*International Journal of Engineering & Technology*, 7 (4.32) (2018) 17-20

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*

***Research paper***

# Data Mining Techniques for Predicting Employability in Morocco

**SAOUABI Mohamed[1]*, EZZATI Abdellah[2]**

*[1]University Hassan the 1st, FST, LAVETE Laboratory*
*[2]University Hassan the 1st, FST, LAVETE Laboratory*
**Corresponding author E-mail: mohamed.saouabi@gmail.com*

**Abstract**

One of the biggest challenges for Big Data applications is to explore large volumes of data and extract valuable information and knowledge for future actions. Employment is the main form of social integration, a factor in improving living conditions and preventing risks of poverty and vulnerability and the most appropriate indicator for assessing the level of social cohesion in a country. Mining employability data will give decision makers a great view of the data and opportunities to make improvement in this sector. In this paper, we presented an experimental study comparing various classification data mining algorithms on employability data in Morocco, which are Decision tree, Logistic regression and Naïve Bayes, which take place in the top 10 data mining algorithms identified by the IEEE International Conference on Data mining. The objective in our experiment is to choose the most efficient and suited algorithm for the employability data.

***Keywords***: *Data mining, Big Data, Employability, Classification, Decision tree, Logistic regression, Naïve Bayes.*

## 1. Introduction

Quite simply, the big data era is in full force today because the world is changing, thanks to advances in communication technologies, people and things are increasingly interconnected and not just part of the time, but almost all the time. People are using more and more social networks, connected objects such as smartphones, vehicles with location sensors, it creates a lot of data -Big Data- that traditional tools and technologies cannot process and analyze.
But storing this amount of data is not the major problem; we need to use this data in order to extract useful information which can be used by decision makers. Data mining can do that, transforming this huge amount of data into valuable information that can be used.
In this experiment, we used RapidMiner Studio Educational Version 8.1.000 in Hadoop, it is used to implement machine learning algorithms, and it includes also Weka extension to implement algorithms designed for Weka mining tool.

## 2. Related Works

Data mining now is used in many fields such as employability, previous works have been done to explore and compare data mining classification algorithms. Few of the related works are listed below.
**M.venkatadri, Lokanatha and C. Reddy [1]** presented a comparative study of different data mining classifications with their limitations, and also they evaluated their performance with experimental analysis based on sample data, although real collected data would be better for mining the data and performing the models created, this collected data sure will have missing values and preparing it in the beginning will be difficult, but the results will be real and it can be used in real life.
**Pooja Thakar, Anil Mehta and Manisha [2]** proposed an empirical study that compares varied classification algorithms on two datasets of MCA (Masters in Computer Applications) students collected from various affiliated colleges of a reputed state university in India.
**Muskan Kukreja, Stephen Albert Johnston and Phillip Stafford [3]** used several classification algorithms to analyze data. They found that Naïve Bayes is far more useful than other widely used methods due to its simplicity, robustness, speed and accuracy. In our paper, we worked on employability data, so performing classification techniques is necessary to find out which is the best for this particular type of data, because every data has his own specificity, and the performance of an algorithm can be different from a specific data to another.

## 3. Experimental Study

In this study, we used Rapid Miner Studio Educational Version 8.1.000, using an employability dataset. Based on the type of data we have and the type of the variable we want to predict, we used classification algorithms, to classify graduates into "working" and "not working". We compared various classification data mining algorithms, which are Decision tree, Logistic regression and Naïve Bayes, and then we chose the most efficient and suited algorithm for the employability data.

### 3.1. Data collection

The data used in this study is collected 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, Training and Scientific Research. Data is large, multivariate, incomplete, heterogeneous and unbalanced in nature. So in the next phase, we will prepare the data and we will clean it, so it can be ready for implementing the classification algorithms.

### 3.2. Data preparation

Preparing the data to suit a data mining task is a crucial phase, and it's the most time-consuming part of the process, so data need to be prepared, the data contains 1752 rows and 22 attributes, but this data needs cleaning and removing the non-pertinent data, like name, phone number, email, etc. Also transforming data and creating new attributes, calculated attributes for example. Final data contains 1208 instances of 13 attributes, here below in Table 1 the list and the description of the attributes.

**Table 1:** The dataset attributes with description

| No. | Attribute | Description |
|---|---|---|
| 1 | Gender | Gender of the graduates |
| 2 | Diploma | Type of the diploma of the graduates |
| 3 | Field | Field of study |
| 4 | Grade | Which grade in the baccalaureate |
| 5 | University | Which university the graduate graduated from |
| 6 | PracticeLevel | The graduate level of practice in his field of study |
| 7 | InformaticLevel | The graduate level of practice in information technology |
| 8 | FrenchLevel | The graduate French level |
| 9 | EnglishLevel | The graduate English level |
| 10 | BaccalaureateSerie | The graduate baccalaureate option |
| 11 | TrainingPeriod | Did the graduate made any training |
| 12 | TheoreticalLevel | The graduate level of theory in his field of study |
| 13 | Employability | Is the graduate working or not working |

### 3.3. Implementation of the classification algorithms

Now, after the data collection and preparation, we implemented the classification algorithms, Decision tree, Logistic regression and Naïve Bayes.

**Table 2:** The class distribution

| Class | Distribution |
|---|---|
| **Working** | 586 (49%) |
| **Not working** | 622 (51%) |

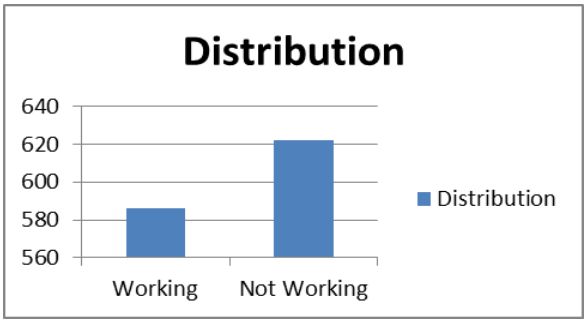


**Fig. 1:** Graph of the class distribution

We used different metrics to compare between Decision tree, Logistic regression and Naïve Bayes:
Accuracy, classification error, Recall, kappa statistics, F measure, sensitivity, precision, ROC (receiver operating characteristic) and the time to build the model, here's a description in **Table 4** of the different metrics we used.

**Table 3:** Confusion matrix for binary classification

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

**Table 4:** Metrics for classification evaluations

| Metrics | Formula | Evaluation Focus |
|---|---|---|
| **Accuracy** | $\frac{tp+tn}{tp+fp+tn+fn}$ | In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated. |
| **Classification error** | $\frac{fp+fn}{tp+fp+tn+fn}$ | Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated. |
| **F measure** | $\frac{2*p*r}{p+r}$ | This metric represents the harmonic mean between recall and precision values. |
| **Sensitivity** | $\frac{tp}{tp+fn}$ | This metric is used to measure the fraction of positive patterns that are correctly classified. |
| **Precision (p)** | $\frac{tp}{tp+fp}$ | Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class. |
| **Recall (r)** | $\frac{tp}{tp+tn}$ | Recall is used to measure the fraction of positive patterns that are correctly classified |

## 4. Results and discussion

### 4.1. Results

After applying Decision Tree and Naïve Bayes, Table 5 describes the results of this experiment.

**Table 5:** Results of the performance comparison of the classifiers

| Algorithm | Accuracy (%) | Precision | Classification error (%) | Recall (%) | Kappa statistics | F measure | Sensitivity (%) | Time to build (ms) |
|---|---|---|---|---|---|---|---|---|
| Decision Tree | 81.70 | 77.06 | 18.30 | 92.92 | 0.631 | 0.84 | 92.92 | 390 |
| Logistic regression | 80.79 | 82.24 | 19.21 | 80.13 | 0.616 | 0.81 | 80.13 | 344 |
| Naïve Bayes | 78.23 | 73.58 | 21.77 | 90.20 | 0.561 | 0.81 | 90.20 | 47 |

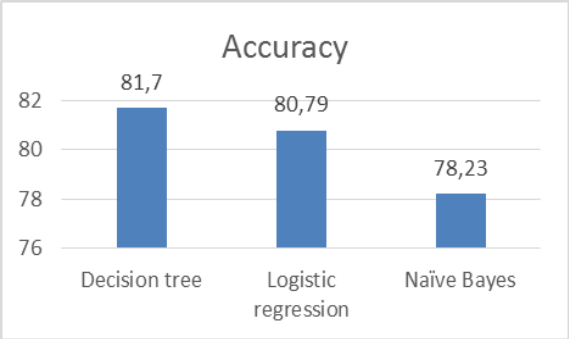


**Fig. 2:** Graph prediction accuracy of classifiers

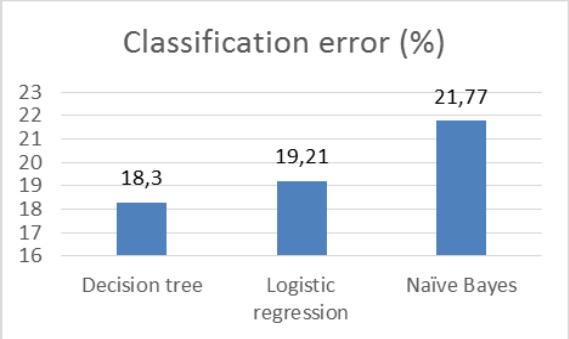


**Fig. 3:** Classification error of the classifiers

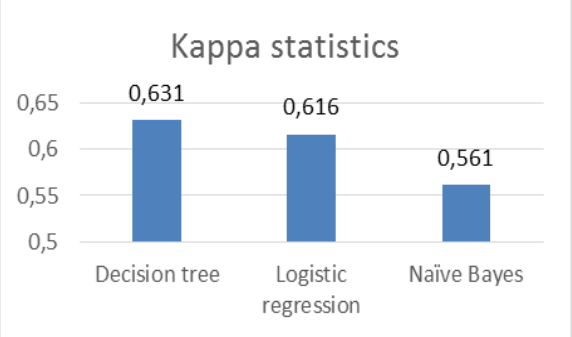


**Fig. 4:** Graph of kappa statistics

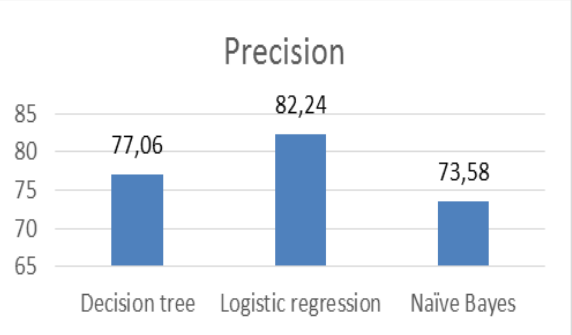


**Fig. 5:** Graph of prediction precision of classifiers

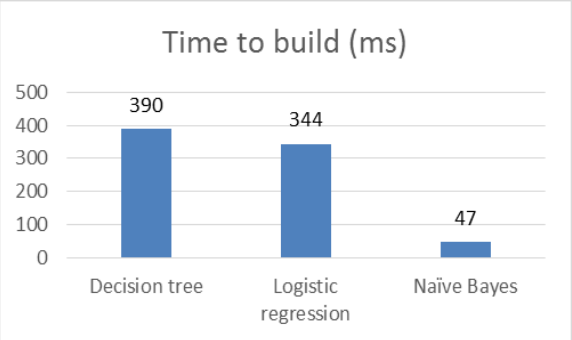


**Fig. 6:** Graph of time to build the models (ms)

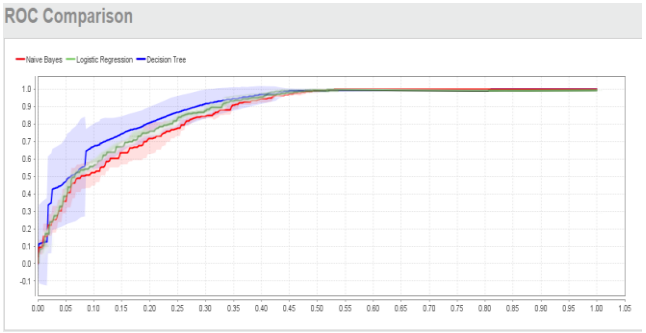


**Fig. 7:** Roc comparison of the classifiers

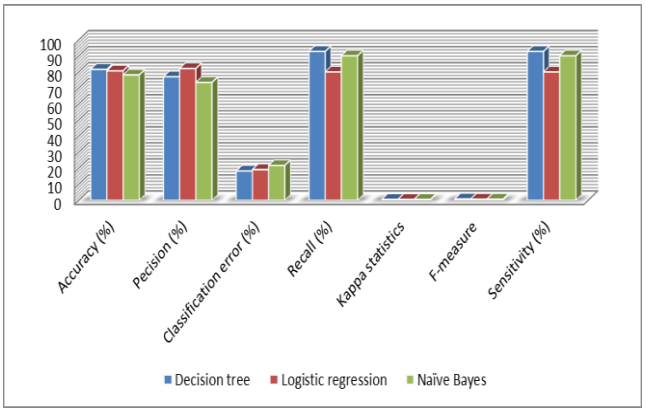


**Fig. 8:** Performance comparison of the three classifiers using the different metrics

### 4.2. Discussion

In our prediction experiment using the classification algorithms Decision Tree, Logistic regression and Naïve Bayes, we have used different metrics in order to compare and choose the most efficient algorithm for employability data, such as Accuracy, Classification error, Recall, F Measure, Kappa statistics, Sensitivity, precision, ROC and time to build the model.

Accuracy represents the percentages of instances correctly classified by the algorithm, based on the results, Figure 2 shows that the accuracy of the model predicted by decision tree (81.70%) is more accurate than the Logistic regression (80.79) and Naïve Bayes (78.23%), which mean also that the classification error of the Decision tree (18.30%) is lower than the Logistic regression (19.21%) and Naïve Bayes (21.77%), who has more unclassified instances. Also, Kappa statistics have shown that the Decision tree model (0.631) is better than the Logistic regression (0.616) and Naïve Bayes (0.561). Based on Cohen interpretation suggestion for Kappa result, Decision tree model and Logistic regression models are substantial, and Naïve Bayes model is moderate.

Sensitivity presents the correctly predicted positive observations to all observations in actual class; the different between decision tree and Naïve Bayes is not very large, 92.92% for Decision Tree and 90.20% for Naïve Bayes, and 80.13% for Logistic regression. Another important metric is F-measure, it tells how precise the classifier is, how many instances are classified correctly, as well as how robust it is, again the results shows that Decision tree is classified correctly with 0.84 F-measure rate, against Logistic regression with 0.81 and Naïve Bayes with 0.81.

Another metric is precision, results shows that Logistic regression has higher precision with 82.24% against decision tree model with 77.06% and Naïve Bayes with 73.58%, and also recall, 92.92% for decision tree, and 90.20% for Naïve Bayes and 80.13% for Logistic regression. In term of time, Naïve Bayes took just 47ms to build the model, while Decision tree took 390ms and Logistic regression 344ms. We used also ROC as metric to compare between the three models, Figure 7 clearly illustrates that the Decision Tree classifier is more accurate than the Logistic regression and the 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.

## 5. Conclusion

Employability is a major problem for all graduates, applying data mining on employability data can help decision makers to take proactive actions, which is why Hassan the 1st University conducted an employability survey; the main objective of this survey was to provide the adequate elements to answer the problem of employability.

The type of data plays an important role for choosing the appropriate data mining algorithm we want to apply. In this study, we wanted to determine which algorithm is the best suited for employability data, and which presents the better model prediction, We applied three classification algorithms, Decision tree, Logistic regression and Naïve Bayes using Rapid Miner studio educational version 8.1.000, and the result have shown that Decision tree is better and more suited for prediction employability comparing with Logistic regression and Naïve Bayes. In fact, Decision tree was better in all metrics, accuracy, classification error, kappa statistics, f-measure, recall, sensitivity, precision and roc, except for time to build the model, Naïve Bayes was faster.

## Acknowledgement

Special thanks to the Prof. Ahmed NEJMEDDINE, president of Hassan the 1st University for his encouragement and his help, and also to Prof. Leila Loukili Idrissi, in charge of mission at the University, for her contribution providing this employability data of graduates to work on in this study.

## References

[1] Venkatadri.m, lokanatha c. reddy A comparative study on decision tree classification algorithms in data mining, 2008, ISSN: 0974-3596
[2] Pooja Thakar, Anil Mehta, Manisha - Role of Secondary Attributes to Boost the Prediction Accuracy of Student's Employability Via Data Mining, IJACSA, 2015, doi: 10.14569/IJACSA.2015.061112
[3] Muskan Kukreja, Stephen Albert Johnston and Phillip Stafford - Comparative study of classification algorithms for immunosignaturing data, 2012. doi: 10.1186/1471-2105-13-139
[4] 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, 2008, doi: 10.1007/s10115-007-0114-2
[5] Mary L. McHugh - Interrater reliability: the kappa statistic, 2012, IJSCE, ISSN: 2231-2307
[6] Hetal Bhavsar, amit Ganatra - An empirical evaluation of data mining classification algorithms, 2016, IJCSIS, 2016, ISSN 1947-5500
[7] Hossin, M.1 and Sulaiman - A review on evaluation metrics for data classification evaluations, IJDKP, 2015, doi: 10.5121/ijdkp.2015.5201
[8] Karimella Vikram, - Data Mining Tools and Techniques: A review, CEIS, 2011, ISSN 2222-2863
[9] Marko Arsenovic - A Comparison of Contemporary Data Mining Tools, ISCIS, 2017
[10] Vijay Kotu and Bala Deshpande, Morgan Kaufmann, Predictive analytics and Data mining, 2014, pp. 17-27, 64-71.
[11] Stephane Tuffery, Data mining and statistics for decision making, Ltd, 2011, pp. 43-72.
[12] Robert Nisbet, Gary Miner, Ken Yale - Handbook of statistical analysis and data mining applications, USA: Academic Press, 2009, pp 39-50, 53-62.
[13] Lemberger, Pirmin, Batty, Marc, Morel, Mederic, Big Data et Data mining: Manuel du data scientist, Malakoff: Dunod, 2015, pp.91-106.
[14] Nong Ye, Handbook of Data mining, New Jersey: Lawrence Erlbaum Associates, 2003, pp. 5-24, 104-125.
[15] Han, Jiawei,Kamber, Micheline,Pei, Jian, Data mining: Concepts and Techniques 3rd edition, USA: Morgan Kaufmann,2012, pp. 44-50, 88-94.
[16] https://rapidminer.com/products/studio/feature-list/
[17] https://docs.rapidminer.com/