Journal of Engineering and Applied Sciences 14 (5): 1690-1694, 2019
ISSN: 1816-949X

Prediction Model for Employability in Morocco Using Data Mining Techniques

Mohamed Saouabi and Abdellah Ezzati
LAVETE Laboratory, FST, University Hassan 1st, Hay El-Farah 02,
Rue Ibne Batouta, NR06 Settat, Morocco

Abstract: Big data is not about just storing data, it's to explore large volumes of data and extract valuable information and knowledge from it 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. Graduates face every year real competitions to ensure their employability. Mining employability data will give decision makers a great view of the data and opportunities to make improvement in this sector. In this study, we presented a data mining model for predicting employability using the classification algorithm decision tree also we presented the variables which have an important role predicting graduate's employability.

Key words: Data mining, big data, employability prediction, decision tree (C4.5), classification, Prediction Mdel

INTRODUCTION

Now a days, people start to use a lot of connected things and objects (Ding *et al*., 2013; Wu *et al*., 2008) such as cars, phones. Data is generated almost all the time which give this huge amount of data ready to be used and stored. Traditional technologies (Silva *et al*., 2016) cannot process this amount of data due to its limited storage capacity, the rigid management, the lack of scalability and flexibility which lead us to the big data revolution (Agrawal and Nyamful, 2016). But big data isn't just about storing the data the extraction of the knowledge is the most important process (Prasad and Aruna, 2016). Data mining advanced algorithms (Aslam and Ashraf, 2014) are needed in order to get accurate results and knowledge to predict future observations. In this experiment, we used Rapid Miner Studio (Kori, 2017) Educational Version 8.1.000 in Hadoop (Kumar *et al*., 2014), it is used to implement machine learning algorithms and it includes also Weka extension to implement algorithms designed for Weka mining tool. Also it provides learning schemes, models and algorithms from R scripts.

Employment (McQuaid *et al*., 2005) is the main form of social integration, a factor of 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. To define employability briefly from a graduate's perspective, it's the capability to gain and obtain a new employment. The enhancement of graduates employability is a significant and persistent problematic. Employability represents a serious problem for graduates they face every year real competitions to ensure their employability, the professional insertion is increasingly difficult. There are many explanations and causes for this matter, factors which take a big responsibility, for example, the poor economic performance of the country, the structure of the economy, the educational system or maybe the university fields of study which makes the professional insertion a bit difficult. This is why we are using data mining in order to propose solutions and opportunities for future prediction (Kaur *et al*., 2015) a model for predicting employability using classification algorithms and the variables which have an important role predicting graduate's employability.

Literature review: Many fields now are taking advantage of the powerful use of data mining and the opportunities offered in order to improve a particular domain here presented below few previous works using data mining classification techniques.

Kaur *et al*. (2015) proposed an experiment focusing on identifying slow learners in the educational field using data mining classification techniques using real world data collected from high schools. Using Weka, they applied several algorithms such as multilayer perception, Naive Bayes, SMO, J48 and REP tree in order to find out the best classifier model. The results have shown that multilayer perception algorithm is the best classifier with 75% accuracy of the model.

Venkatadri and Lokanatha (2010) presented an experiment using various data mining classification algorithms, the aim of this experiment is to choose the best suited algorithm performing the best accuracy model

Corresponding Author: Mohamed Saouabi, LAVETE Laboratory, FST, University Hassan 1st, Hay El-Farah 02,
Rue Ibne Batouta, NR06 Settat, Morocco

and to evaluate the models using different metrics, they used sample data but using real world data would be better in order to use the results in real life and take advantage of the extract knowledge to make improvements.

Chitra and Subashini (2013) presented an experiment in the bank sector, the aim of this experiment is to discover and anticipate in advance fraud prevention and detection, customer retention, marketing and risk management.

Mirmozaffari *et al.* (2017) proposed a system implemented in Weka using data mining techniques in order to predict heart disease, they used several classification algorithms and the results have shown that the random tree algorithm achieved 97.6077% of the model accuracy and the lowest errors and therefore, considered as the highest algorithm performance.

MATERIALS AND METHODS

In this study, we used Rapid Miner Studio Educational Version 8.1.000 using an employability dataset. In our previous research, data mining techniques for predicting employability. In Morocco (Mohamed and Abdellah, 2018), we did an experiment comparing three classification algorithms (Kaur *et al.*, 2015), decision tree, logistic regression and Naive Bayes and the results have shown that decision tree model is better than logistic regression and Naive Bayes in all metrics (Hossin and Sulaiman, 2015), accuracy, classification error, Kappa statistics, F-measure, recall, sensitivity, precision and ROC, except for time to build the model, Naive Bayes was faster. Here, presented below a brief summary of the results of the comparative between the data mining algorithms decision tree (Song and Lu, 2015), Logistic regression and Naive Bayes (Pisote and Bhuyar, 2015) (Table 1).

And now, we will present the most efficient model which is decision tree (C4.5) which we used to classify graduates into working and not working for predicting employability (Fig. 1).

Data collection: In this phase, we collect the data we're going to use 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. This survey took 3 years which means it started in 2013 and finished in 2016. Data is large, multivariate, incomplete, heterogeneous and unbalanced in nature (Lee *et al.*, 2017). In the next phase, we will prepare the data and we will clean it, so, it can be ready for modeling and applying the classification algorithm decision tree.

Data preparation: Once we collected the data, this data need to be selected, cleaned, built into the desirable format and formatted. This is one of the crucial phases in the data mining process and it's the most time-consuming part of the process. The data at first contained 1752 rows and 22 attributes, after cleaning and removing the non-pertinent data, like name, phone number, email, etc. The final data contains 1208 instances of 13 attributes as described (Table 2).

Table 1: Results of the performance comparison of the classifiers

Algorithm	Accuracy (%)	Precision	Classification error (%)	Recall (%)	Kappa statistics	F-measure	Sensitivity (%)	Time to build (msec)
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
Naive Bayes	78.23	73.58	21.77	90.20	0.561	0.81	90.20	47

Table 2: Description of the attributes

Attribute	Types	Description
Gender	Binary	Gender of the graduates (male, female)
Diploma	Nominal	Type of the diploma of the graduates
Field	Nominal	Field of study
Grade	Ordinal	Which grade in the baccalaureate
University	Nominal	Which university the graduate graduated from
Practice level	Ordinal	The graduate level of practice in his field of study
Informatic level	Ordinal	The graduate level of practice in information technology
French level	Ordinal	The graduate French level
English level	Ordinal	The graduate English level
Baccalaureate serie	Nominal	The graduate baccalaureate option
Training period	Binary	Did the graduate made any training (yes or no)
Theorical level	Ordinal	The graduate level of theory in his field of study
Employability	Binary	Is the graduate working or not working

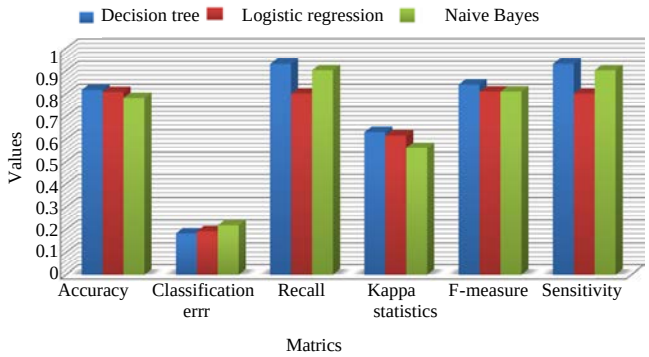

Fig. 1: Performance comparison of the three classifiers using the different metrics

Fig. 2: Graph of the class distribution

Modeling: Now, after the data collection and preparation, and based on the type of data we have and the type of the variable we want to predict, we implemented the classification algorithm Decision tree (Table 1 and Fig. 2).

We used different metrics in order to evaluate the model's performance: Accuracy, classification error, recall, Kappa statistics, F-measure, sensitivity, precision and the time to build the model, here's a description below of the different metrics we used.

RESULTS AND DISCUSSION

After applying decision tree this is the results f the experiment (Table 3-6). As we mentioned before, in our previous research, data mining techniques for predicting employability. In Morocco, we did an experiment comparing three classification algorithms, decision tree, logistic regression and Naive Bayes and the results have shown that decision tree model is better than Naive Bayes in all metrics, accuracy, classification error, Kappa statistics, F-measure, recall, sensitivity, precision and ROC, except for time to build the model, Naive Bayes was faster. And now, we will present the decision tree (C4.5) model for predicting employability and we'll describe the results.

Accuracy Galdi and Tagliaferri (2018) represents the percentages of instances correctly classified by the

Table 3: The class distribution

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

Table 4: Confusion Matrix for binary classification

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

Table 5: Metrics for classification evaluation

Metrics	Formula	Evaluation focus
Accuracy	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	FP+FN/TP+ FP+TN+FN	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated
F-measure	2×P×R/P+R	This metric represents the harmonic mean between recall and precision values
Sensitivity	TP/TP+FN	This metric is used to measure the fraction of positive patterns that are correctly classified
Precision (p)	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)	TP/TP+TN	Recall is used to measure the fraction of positive patterns that are correctly classified

algorithm, based on the results, Table 6 shows that the accuracy of the model predicted by decision tree is 81.70 and 18.30 % of the classification error representing the misclassified instances. Also, Kappa statistics (McHugh, 2012) have shown that the decision tree model (0.631) is a substantial model based on Cohen interpretation suggestion for Kappa results.

Sensitivity by Yao (2003) presents the correctly predicted positive observations to all observations in actual class 92.92% for decision tree. Another important metric is F-measure (Powers, 2011), it tells how precise the classifier is and how many instances are classified

Table 6: Performance of the decision tree model

Algorithm/metrics	Decision tree
Accuracy	81.70
Precision	77.06
Classification error	18.30
Kappa statistics	0.631
F measure	0.84
Recall	92.92
Sensitivity	92.92
Time to build the model (msec)	390

Table 7: Confusion matrix of decision tree model

Confusion matrix	True working	True not working	Class precision (%)
Pred. working	409	44	90.29
Pred. not working	177	578	76.56
Class recall	69.80 (%)	92.93 (%)	

correctly as well as how robust it is again the results shows that decision tree is classified correctly with 0.84 F-measure rates.

Another metric is precision, Alvarez (2002) results shows that decision tree model has a high precision with 77.06% and also recall with 92.92%. And in term of time to build the model, it took 390 m sec.

Based on the decision tree model applied, the variables which have an important role predicting graduate's employability are: University {encg, ensa, fst}, Grade {Very Good, Good, Pretty Good}, Training period {Yes}, French level {Excellent, Good, Medium}.

Model developed by decision tree algorithm: Here, presented below the decision tree model, these rules describe and give a clear insight of the attributes that affect the graduate's employability.

Algorithm 1; Decision tree algorithm:

University = ENCG: Working {Working=143, NotWorking=5}
University = ENSA
| | Grade = Good: Working {Working=2, NotWorking=1}
| | Grade = Passable: NotWorking {Working=0, NotWorking=8}
| Gender = Male: Working {Working=20, NotWorking=0}
University = ESTB
| EnglishLevel = High
| | TrainingPeriod = No: NotWorking {Working=0, NotWorking=11}
| | TrainingPeriod = Yes
| | | FrenchLevel = Low: NotWorking {Working=0, NotWorking=3}
| | | FrenchLevel = Medium: Working {Working=2, NotWorking=0}
| | | FrenchLevel = Vey Low| | | | InformaticLevel = Very Good: Working {Working=13, NotWorking=1}
| | | | InformaticLevel = Medium: Working {Working=7, NotWorking=0}
| | | | InformaticLevel = Excellent: Working {Working=7, NotWorking=0}
| EnglishLevel = Excellent: Working {Working=5, NotWorking=0}
| EnglishLevel = Low: NotWorking {Working=0, NotWorking=3}
| EnglishLevel = Medium: Working {Working=3, NotWorking=0}
University = FPK: NotWorking {Working=100, NotWorking=212}
University = FSJES: Working {Working=161, NotWorking=105}
University = FST
| Grade = Pretty Good
| | Diploma = Diplôme d'ingenieur: Working {Working=10, NotWorking=0}
| | Diploma = Doctorat: Working {Working=2, NotWorking=0}
| | Diploma = Licence professionnelle: NotWorking {Working=5, NotWorking=13}
| | Diploma = Licence sciences et techniques
| | | FrenchLevel = Medium: Working {Working=10, NotWorking=3}
| | | FrenchLevel = Low: NotWorking {Working=0, NotWorking=13}
| | | FrenchLevel = Excellent: Working {Working=6, NotWorking=0}
| | | FrenchLevel = Vey Low
| | | | EnglishLevel = High
| | | | | PracticeLevel = Low: NotWorking {Working=0, NotWorking=10}
| | | | | PracticeLevel = Very Good
| | | | | | Gender = Female: Working {Working=12, NotWorking=0}
| | | | | | Gender = Male: Working {Working=4, NotWorking=0}
| | | | | PracticeLevel = Medium
| | | | | | InformaticLevel = Medium: Working {Working=6, NotWorking=1}
| | | | | | InformaticLevel = Very Good: Working {Working=7, NotWorking=0}
| | | | | PracticeLevel = Excellent
| | | | | | BaccalaureateSerie = SVT: Working {Working=7, NotWorking=0}
| | | | | | BaccalaureateSerie = Sciences et technologie électrique: Working {Working=2, NotWorking=0}
| | | | EnglishLevel = Excellent
| | | | | Field = Gestion: NotWorking {Working=0, NotWorking=12}
| | | | | Field = Génie Electrique: Working {Working=4, NotWorking=0}
| | | | | Field = Génie Mécanique: NotWorking {Working=0, NotWorking=4}
| | | | | Field = Management Logistique et Transport: Working {Working=2, NotWorking=0}
| | | | | Field = Protection de l'environnement: Working {Working=2, NotWorking=0}
| | | | | Field = Techniques d'Analyse et Contrôle de Qualité: NotWorking {Working=0, NotWorking=5}
| | | | EnglishLevel = Medium: Working {Working=10, NotWorking=1}
| | Diploma = Master recherche
| | | PracticeLevel = Very Good: Working {Working=14, NotWorking=7}
| | | PracticeLevel = Low: NotWorking {Working=0, NotWorking=12}
| | | PracticeLevel = Medium: Working {Working=11, NotWorking=6}
| | | PracticeLevel = Excellent: Working {Working=4, NotWorking=0}
| | Diploma = Master spécialisé: Working {Working=6, NotWorking=1}
| Grade = Good: Working {Working=43, NotWorking=24}
| Grade = Passable: NotWorking {Working=21, NotWorking=94}
| Grade = Excellent: Working {Working=2, NotWorking=0}

CONCLUSION

Determining the factors that influence the employability of the graduates will give decision makers a great view and opportunities to make improvements in this sector. The objective of this paper is to present in details the model of decision tree algorithm using Rapid Miner Studio Educational Version 8.1.000 in Hadoop. In this study we used real data, it's 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. We presented first the model's performance evaluation using different metrics, accuracy, classification error, recall, F-measure, Kappa statistics,

sensitivity, precision and time to build the model. The results have shown that decision tree model is accurate with 81.70% with 987 correctly classified instances and just 221 misclassified instances based on the generated confusion matrix (Sammut and Webb, 2017). And then, we presented the variables which have an important role predicting graduate's employability which are University {encg, fsjes, fst}, Grade {Very Good, Good, Pretty Good}, Training period {Yes}, French level {Excellent, Good, Medium}.

ACKNOWLEDGEMENTS

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

REFERENCES

Agrawal, R. and C. Nyamful, 2016. Challenges of big data storage and management. Global J. Inf. Technol., 6: 01-10.

Alvarez, S.A., 2002. An exact analytical relation among recall, precision and classification accuracy in information retrieval. Comput. Sci., 1: 1-22.

Aslam, S. and I. Ashraf, 2014. Data mining algorithms and their applications in education data mining. Intl. J. Adv. Res. Comput. Sci. Manage. Stud., 2: 50-56.

Chitra, K. and B. Subashini, 2013. Data mining techniques and its applications in banking sector. Intl. J. Emerging Technol. Adv. Eng., 3: 219-226.

Ding, G., L. Wang and Q. Wu, 2013. Big data analytics in future internet of things. Res. J. Internet, 1: 1-6.

Galdi, P. and R. Tagliaferri, 2018. Data Mining: Accuracy and Error Measures for Classification and Prediction. In: Reference Module in Life Sciences, Roitberg, B.D. (Ed.). Elsevier, New York, USA., pp: 1-14.

Hossin, M. and M.N. Sulaiman, 2015. A review on evaluation metrics for data classification evaluations. Int. J. Data Mining Knowledge Manage. Process, 5: 1-11.

Kaur, P., M. Singh and G.S. Josan, 2015. Classification and prediction based data mining algorithms to predict slow learners in education sector. Proc. Comput. Sci., 57: 500-508.

Kori, A., 2017. Comparative study of data classifiers using rapidminer. Intl. J. Eng. Dev. Res., 5: 1041-1043.

Kumar, R., B.B. Parashar, S. Gupta, Y. Sharma and N. Gupta, 2014. Apache Hadoop, NoSQL and NewSQL solutions of big data. Intl. J. Adv. Found. Res. Sci. Eng., 1: 28-36.

Lee, C.K.M., Y. Cao and K.H. Ng, 2017. Big Data Analytics for Predictive Maintenance Strategies. In: Supply Chain Management in the Big Data Era, Chan, H.K., N. Subramanian and M.D.A. Abdulrahman (Eds.). IGI Global, Pennsylvania, USA., pp: 50-74.

McHugh, M.L., 2012. Interrater reliability: The kappa statistic. Biochemia Med., 22: 276-282.

McQuaid, R.W., A. Green and M. Danson, 2005. Introducing employability. Urban Stud., 42: 191-195.

Mirmozaffari, M., A. Alinezhad and A. Gilanpour, 2017. Data mining classification algorithms for heart disease prediction. Intl. J. Comput. Commun. Instrum. Eng., 4: 11-15.

Mohamed, S. and E. Abdellah, 2018. Data mining techniques for predicting employability in morocco. Intl. J. Eng. Technol., 7: 17-20.

Pisote, A. and V. Bhuyar, 2015. Review article on opinion mining using naive bayes classifier. Adv. Comput. Res., 7: 259-261.

Powers, D.M.W., 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol., 2: 37-63.

Prasad, R.S.R. and C. Aruna, 2016. Scalable and Flexible Big Data Analytic Framework (SFBAF) for big data processing and knowledge extraction. Proceedings of the International Conference on Engineering Technologies and Big Data Analytics (ETBDA'2016), January 21-22, 2016, Institute of International Education, Bangkok, Thailand, pp: 51-55.

Sammut, C. and G.I. Webb, 2017. Encyclopedia of Machine Learning and Data Mining. 2nd Edn., Springer, Berlin, Germany, ISBN:9781489976857, Pages: 1335.

Silva, Y.N., I. Almeida and M. Queiroz, 2016. SQL: From traditional databases to big data. Proceedings of the 47th ACM Technical Symposium on Computing Science Education (SIGCSE'16), March 2-5, 2016, ACM, New York, USA., ISBN:978-1-4503-3685-7, pp: 413-418.

Song, Y.Y. and Y. Lu, 2015. Decision tree methods: Applications for classification and prediction. Shanghai Arch. Psychiatry, 27: 130-135.

Venkatadri, M. and C.R. Lokanatha, 2010. A comparative study on decision tree classification algorithms in data mining. Intl. J. Comput. Appl. Eng. Technol. Sci., 2: 24-29.

Wu, X., V. Kumar, J.R. Quinlan, J. Ghosh and Q. Yang *et al*., 2008. Top 10 algorithms in data mining. Knowledge Inform. Syst., 14: 1-37.

Yao, J.T., 2003. Sensitivity analysis for data mining. Proceedings of the 22nd International Conference on North American Fuzzy Information Processing Society (NAFIPS 2003), July 24-26, 2003, IEEE, Chicago, Illinois, USA., ISBN:0-7803.