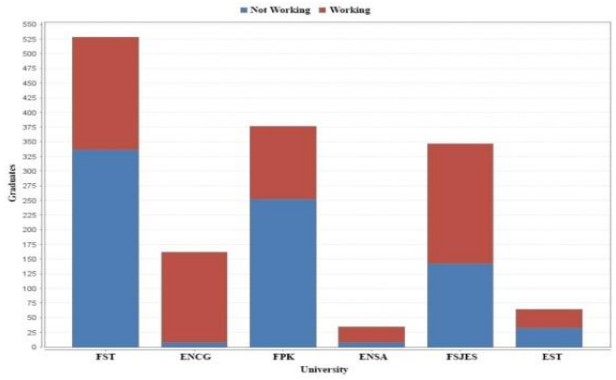

Figure 4. Graduates' employability of every university

3.3. Modeling

The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

Define target: Now, we define the target, which means the variable we want to predict; employability is our variable, with two classes, working and not working.

Which model should we apply? (Decision Tree, Logistic regression and Naïve Bayes): Based on the type of the data we have and the type of the variable we want to predict, we choose the models we will apply during this phase.

We will work with a supervised learning, since we have a target, which is the variable to predict, employability, with two classes: working and not working. The type of our variable to predict is qualitative, containing two classes, which mean the type of the variable is binary; we will work with a classification technique. We will apply the classification algorithm Decision Tree, Logistic regression and Naïve Bayes, we will assess each model, and finally we choose which model present to best and accurate model by evaluating the models' performances in the next phase.

Split Data: Now we need to apply the models, Decision Tree, Logistic regression and Naïve Bayes, we will split the data into two sets, training data with 80% and a test data with 20%. In the training data, we build the models and then applied on the test data. Because the test data has never been seen by the model, so the performance will be a good guide to what will be seen when the model is applied to unseen data.

3.4. Evaluation

Now, we will evaluate the strengths and weaknesses of our classification algorithms. In this phase, we evaluate the models. If the models do not satisfy the expectations, we go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved.

Confusion Matrix: A binary classification makes predictions where the outcome has two possible values: positive and negative. Moreover, the prediction for each example may be right or wrong, leading to a 2x2 confusion matrix as explained in Table 2.

Table 2. 2X2 Confusion Matrix

	Positive Class	Negative Class
Predicted Positive Class	True positive (TP)	False negative (FN)
Predictive Negative Class	False positive (FP)	True negative (TN)

TP: the number of "true positives", positive instances that have been correctly identified by the algorithm.
FP: the number of "false positives", negative instances that have been incorrectly identified by the algorithm.
FN: the number of "false negatives", positive instances that have been incorrectly identified by the algorithm.
TN: the number of "true negatives", negative instances that have been correctly identified by the algorithm.

Metrics: Different metrics can be used in order to evaluate the models' performance and to choose which provide the best accurate model: Accuracy, classification error, recall, kappa statistics, f-measure, sensitivity, precision, ROC curve and the time to build the model. His is an example of a roc curve comparison between two classifiers.

Based on ROC curve, it clearly illustrates that the Decision Tree classifier is more accurate than Logistic regression and Naïve Bayes, the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the model, which means that the decision tree model is more accurate.

As result of the evaluation phase, we evaluate the models, Decision tree, Logistic regression and Naïve Bayes, and we choose the model with the best accurate model based on the different metrics. Figure 5 is shown in ROC comparison between the classifiers.

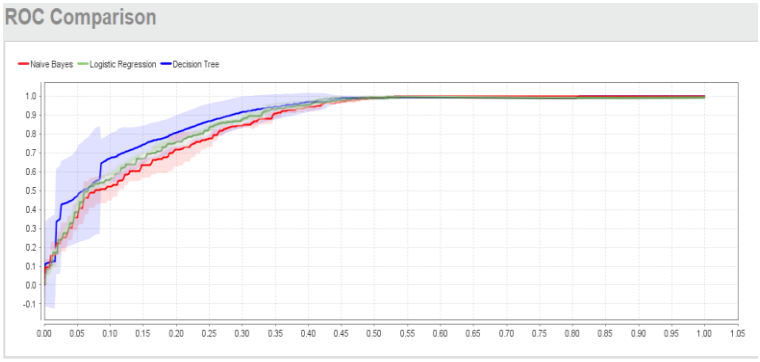

Figure 5. ROC comparison between the classifiers

Variables predicting graduates: In this step, we present the variables which have an important role predicting graduates' employability.

3.5. Deployment

After all these phases, this is the final phase, where we use methods for integrating data mining discoveries into use. The experts and decision makers will be able to use the results by exporting into databases or applications for reporting and reviewing.

4. CONCLUSION

Data mining offers many techniques in order to discover hidden patters between the data. These hidden patterns can be used to predict future behavior. In this paper, we presented the process of data mining from data collection phase to the deployment phase using Rapid Miner Studio Educational Version 8.1.000, using an employability dataset. We also presented a brief overview about Rapid Miner and few other tools dedicated for data mining in order to choose which tool to work with.

REFERENCES

[1] Comparative Study of Data Mining Tools - Kalpana Rangra Dr. K. L. Bansal
[2] https://orange.biolab.si/
[3] http://www.rdatamining.com/
[4] http://rapidminer.com/
[5] https://www.cs.waikato.ac.nz/ml/weka/
[6] Handbook of Data Mining - Ye, Nong
[7] Data Mining and Statistics for Decision Making - Tufféry, Stéphane
[8] Handbook of Statistical Analysis and Data Mining Applications - Nisbet, Robert,Elder, John, IV,Miner, Gary
[9] Distribution patterns of energy consumed in classified public buildings through the data mining process - Yibo Chena,b , Jianzhong Wu
[10] A new web-based solution for modelling data mining processes - Viktor Medvedeva, Olga Kurasovaa, Jolita Bernatavicien, Povilas Treigys, Virginijus Marcinkevicius, Gintautas Dzemyda
[11] Extract Business Process Performance using Data Mining - Faisal M. Nafie, Mergani.A.Eltahir
[12] A Data Mining & Knowledge Discovery Process Model - Óscar Marbán, Gonzalo Mariscal and Javier Segovia
[13] Applying process mining techniques in software process appraisals - Arthur M. Vallea,b, Eduardo A.P. Santos, Eduardo R. Loures
[14] A survey of Knowledge Discovery and Data Mining process models – Lukasza Kurgan and Petrmusilek
[15] Using Semantic Lifting for improving Process Mining: a Data Loss Prevention System case study - Antonia Azzini, Chiara Braghin, Ernesto Damiani, Francesco Zavatarelli
[16] Applying process mining techniques in software process appraisals - Arthur M. Vallea,b, Eduardo A.P. Santos, Eduardo R. Loures