A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms

https://doi.org/10.3991/ijoe.v17i02.20049

Inssaf El Guabassi (✉)
Abdelmalek Essaadi University, Tetouan, Morocco
elguabassi@gmail.com

Zakaria Bousalem
Hassan 1st University, Settat Morocco

Rim Marah
Abdelmalek Essaadi University, Tetouan, Morocco

Aimad Qazdar
Cadi Ayyad University, Marrakech, Morocco

Abstract—In the 21st century, University educations are becoming a key pillar of social and economic life. It plays a major role not only in the educational process but also in the ensuring of two important things which are a prosperous career and financial security. However, predicting university admission can be especially difficult because the students are not aware of admission requirements. For that reason, the main purpose of this research work is to provide a recommender system for early predicting university admission. Therefore, the contributions are threefold: The first is to apply several Supervised Machine Learning algorithms namely Linear Regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression. The second purpose is to compare and evaluate algorithms used to create a predictive model based on various evaluation metrics. The last purpose is to determine the most important parameters that influence the chance of admission. The experimental results showed that the Random Forest Regression is the most suitable Machine Learning algorithm for predicting university admission. Also, the Cumulative Grade Point Average is the most important parameter that influences the chance of admission.

Keywords—Machine Learning, Educational Data Mining, Linear Regression, Decision Tree, Support Vector Regression, Random Forest Regression

1 Introduction

Machine Learning is a subset of artificial intelligence (AI) that enable computers to automatically improve through experience, today it is becoming an increasing part of

our daily lives because its applications are extending to different areas like agriculture, finance, electronic commerce, logistics, marketing, security, shopping, etc. Moreover, Machine Learning enables an increasing number of applications that were not possible before. In the area of education, the adoption of Machine Learning is also accelerating. In recent decades, several researchers and scientists show their interest in the application of Machine Learning in an educational context [1] [2] [3]. Therefore, some examples of the topics that are studied by different researchers are:

- Dropout prediction in e-learning courses
- Early prediction of student outcomes
- Prediction of academic performance associated with internet usage behaviors
- Predicting at-risk university students in a virtual learning environment
- Mental stress detection in university students
- Dropout early warning systems for high school students
- Analyze and support mediation of student e-discussions
- Predicting MOOC dropout over weeks
- Identifying students' inquiry planning

Detection of learner learning style etc.In this sense, early predicting university admission is also considered an important topic for not only the new graduate students but also for the university. Unfortunately, newly graduate students usually are not aware of admission requirements to a postgraduate program at the university. Thus, they, therefore, spend their precious time and money focusing on things that won't increase their chances of admission to graduate programs.

The real and major issues of this topic are: Firstly, which the best Machine Learning algorithm for predicting university admission? Secondly, what most important parameters can affect the chance of admission?

The main aim of this paper is to provide a Recommender System for Early Predicting University Admission based on Machine Learning algorithms namely. Hence, the purposes are threefold. The first is to apply several supervised Machine Learning algorithms (i.e., Linear Regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression) to our dataset. The second purpose is to compare and evaluate algorithms used to create a predictive model based on various evaluation metrics. The last purpose is to determine the most important parameters that influence the chance of admission. It should be noted that the present research work uses the dataset for graduate students at the University of California in Los Angeles.

The outline of the present paper is as follows: Section 2 presents recent studies regarding the specified area. The settings and basic definitions are briefly described in Section 3. Section 4 concentrates on the proposed methodology. In Section 5 our proposed recommender system is presented. Section 6 contains results and discussion. Finally, Section 7 presents the main conclusions considering some future research directions.

2 Related Work

Due to the rapidly growing interest in the field of education[4][5][6], there are several research studies have been conducted on predicting university admission based on several factors using supervised or unsupervised Machine Learning algorithms. Xiaojun Wu & Jing Wu [7] conducted a study on predicting students' selection criteria in non-native language MBA admission based on Ridge regression, SVM, Random forest, GBDT. AlGhamdi and al [8] applied three classification methods which are linear regression, decision tree, and logistic regression model to automatically predict the postgraduate admission. Nandal [9] developed Student Admission Predictor (SAP) based on Deep Learning techniques namely Decision Tree, Support Vector Machine, Gaussian Naive Bayes, Linear Regression, Random Forest, and Deep Neural Networks. Zhao et al [10] explained their study based on a quantitative Machine Learning approach to predict master students' admission in professional institutions. They used standard SVM, S3VM, and SVM+, as well as their dataset, which is collected from Northeastern University's MS in Computer Science (MSCS) program.

3 Settings and Basic Definitions

In this section we will discuss the three main elements to achieve our purpose; these elements are Recommender Systems, Algorithms used, and evaluations methods.

3.1 Recommender systems

Simply put, a recommender system is an application intended to offer the user items that may be of interest to him according to his profile. Recommender systems are used in particular on online sales websites. They allow e-commerce merchants to automatically highlight products likely to interest visitors. Recommender algorithms can be divided into three categories which are:

- Content-Based Filtering [11]: This type of recommender system is based on profiles. We build profiles for users as well as for products
- Collaborative Filtering [12]: This method tries to find a group of users who have the same tastes and preferences as the target user
- Knowledge-based systems [13]: This type of recommender system recommends items based on the knowledge user.

3.2 Algorithms

There are many algorithms for predictive modeling Machine Learning. In the next sections, we will present the algorithms used to build predictive models which are Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR), and Random Forest Regression (RFR).

Linear Regression (LR) [14] is the most important algorithm in the field of Machine Learning, especially supervised learning. It is a way to model a relationship between a dependent variable and one or more independent variables. It consists of finding a regression line straight line through the points.

Decision Tree (DT) [15] is the most widely used classification and prediction technique. It is a tree structure, where each internal node with outgoing edges indicates a condition on an attribute, each branch is an outcome of the test, and each leaf terminal node represents a class label.

Support Vector Regression (SVR) [16] is also a very popular Machine Learning technique used in both classification and regression. It is similar to Linear Regression with only a few minor differences. SVR allows defining how much error is acceptable in our predictive model and will find an appropriate line to fit the data.

Random Forest Regression (RFR) [17] is an ensemble learning method that constructs a multitude of decision trees at training time and uses the average prediction of the individual trees to improve the prediction.

3.3 Evaluation methods

Evaluating a model is a core part of building an effective Machine Learning model. There are many methods of evaluation that can be used. In the following, we will discuss the three main metrics which we will use in our evaluation, named R-squared (R2), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

R-Squared (R2 or the coefficient of determination) [18] is an indicator that allows judging the quality of simple linear regression. It measures the fit between the model and the observed data or how well the regression equation is to describe the distribution of points. In short, the closer the coefficient of determination is to 0, the more the scatter plot disperses around the regression line. On the contrary, the more the R^2 tends towards 1, the more the cloud of points narrows around the regression line. When the points are exactly aligned on the regression line, then $R^2 = 1$.

Mean Square Error (MSE) [19] is the arithmetic mean of the squares of the predictions between the model and the observations. This is the value to be minimized in the context of a single or multiple regressions. The method is based on the nullity of the mean of the residuals. But the average of their squares is generally not zero.

Root Mean Square Error (RMSE) is a standard way to measure the error in model evaluation studies. It is the square root of the mean of the square of all of the errors.

After briefly determining the settings and describing basic definitions, in the next section, we will present the methodology used to develop our recommender System RSEPUA.

4 Proposed Methodology

University education is becoming the most important thing in today's world. It gives us useful skills and a huge experience. In most cases, having a higher education provides a good job and the best career in the future. Admission to graduate schools is the

most difficult and important step. The major problem students often fall into when applying to graduate schools as they are not aware of admission requirements. The purpose of this research work is to provide a recommender system for early predicting university admission based on Machine Learning algorithms. The methodology used in this work is composed of six steps. Figure 1 presents an overview of this methodology.

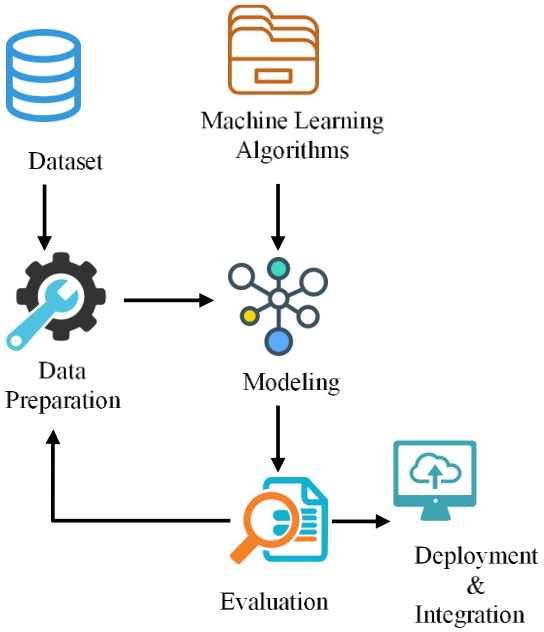

Fig. 1. Methodology used to our recommender System RSEPUA

Dataset: The data used in this research work is collected from a second version of the data set named "Graduate Admission 2" [20]. It is inspired by the University of California in Los Angeles Graduate Dataset. This data is available publicly even on Kaggle which offers free datasets for training and evaluation. It is comprised of 500 rows with 8 features. The dataset contains various features which are considered important during the enrollment and admission for master's programs. Table 1 represents an overview of dataset features used for training and testing.

Table 1. Dataset Features

Feature	Description	Type
GRE Score	Graduate Record Examinations (i.e., from 260 to 340)	Quantitative
TOEFL Score	Test of English as a Foreign Language (i.e., from 0 to 120)	Quantitative
University Rating	University Rating (i.e., from 1 to 5)	Quantitative
SOP	Statement of Purpose (i.e., from 1 to 5)	Quantitative
LOR	Letter of Recommendation (i.e., from 1 to 5)	Quantitative
CGPA	Cumulative Grade Point Average (i.e., from 0 to 10)	Quantitative
Research	Research Experience (i.e., 0 or 1)	Quantitative
Chance of Admit	Probability of getting admitted (i.e., from 0 to 1)	Quantitative

Data Preparation: It is referred to as "data preprocessing". It represents one of the most crucial steps in all Machine Learning projects because it involves data collection, formatting data, Improving data quality, feature engineering, and labeling

Modeling: This step involves the conception of different Machine Learning algorithms (e.g., regression, classification, clustering, etc.) that can be used for predicting university admission.

Machine Learning Algorithms: It represents the algorithms used to build our predictive model, which are Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR), and Random Forest Regression (RFR).

Evaluation: This step is a core part of building our Machine Learning model. There are different metrics of evaluations that can be used. The evaluation metrics used in this research work are Mean Square Error, Root Mean Square Error, and R-squared.

Deployment & Integration: It is all of the tasks that make our recommender system RSEPUA available for use.

5 Proposed Recommender System

The recommender system proposed in this paper is articulated in two parts as shown in Figure 2. The first part is the User-based datamining that contains sample data. The second part focuses on Machine Learning algorithms, that use different regression algorithms (i.e., Linear Regression, Decision Tree, Support Vector Regression, and Random Forest Regression) to Build a predictive model for predicting students' admission in higher education. It should be noted that the parameters used in this study are GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, and Research Experience.

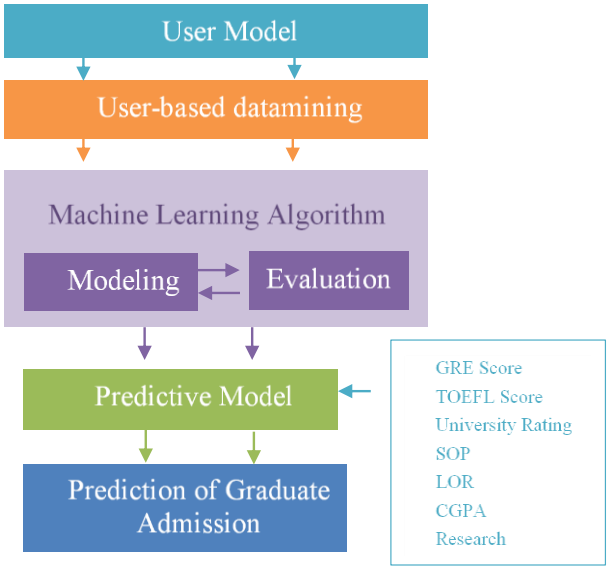

Fig. 2. Proposed Recommender System RSEPUA

After presenting the proposed recommender system RSEPUA, in the next section we will present the discussion of the obtained results.

6 Results and Discussion

In this present work, we used Supervised Machine Learning algorithms to build a predictive model for early predicting university admission. Hence, the prediction model is based on four regression algorithms which are Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR), and Random Forest Regression (RFR). All those algorithms implementation is available in the XLSTAT environment [21].

6.1 Correlation between the parameters profile and the chance of admission

We used the Pearson Correlation Coefficient (PCC) [22] to evaluate the linear correlation between the variables in our dataset. Indeed, this coefficient allows us to reduce the number of features and determine the relationship between different parameters profile (i.e., GRE Score, TOEFL Score, University Rating (UR), SOP, LOR, CGPA, and Research Experience) and the chance of admission. Table 2 illustrates the results obtained for the Pearson Correlation Coefficients. Figure 3 illustrates a correlation map of the table values.

Table 2. Correlation matrix (Pearson)

Variables	GRE Score	TOEFL Score	UR	SOP	LOR	CGPA	Research	Chance of Admit
GRE Score	1	0.827	0.635	0.613	0.525	0.826	0.563	0.810
TOEFL Score	0.827	1	0.650	0.644	0.542	0.811	0.467	0.792
UR	0.635	0.650	1	0.728	0.609	0.705	0.427	0.690
SOP	0.613	0.644	0.728	1	0.664	0.712	0.408	0.684
LOR	0.525	0.542	0.609	0.664	1	0.637	0.373	0.645
CGPA	0.826	0.811	0.705	0.712	0.637	1	0.501	0.882
Research	0.563	0.467	0.427	0.408	0.373	0.501	1	0.546
Chance of Admit	0.810	0.792	0.690	0.684	0.645	0.882	0.546	1

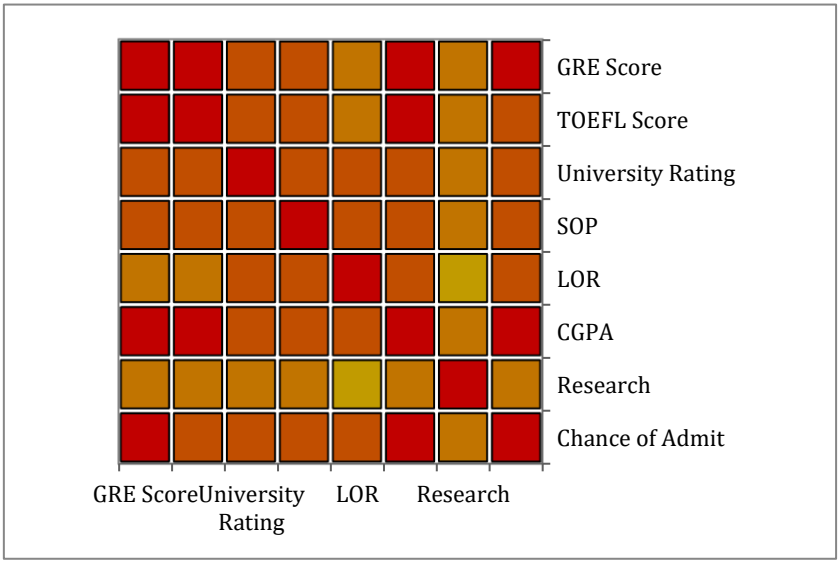

Fig. 3. Correlation maps

The results obtained demonstrate a high correlation between CGPA, GRE score, and TOEFL score and the chance of admission.

6.2 Comparison of machine learning algorithms for predicting university admission

Table 3 represents a comparison of four Machine Learning algorithms namely Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR), and Random Forest Regression (RFR) for predicting university admission. The evaluation metrics used in this comparison are Mean Square Error (MSE), Root Mean Square Error (RMSE) and R-squared (R^2).

Table 3. Evaluation of the results

	MSE	RMSE	R2
LR	0.003322947	0.057645010	0.82190074
DT	0.003992939	0.063189710	0.784975338
SVR	0.003636979	0.060307369	0.821388466
RFR	0.003004176	0.065606216	0.885856381

As table 2 shown, it is clear that Random Forest Regression (RFR) provides better performance because it has a low MSE, low RMSE, and high R2 score. The following chart (see figure 4) indicates the predicted values versus the observed values.

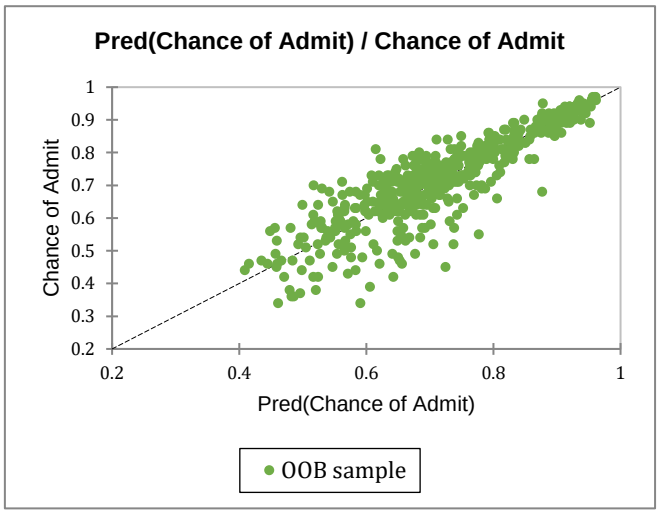

Fig. 4. Predicted values versus the observed values

6.3 Most important parameters influence the chance of admission

After identifying the most suitable Machine Learning algorithm for predicting university admission, it is therefore evident to ask which parameters have the most predictive power. Indeed, the parameters with high importance are the engines of the prediction and their values have a strong and significant impact on the outcome values. For that reason, after training and validating a random forest, the variable importance (VIMP) is calculated based on Mean increase error. The normalized variable importance measure for each variable is represented in the following Table 4. Figure 5 illustrates a graphical representation of the table values.

Table 4. Variable Importance

Variables	Mean increase error
GRE Score	14.076
TOEFL Score	6.006
University Rating	4.614
SOP	9.692
LOR	2.929
CGPA	36.589
Research	10.516

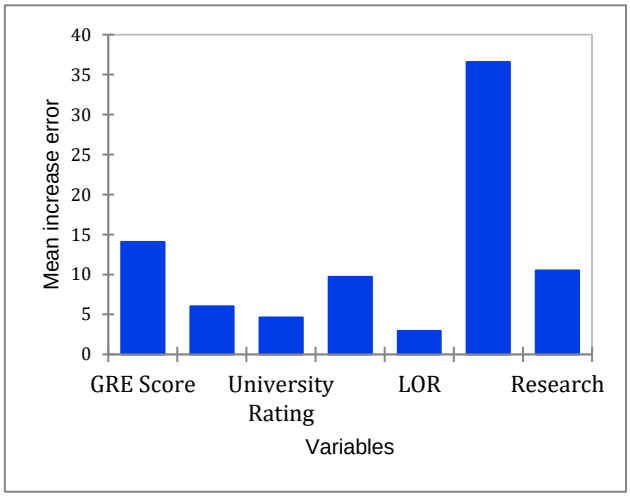

Fig. 5. Parameters influence the chance of admission

As we can see, the most important parameter is the Cumulative Grade Point Average (CGPA). Hence, it is clear that there is a very strong link between the CGPA and the chance of admission.

6.4 Proposed interfaces

In this section, we will present some interfaces of our recommender system RSEPUA, starting with the first form-based interface as shown in Figure 6. In it, the user interacts with our recommender system by entering information into the fields that accept it. This interface contains the flowing information: CGPA, GRE, TOEFL, University Rating, SOP, LOR, and Research.

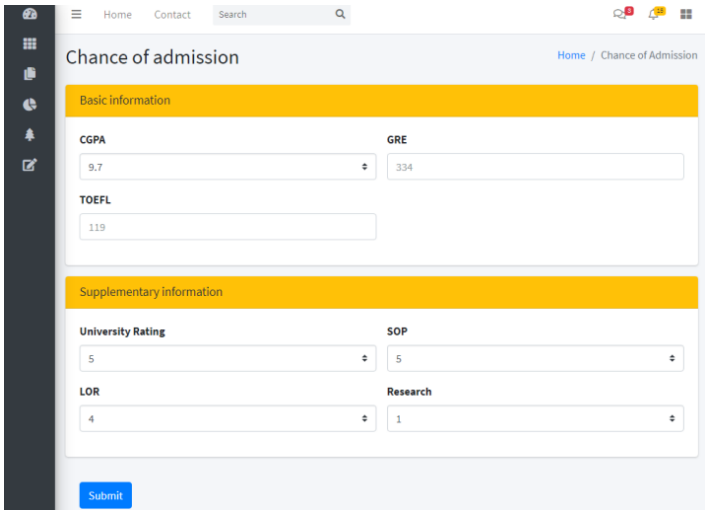

Fig. 6. Principal interface

The result changes according to the user's profile, here for this learner the system predicts that the chance of admission of this user is 95% as shown in figure 7.

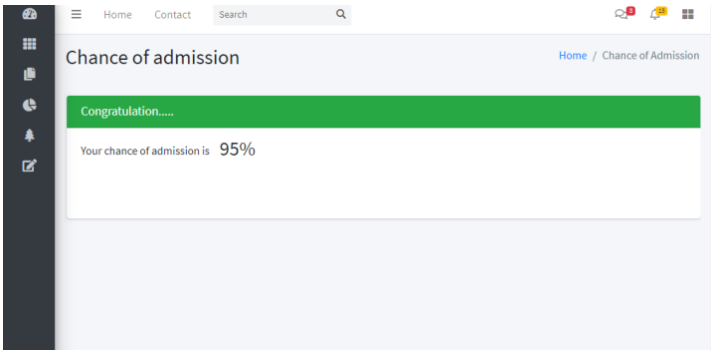

Fig. 7. Chance of admission

7 Conclusion and Future Work

Machine Learning allows us to reduce the human error probability by providing very strong recommendations, predictions, and decisions based on only the input data. For that reason, it has become one of the most important and common aspects of the digital world. Different application areas adapt and adopt Machine Learning techniques in their systems such as medicine, finance, marketing, business intelligence, healthcare, etc. In our case, we aim to design a recommender system based on Machine Learning techniques in the field of Education. Thus, the contributions were threefold: The first was to apply several Supervised Machine Learning algorithms (i.e., Linear Regression,

Support Vector Regression, Decision Tree Regression, and Random Forest Regression) on our dataset. The second purpose was to compare and evaluate algorithms used to create a predictive model based on various evaluation metrics. The last purpose was to determine the most important parameters that influence the chance of admission. The experimental results showed that the Random Forest Regression is the most suitable Machine Learning algorithm for predicting university admission. Also, the Cumulative Grade Point Average is the most important parameter that influences the chance of admission.

The major directions for future work are:

- Applying techniques such as clustering and artificial neural networks to have better predicting
- Utilizing dataset with massive size and diverse features to tackle the issue of scalability
- Exploiting few hybrid feature selection algorithms.

8 References

[1] Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. Computers in Human Behavior, 107, 105584. https://doi.org/10.1016/j.chb.2018.06.032

[2] Qazdar, A., Er-Raha, B., Cherkaoui, C., &Mammass, D. (2019). A machine learning algorithm framework for predicting students' performance: A case study of baccalaureate students in Morocco. Education and Information Technologies, 24(6), 3577-3589. https://doi.org/10.1007/s10639-019-09946-8

[3] El Guabassi, I., Al Achhab, M., Jellouli, I., & El Mohajir, B. E. (2016, October). Recommender system for ubiquitous learning based on decision tree. In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) (pp. 535-540). IEEE. https://doi.org/10.1109/cist.2016.7805107

[4] Guabassi, I. E., Achhab, M. A., Jellouli, I., & Mohajir, B. E. E. (2016). Towards adaptive ubiquitous learning systems. International Journal of Knowledge and Learning, 11(1), 3-23. https://doi.org/10.3991/ijet.v13i12.7918

[5] El Guabassi, I., Al Achhab, M., Jellouli, I., &Mohajir, B. E. E. (2018). Personalized ubiquitous learning via an adaptive engine. International Journal of Emerging Technologies in Learning (iJET), 13(12), 177-190. https://doi.org/10.3991/ijet.v13i12.7918

[6] Bousalem, Z., El Guabassi, I., &Cherti, I. (2018, July). Toward adaptive and reusable learning content using XML dynamic labeling schemes and relational databases. In International Conference on Advanced Intelligent Systems for Sustainable Development (pp. 787-799). Springer, Cham. https://doi.org/10.1007/978-3-030-11928-7_71

[7] Wu, X., & Wu, J. (2019). Criteria evaluation and selection in non-native language MBA students admission based on machine learning methods. Journal of Ambient Intelligence and Humanized Computing, 1-13. https://doi.org/10.1007/s12652-019-01490-0

[8] AlGhamdi, A., Barsheed, A., AlMshjary, H., &AlGhamdi, H. (2020, March). A Machine Learning Approach for Graduate Admission Prediction. In Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing (pp. 155-158). https://doi.org/10.1145/3388818.3393716

[9] Nandal, P. (2020). Deep Learning in diverse Computing and Network Applications Student Admission Predictor using Deep Learning. Available at SSRN 3562976. https://doi.org/10.2139/ssrn.3562976

[10] Zhao, Y., Lackaye, B., Dy, J. G., & Brodley, C. E. (2020). A Quantitative Machine Learning Approach to Master Students Admission for Professional Institutions. International Educational Data Mining Society.
[11] Di Noia, T., Mirizzi, R., Ostuni, V. C., Romito, D., & Zanker, M. (2012, September). Linked open data to support content-based recommender systems. In Proceedings of the 8th international conference on semantic systems (pp. 1-8). https://doi.org/10.1145/2362499.2362501
[12] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems (TOIS), 22(1), 5-53. https://doi.org/10.1145/963770.963772
[13] Burke, R. (2000). Knowledge-based recommender systems. Encyclopedia of library and information systems, 69(Supplement 32), 175-186.
[14] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.
[15] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674. https://doi.org/10.1109/21.97458
[16] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222. https://doi.org/10.1023/b:stco.0000035301.49549.88
[17] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
[18] Miles, J. (2014). R squared, adjusted R squared. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06627
[19] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82. https://doi.org/10.3354/cr030079
[20] Acharya, M. S., Armaan, A., & Antony, A. S. (2019, February). A comparison of regression models for prediction of graduate admissions. In 2019 International Conference on Computational Intelligence in Data Science (ICCIDS) (pp. 1-5). IEEE. https://doi.org/10.1109/iccids.2019.8862140
[21] Addinsoft, X. (2015). Data analysis and statistics with MS Excel. Addinsoft, NY, USA. xlstat available at http://www. xlstat. com/en/home.
[22] Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5

9 Authors

Inssaf El Guabassi received his PhD in February 2019 from the Abdelmalek Essaadi University, Faculty of Sciences, Morocco, Tetouan.

Zakaria Bousalem is a Phd student at the Faculty of Sciences and Technology, Settat Morocco.

Rim Marah received his PhD in July 2018 from the Abdelmalek Essaadi University, Faculty of Sciences, Morocco, Tetouan.

Aimad Qazdar is an Assistant Professor at the Faculty of Sciences Semlalia, ISI Laboratory – Cadi Ayyad University in Marrakech, Morocco.

Article submitted 2020-11-25. Resubmitted 2020-12-22. Final acceptance 2020-12-23. Final version published as submitted by the authors.