Machine Learning for Students Employability Prediction

Aniss Moumen[1,*], Imane El Bakkouri[1], Hamza Kadimi[1], Abir Zahi[1], Ihsane Sardi[1], Mohammed Saad Tebaa[1], Ziyad Bousserrhine[1], Hanae Baraka[2]
[1]National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
[2]National School of Applied Sciences, Hassan I University, Berrechid, Morocco

Keywords: Employability, Machine learning, Students, review

Abstract: Nowadays, students' employability is a major concern for the institutions, and predicting their employability can help take timely actions to increase the institutional placement ratio. Data mining techniques such as classification is best suited for predicting the employability of students. Knowing weaknesses before appearing can help students work in areas that they need to improve to best match the company's skillset. Moreover, predict student employability can help educational staff in elaborating curriculum programs. This paper presents a systematic and exploratory literature review on Machine learning algorithms for students employability from Scopus Database.

1 INTRODUCTION

The use of machine learning techniques in the education field is increasing nowadays. Many techniques are used in educational data mining like Decision trees, neural networks, Naive Bayes, K-nearest neighbors, support vector machines, etc. Using these techniques can automate some tasks (Kamath et al., 2016; Peña-Ayala, 2014), such as tracking variables of the students correlated with student performance to find the at-risk students. So, a comparative analysis is done to find the algorithm with the highest accuracy.

The prediction technique in learning management system data is a tool that can improve the governance of the educational system. The training set of data is used to guide the learning process, and a test set is used to analyze students' performance (Bai & Hira, 2021).

According to (Mezhoudi et al., 2021; Moumen et al., 2020) Many researchers study the problems related to employability prediction; simultaneously, they are concerned about developing automated techniques required for the analysis phase of student performance.

This paper presents a systematic exploratory literature review about student employability prediction systems based on Machine learning algorithms. We started from the Scopus database to find previous works; then, we analyzed all collected papers through two levels: a meta-analysis and a thematic analysis to elaborate a comparative study between models depending on accuracy.

2 METHODS

According to (MacKenzie et al., 2012) and (Kitchenham, 2014), the Systematic Literature Review is defined as "a systematic literature review is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest". The researchers conducted a SLR to summarise the existing evidence concerning technology. In this paper, we focus on the topic "Students Employability Prediction", and we have identified the questions :

- Which features characterize student employability?
- Which machine learning algorithms are used to predict student employability? and with which tools?
- Which machine learning algorithm performs a better accuracy?

To address this systematic literature review, we start by defining the query defining by the keywords: "Student", "Employability", and "Machine learning". Then we have launched our research through the

Moumen, A., El Bakkouri, I., Kadimi, H., Zahi, A., Sardi, I., Tebaa, M., Bousserrhine, Z. and Baraka, H.
Machine Learning for Students Employability Prediction.
DOI: 10.5220/0010732400003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 274-278
ISBN: 978-989-758-559-3

Scopus database. We have found 28 references. For this paper, we have selected 15 articles. The figure below summarizes all the steps.

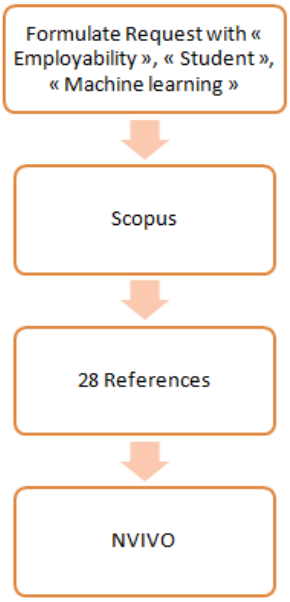

Figure 1: Steps of systematic literature review on "Student Employability prediction"

4 RESULTS AND DISCUSSIONS

After cleaning our corpus containing 25 references, which will be presented in the meta-analysis section below, we will proceed with a thematic analysis from only 15 available references to discuss data & features, tools and machine learning algorithms and evaluation methods.

4.1 Meta-analyze

In our corpus, we have 14 conference papers and 11 journal articles. The publications start from 2016 to 2021. The figure below indicates the number of papers by year.

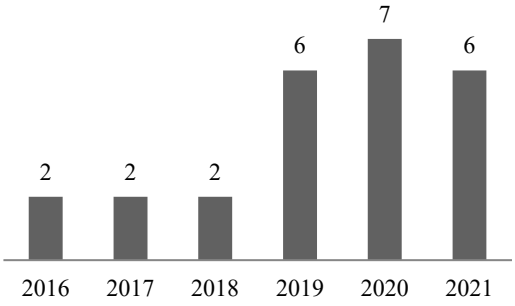

Figure 2: Number of papers from Scopus by year.

From this figure, we can conclude that 2019 was the year when the publications start to increase. The table below summarizes the minimum, maximum, mean and median of publications :

Table 1: Summary of publications

Min	Max	Mean	Median
2	7	4	4

Concerning the journals that publish on this topic, we can mention: Journal of Technical Education and Training, Journal of Ambient Intelligence and Humanized Computing, International Journal of Performability Engineering.

4.2 Features and Data Collection

From this corpus, we have found two categories of technics used by the researchers to collect data: 1) Questionnaires addressed to students and employers or 2) Data gathered from institutions with or without survey (University, career centre, registrar, guidance, etc.).

Concerning the features for the first category, we conclude that the authors focus on measuring Employability skills and study their relationship with individual characteristics. For example, the authors (Bai & Hira, 2021; Dubey & Mani, 2019; Karim & Maat, 2019; M. M. Almutairi & M. H. A. Hasanat, 2018) use surveys to collect data about employability skills such as: Problem-solving, Reasoning ability, Teamwork, Self-management and Time management, Communication skills, Technical skills, ICT Skills.

The second category of researchers (Casuat, Festijo, et al., 2020; Casuat, Sadhiqin Mohd Isira, et al., 2020; Febro, 2019; Nagaria & Senthil Velan, 2020; Piad et al., 2016) study the relationship between the academic performance grades and scores in various subjects (Math, Natural science, Language and Humanities, Social science) and personal information (gender, educational background, language usage, location, degree type, specialization, work experience), family information (parents educational background and job, parent's income, number of sisters and brother) and some orientation and preferences (religion, music and sport).

Another work from (Vignesh et al., 2020) uses real-time data collected from Twitter, Kaggle, UCI, Data.gov and Google form to study an XGBoost Classifier of students features, such as student academic scores, specialization, programming and analytical capabilities, personal details.

We have found that researchers try to increase and randomize their samples. So, (Dubey & Mani, 2019) use bootstrapping technique to increase her sample from 95 to 195 high school students. (Thakar et al., 2017) use 9459 instances and 160 attributes. (Casuat, Festijo, et al., 2020) collect 3000 observations of 12 students attributes from Mock job interview evaluation from 2015 to 2018 and use the synthetic minority over-sampling technique (SMOTE) to resolve the problem of the imbalanced dataset. And also (Bai & Hira, 2021) use 10000 samples as input of their model based on Deeplearning algorithms.

Moreover, the authors split their datasets according to the proportionality 80%/20% or 70%/30% for training and validation steps. Table below synthesize this review :

Table 2: Sampling techniques

Authors	Sample size (original)	Sampling
Piad, K.C. et al. (2016)	515 students	Random
Bharambe, Y. et al. (2017)	2100 students	Random
M. M. Almutairi et al. (2018)	55 recruters 194 students	Random
A. Dubey et al.(2019)	95 students	Random with Boostrapping
Nagaria, J. et al. (2020)	215 students	Random
Casuat, C.D. et al. (2020)	3000 observations	Random with SMOTE
Thakar, P. et al. (2017)	9459 students	Random
Bai, A. et al. (2021)	10000 students	Random

4.3 Machine Learning Approaches

This review shows that the authors mostly use supervised machine learning models to build their proposals. Therefore, they use data mining techniques or meta-heuristic algorithms in the data preprocessing, and selection features stage to improve the model's accuracy. The figure below resume the stages commonly followed by the researchers :

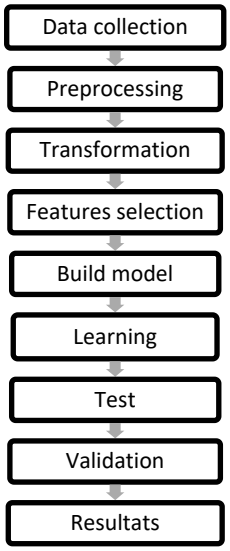

Figure 3: Stages of Machine learning modelling.

The authors use a set of Machine Learning Algorithms to conduct their comparative study or build a proposed model. From the table below, we conclude that the most used algorithms are: Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Logistic Regression (LR). The authors implement these algorithms through : Python, Matlab and R. Some authors also deploy their studies with WEKA, SPSS or Rapide Miner Studio.

Table 3: Machine learning algorithms and tools

Authors	ML algorithms	Tools
Piad, K.C. et al. (2016)	Naïve Bayes, J48, SimpleCart, Logistic Regression, Chaid	WEKA and SPSS
Bharambe, Y. et al. (2017)	Naive Bayes, Decision Tree, Logistic Regression, KNN, SVM, Random Forest, Multi-class Ada Boosted and Quadratic Discriminant Analysis (QDA)	Python
M. M. Almutairi et al. (2018)	J4.8, Naïve Bayes and KNN	WEKA
A. Dubey et al.(2019)	Random Forest, KNN, SVM, Logistic Regression and Decision Tree	Scikit-learn (Python)

Nagaria, J. et al. (2020)	Decision Tree; Exploratory Data Analysis; Random Forest	R
Casuat, C.D. et al. (2020)	DT, RF, SVM, KNN, LR	Python
Thakar, P. et al. (2017)	Simple Cart, KNN, VFI, VotedPeceptron, REP Tree, LMT, J48, Graft J48, ADT Tree, Random Forest, Random Tree, IB1, Kstar,Ibk, DTNB	RapidMiner Studio
Bai, A. et al. (2021)	Deep Belief Network and Soft max Regression, Deep Autoencoder-Softmax Regression, Deep AutoEncoder, Deep Neural Network.	Matlab

4.4 Evaluation of the Models

To measure and evaluate the performance of the models, the authors utilized the Confusion Matrix resulting from testing the trained models using the Test dataset.

Also, the authors used several metrics such as Accuracy, Precision, Recall, and F1Score (Géron, 2019), which are expressed as :

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1-score = 2(Precision * Recall)/(Precision + Recall)

The table below concludes that the accuracy of the logistic regression, random forest, and SVM give the best results.

Deep learning models seem to be a promising perspective to explore by researchers in this topic.

Table 4: Comparative study

Authors	Best algorithm	Accuracy (%)
Piad, K. C. et al. (2016)	Logistic Regression	78.4
Bharambe, Y. et al. (2017)	Random Forest	99
M. M. Almutairi et al. (2018)	Naïve Bayes	69
A. Dubey et al.(2019)	Logistic Regression	93
Nagaria, J. et al. (2020)	Random Forest	85
Casuat, C.D. et al. (2020)	SVM	91
Thakar, P. et al. (2017)	Model based on : Simple Cart, kStar, Random Forest and Random Tree	87
Bai, A. et al. (2021)	DBN-SR	98

5 CONCLUSIONS

This study presents an exploratory literature review about the usage of machine learning algorithms to predict student employability. Future research needs to be elaborated to predict student employability, especially in Morocco.

REFERENCES

Bai, A., & Hira, S. (2021). An intelligent hybrid deep belief network model for predicting students employability. *Soft Computing*, *25*(14), 9241-9254. Scopus. https://doi.org/10.1007/s00500-021-05850-x

Casuat, C. D., Festijo, E. D., & Alon, A. S. (2020). Predicting students' employability using support vector machine : A smote-optimized machine learning system. *International Journal of Emerging Trends in Engineering Research*, *8*(5), 2101-2106. Scopus. https://doi.org/10.30534/ijeter/2020/102852020

Casuat, C. D., Sadhiqin Mohd Isira, A., Festijo, E. D., Sarraga Alon, A., Mindoro, J. N., & Susa, J. A. B. (2020). A Development of Fuzzy Logic Expert-Based Recommender System for Improving Students' Employability. *2020 11th IEEE Control and System Graduate Research Colloquium, ICSGRC 2020 - Proceedings*, 59-62. Scopus. https://doi.org/10.1109/ICSGRC49013.2020.9232543

Dubey, A., & Mani, M. (2019). Using machine learning to predict high school student employability—A case study. *Proceedings - 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019*, 604-605. Scopus. https://doi.org/10.1109/DSAA.2019.00078

Febro, J. D. (2019). Utilizing feature selection in identifying predicting factors of student retention. *International Journal of Advanced Computer Science and Applications*, *10*(9), 269-274. Scopus. https://doi.org/10.14569/ijacsa.2019.0100934

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts,*

Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

Kamath, R. S., Kamath, R. S., & Kamat, R. K. (2016). *Educational data mining with R and Rattle*. River Publishing.

Karim, Z. I. A., & Maat, S. M. (2019). Employability skills model for engineering technology students. *Journal of Technical Education and Training*, *11*(2), 79-87. Scopus. https://doi.org/10.30880/jtet.2019.11.02.008

Kitchenham, B. (2014). *Procedures for Performing Systematic Reviews* (p. 33). Keele University.

M. M. Almutairi & M. H. A. Hasanat. (2018). Predicting the suitability of IS students' skills for the recruitment in Saudi Arabian industry. *2018 21st Saudi Computer Society National Computer Conference (NCC)*, 1-6. IEEE. https://doi.org/10.1109/NCG.2018.8593016

MacKenzie, H., Dewey, A., Drahota, A., Kilburn, S., Kalra, P. R., Fogg, C., & Zachariah, D. (2012). Systematic Reviews : What They Are, Why They Are Important, and How to Get Involved. *Systematic Reviews*, *4*, 10.

Mezhoudi, N., Alghamdi, R., Aljunaid, R., Krichna, G., & Düştegör, D. (2021). Employability prediction : A survey of current approaches, research challenges and applications. *Journal of Ambient Intelligence and Humanized Computing*. Scopus. https://doi.org/10.1007/s12652-021-03276-9

Moumen, A., Bouchama, E. H., & El bouzakri el idrissi, Y. (2020). Data mining techniques for employability : Systematic literature review. *2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, 1-5. IEEE. https://doi.org/10.1109/ICECOCS50124.2020.9314555

Nagaria, J., & Senthil Velan, S. (2020). Utilizing Exploratory Data Analysis for the Prediction of Campus Placement for Educational Institutions. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*. Scopus. https://doi.org/10.1109/ICCCNT49239.2020.9225441

Peña-Ayala, A. (Éd.). (2014). *Educational Data Mining* (Vol. 524). Springer International Publishing. https://doi.org/10.1007/978-3-319-02738-8

Piad, K. C., Dumlao, M., Ballera, M. A., & Ambat, S. C. (2016). Predicting IT employability using data mining techniques. *2016 3rd International Conference on Digital Information Processing, Data Mining, and Wireless Communications, DIPDMWC 2016*, 26-30. Scopus. https://doi.org/10.1109/DIPDMWC.2016.7529358

Thakar, P., Mehta, A., & Manisha. (2017). A unified model of clustering and classification to improve students' employability prediction. *International Journal of Intelligent Systems and Applications*, *9*(9), 10-18. Scopus. https://doi.org/10.5815/ijisa.2017.09.02

Vignesh, A., Yokesh Selvan, T., Gopala Krishnan, G. K., Sasikumar, A. N., & Ambeth Kumar, V. D. (2020). *Efficient Student Profession Prediction Using XGBoost Algorithm* (Vol. 35, p. 148). Springer Science and Business Media Deutschland GmbH; Scopus. https://doi.org/10.1007/978-3-030-32150-5_15