ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse57942020.pdf
https://doi.org/10.30534/ijatcse/2020/57942020

The Impact of Data Normalization on Predicting Student Performance: A Case Study from Hashemite University

Esra'a Alshdaifat[1], Ala'a Al-shdaifat[2], Aisha Zaid[3], Ahmad Aloqaily[4]
[1]Computer Information System Department, Hashemite University, Jordan, esraa@hu.edu.jo
[2,4]Computer Science Department, Hashemite University, Jordan, alaa_shdaifat, aloqaily @hu.edu.jo
[3]Software Engineering Department, Hashemite University, Jordan, Email: aisha@hu.edu.jo

ABSTRACT

The main goal of mining and analyzing educational data is to identify interesting patterns that might influence learning quality and student performance. Supervised machine learning can be utilized to predict student performance. More specifically, classification algorithms are used to construct a student performance prediction model. One of the main challenges to generate the desired model is adopting appropriate data preprocessing techniques. The central idea of the work presented in this paper is to generate a student performance prediction model and investigate the impact of various data normalization techniques on its effectiveness. The influence of various data normalization techniques is assessed on three novel academic datasets collected from the Hashemite University in Jordan. The obtained results provided empirical evidence that the adopted data normalization technique has a significant effect on classification model effectiveness. The SVM classifier coupled with the Z-score normalization technique produced the most effective student performance prediction model.

Keywords: Classification, Data normalization, Educational Data, Grade Prediction, Student Performance.

1. INTRODUCTION

Mining educational data is concerned with applying machine learning algorithms to educational datasets in order to discover interesting patterns that might be useful for improving the education field. Numerous researchers directed their work to analyze educational data due to the attractiveness of this research area. One of the most interesting goals for mining educational data is predicting student performance. The ability to predict student performance would reduce student failure or low grades achievements, thus improving the education and the reputation of educational institutions. However, generating an effective model to predict student performance is challenging and requires extensive efforts. Several issues cause predicting student performance to be a challenging research area, such as the size of educational data, the "uncleanness" of educational data and the determination of the "key" features affecting student performance. Consequently, a comprehensive methodology should be adopted to generate an effective student performance prediction model.

Generally, the first phase in generating performance prediction model is collecting educational data. Then the most sophisticated phase will be commenced by applying several data preprocessing techniques to the collected data in order to obtain a "clean" and "convenient" dataset that is ready to be fed into machine learning algorithms. Handling missing values, solving inconsistency, removing redundancy, feature selection and data normalization are considered during this phase. Among these, data normalization is essential with regard to educational datasets due to the inclusion of numeric features with varying ranges. For example, first exam mark, second exam mark, midterm exam mark, course project mark and GPA have different ranges. Thus, data normalization is needed to avoid that one feature hiding the effect of others. Data normalization is concerned with transforming numeric data into new data that has smaller values and a predetermined range. Feature selection is also a crucial issue with respect to predicting student performance. Researchers have spent abundant efforts to identify the "key" features that could generate an effective performance prediction model [1]. After preparing the dataset, the succeeding phase is to generate the prediction model by training the classification algorithm on the processed dataset. Note here that many algorithms can be utilized for this purpose such as Decision Tree, Naive Bayes, Support Vector Machine SVM and Artificial Neural Network ANN. Then an evaluation of the generated models should be accomplished. Several assessment measures might be used, and the most widely used measures are accuracy and Area Under the ROC Curve (AUC) [2]. The reason behind models evaluation is to decide their applicability to be used for predicting student performance.

The work presented in this paper is concentrated on constructing an effective performance prediction model. More specifically, the aims of the work can be summarized as follows:

1) Investigate the impact of data normalization on predicting student performance.

2) Identify the key features to be fed into classification algorithms to get an effective student performance prediction model.
3) Explore the chance of predicting student performance for any course with acceptable prediction effectiveness.

With respect to the first aim, the work presented in this paper provides a comparative study of three different data normalization techniques: (i) Min-Max normalization, (ii) Z-score normalization and (iii) Decimal normalization. The second aim is addressed by considering two different approaches. The first approach is to rely on our knowledge in the educational domain to eliminate the "irrelevant" features. The second approach is to adopt a feature selection measure to spot the dominant features. The well-known information gain measure [3] is adopted for this purpose. Regarding the third aim, three different courses are considered in the work presented in this paper: (i) Data Mining, (ii) System Analysis and Design, and (iii) Multimedia. In other words, three novel datasets are collected and utilized for the research presented in this paper.

The research work is presented in six sections, together with this introduction. The next section supplies the reader with the essential background to the work presented in this paper. The third section clarifies the adopted methodology to generate the desired performance prediction model. The fourth section presents an outline of the main characteristics of the evaluation datasets. The fifth section presents the adopted experimental settings and discusses the obtained results. Finally, and within the last section, the results are concluded.

2. LITERATURE REVIEW

In this section, we offer the reader with the essential background to the work presented in this paper. Since the main goal of the work described in this paper is to investigate the effect of normalization techniques on predicting student performance, this section is divided into two Sub-sections. Sub-section 2.1 presents an overview of predicting student performance and discusses the latest previous work associated with this issue, while Sub-section 2.2 provides a brief background of main data normalization techniques and the importance of applying data normalization prior to applying machine learning algorithms.

2.1 Previous work on predicting student performance

Recently, students, instructors and academic organizations desire to predict student performance due to its expected advantages. The potential benefits of predicting student performance can be summed up as follows [4]:

- With respect to academic organizations, the capability to predict student performance will enhance the rating and reputability of the organization.
- With respect to instructors, the capability to predict student performance will enable updated teaching plans to be considered and applied.
- With respect to students, the capability to predict student performance will reduce student failure or achieving low grades.

Consequently, many research studies tried to predict student performance using machine learning algorithms [5]-[13]. More specifically, classification algorithms were used to generate prediction models, based on historical educational data, which are able to predict the grade for new students. Several machine learning algorithms were utilized to generate the desired grade prediction model such as Decision Tree (DT), Naïve Bayes(NB), Support Vector Machines(SVM) and Artificial Neural Network ANN.

Regarding the used educational data, some researchers tried to predict student performance utilizing relatively small datasets, which include less than one hundred or few hundreds of samples, such as the work conducted by Kumar and Pal [14], Mueen et al. [15], Natek and Zwilling [16], Abu Zohair [17] and Shanthini et. al. [18]. Whereas other researchers utilized relatively big datasets that feature thousands of samples such as the work carried by Ahmed and Elaraby [19] and Kabakchieva [20].

Regardless of the size of the considered dataset, the work on predicting student performance can be differentiated according to the considered features used to generate the considered model. According to the literature, grade point average (GPA) and internal assessments (such as exam mark, assignment mark and quizzes) are the most widely used followed by student demographic (such as gender and age) and external assessments (such as final exam mark for specific a subject) [1]. In addition, many researchers utilized high school background, scholarship, extra-curricular activities and social interaction networks to predict student performance [1].

With respect to the final prediction, the work on predicting student performance can be categorized into: (i) predicting student status (pass, fail and/or dropout), (ii) predicting student dropout (yes, no) and (iii) predicting student grade (A, B, C, D, E and F). Among the previous categories, predicting student grade is challenging because of the higher number of class labels compared to other categories.

2.2 Data normalization

A significant issue when applying machine learning algorithms to a dataset including numeric attributes is that some attributes conceal the effect of others, primarily when the adopted algorithm applies some calculations on the considered features. To address this issue, normalization techniques are utilized prior to applying machine learning algorithms. Normalization is concerned with transforming the original feature values to new ones by applying a specific equation. The previous work that studied the effect of normalization on classification effectiveness showed that classification accuracy is highly affected by the adopted normalization technique [21]-[26]. Three well-known data

normalization techniques, which also implemented in most data mining tools are listed below:

• Min-Max normalization. Using the Min-Max normalization technique, the transformation process commences with identifying a "small range" desired for the new transformed values ([0,1] is commonly adopted), then the Min-Max normalization equation is applied to each old attribute value, thus a new value will be produced [3].

• Z-score normalization. Using the Z-score normalization technique, the mean and standard deviation are initially calculated for considered attribute to be used with the Z-score normalization equation to calculate a new value for each old attribute value.

• Decimal normalization. Using the Decimal normalization technique, the number of decimal point movement is determined first, then each value in the considered attribute is transformed to a new value by moving its decimal point according to the predefined movement number.

With respect to the work presented in this paper, the effect of the above three normalization techniques on predicting student grade, using three novel datasets, is investigated.

3. RESEARCH METHODOLOGY

In this section, the adopted methodology to generate the desired student performance prediction model is presented. The overall methodology is given in Figure 1, so the reader can realize the required process to build the model. From the figure, a range of phases have to be considered, these can be summarized as follows:

1) Acquiring an educational dataset. Section 4 will present the main characteristics of the three novel datasets utilized in this research.
2) Preprocessing the considered dataset, where several tasks are considered including: (i) handling missing values, (ii) solving inconsistency, (iii) removing redundancy, (iv) feature selection and (v) normalization. The details of the performed preprocessing are discussed in Section 4.
3) Applying a classification algorithm to generate the prediction model using the pre-processed educational data. Two high performance classification algorithms were utilized for this purpose: (i) Support Vector Machine SVM and (ii) Artificial Neural Network ANN.
4) Assessing the accuracy of the produced model to determine if it can be utilized to predict student grade. As noted in the introduction to this paper, accuracy and Area Under the ROC Curve (AUC) are the most widely used metrics to evaluate the prediction models, thus they were adopted for assessing the generated performance prediction models.
5) Utilizing the model to predict the grades for new students.

4. DATASET DESCRIPTION AND PREPARATION

In this section, an overview of the main characteristics of the datasets is presented. Three datasets that represent three different courses for computer science students at Hashemite University are considered in this paper. The collected data covers the time interval from the fall semester (2013/2014) to the summer semester (2016/2017). The three considered courses are: (i) Data Mining, (ii) System Analysis and Design and (iii) Multimedia. The Data Mining course dataset includes 237 records. Regarding the System Analysis and Design course dataset, 316 records are included. Moreover, the Multimedia course dataset features 467 records. It is necessary to note here that the reason for selecting these three courses is the adopted grading approach. More specifically, the three chosen datasets evaluate students based on first and second exams or midterm exam, course project and final exam. In the context of Data Mining and Multimedia, the project weight is 30 marks, while System Analysis and Design project weight is 10 marks. This will help us in evaluating the effect of project mark on predicting the final grade for students.

With respect to data preprocessing, a set of phases have been considered commencing from data understanding and cleaning to feature selection. More specifically, the first phase in the adopted preprocessing was data cleaning (removing redundancy and handling inconsistency). The second phase was handling missing values, and the third phase was feature selection as it is recognized that the generation of an effective classification model is highly affected by the features considered to build it. Thus, we adopted two methods to select the features that will be used later to build the desired student performance model. The first method was to employ our experience in the academic field to prune unnecessary features that are clearly expected to not play a role in forecasting student grade. Examples of such features are student name, student number, admission year and national number. Additionally, the final exam mark and student mark out of 100 were eradicated because the main objective of this research is to predict student grade prior to the final exam. Table 1 presents the remaining features, with a brief description of each, after adopting the first feature selection method.

The second method was to adopt the well-known information gain measure [3] to evaluate the features. Since three datasets are considered in the work presented in this paper, the information gain evaluator was applied three times. The results obtained from applying information gain evaluator for the Data Mining course dataset are presented in Table 2. The evaluator orders features in descending order, features associated with higher information gain are shown at the top. From the table, it can be noted that the highest three ranks were given to Mid Exam Mark, GPA and Project Mark features, while zero information gain value was given to Secondary Average, Taken Hours, Semester Hours which were removed.

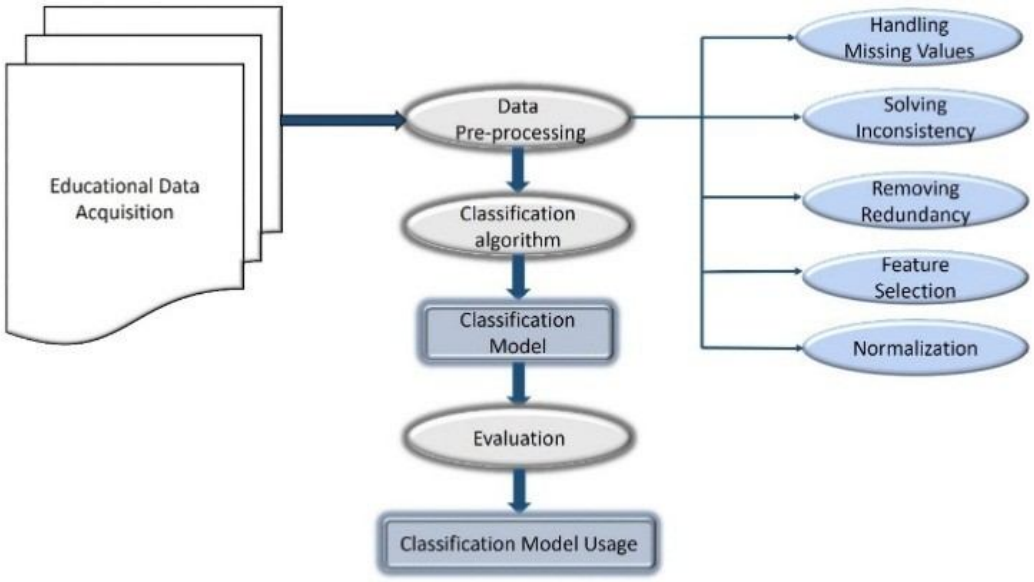

Figure 1: Generating student performance prediction model

The results obtained from applying the information gain evaluator for the Multimedia course dataset are presented in Table 3. Again, the same as the case of the Data Mining dataset, Mid Exam Mark, GPA and Project Mark features were assigned the highest information gain. Only Taken Hours, Semester Hours were assigned a value of zero information gain, thus they were eliminated. The results obtained from applying the information gain evaluator for the System Analysis and Design course dataset are presented in Table 4. From the table, it can be noted that GPA, Second Exam Mark and First Exam Mark were assigned the highest information gain. While Taken Hours, Project Mark and Semester Hours were assigned a value of zero information gain, thus they were eliminated. Unlike Data Mining and Multimedia datasets, the project mark for System Analysis and Design dataset was not considered a dominant feature. The reason behind that is the weight of the project. More specifically, the weight assigned to the project with respect to Data Mining and Multimedia datasets was higher than the System Analysis and Design dataset.

The last and the significant step in the adopted preprocessing methodology is data normalization. In this phase, three different normalization techniques were applied
to each dataset: (i) Decimal normalization, (ii) Z-score normalization and (iii) Min-Max normalization. Consequently, three variations are generated for each dataset.

5. EXPERIMENTS AND EVALUATION

This section presents an overview of the adopted experimental settings and the obtained results. The experiments were executed utilizing the Weka data mining tool [27] and Microsoft Excel. In order to acquire precise results, Ten-fold Cross Validation (TCV) was adopted during evaluation. As noted earlier accuracy and AUC measures were considered to evaluate the resulting classification models, however, the analytical study of the obtained results will be based on AUC measure. The reason behind that back to the recent studies that recommend adopting AUC measure to evaluate machine learning algorithms [2].

Commencing with the results obtained when using the two classification algorithms coupled with the alternative data normalization techniques with respect to the Data Mining course dataset. Table 5 presents the results in terms of average accuracy and AUC measures. Note here that "No normalization" was also considered when generating classification models to be used as a baseline standard for generally evaluating the effect of normalization on classification effectiveness. From Table 5, it is clearly noted that the adopted normalization technique has a significant effect on classification effectiveness. More specifically, the obtained AUC results with respect to ANN classifier range from 0.49, where no normalization was applied to the considered data, to 0.84 where Z-score normalization was applied. Additionally, the AUC results range from 0.69 to 0.86 with respect to the SVM classifier. Regarding comparing the three considered normalization techniques, Z-score normalization generated the best AUC results regardless of the used classification algorithm. It is interesting to note here that the ANN classifier was more affected by the adopted normalization technique than the SVM classifier.

Table 1: The evaluation datasets description

Feature	Brief Description	Type	Values/Range
GPA	Student Grade Point Average (GPA).	Numeric	[0.0 - 4.00]
Mid	Mid-term exam mark.	Numeric	[0 – 30]
First	First exam mark.	Numeric	[0 – 25]
Second	Second exam mark.	Numeric	[0 – 25]
Project	Course project mark.	Numeric	[0 – 10] or [0 – 30]
Secondary Average	Secondary school average grade.	Numeric	[65.0 – 99.7]
Taken Hours	The total number of hours that were completed by the student when he\she enrolled in this course.	Numeric	[33 – 132]
Semester Hours	The total number of hours that were registered by the student in the semester, when he\she enrolled in this course.	Numeric	[6-19]
Taken	Indicates whether the course was taken previously by the student.	Nominal/Boolean	{ Yes, No}
Warned	Indicates whether the student was given an academic warning during his study.	Nominal/Boolean	{ Yes, No}
Gender	Student gender.	Nominal/Boolean	{Female, Male}
Secondary Certificate Branch	With respect to secondary education in Jordan, there are different specialization.	Nominal	(Scientific, Informatics, Industrial}
Honor List	Indicates whether the student was named on the faculty honor list.	Nominal/Boolean	{ Yes, No}
Admission Type	Jordan universities have different admission types, these are mainly differentiated according to the origin of student secondary certificate.	Nominal	{Regular, others}
Semester	The semester in which the student enrolled the course.	Nominal	{ Fall, Spring, Summer1, Summer2 }
Major	Student specialization.	Nominal	{Computer Information System (CIS), Computer Science (CS), Software Engineering (SW)}
International/Local Secondary	Indicates whether student secondary certificate is international or Jordanian (local).	Nominal/Boolean	{International, Local}
Birth Place	Student birth place.	Nominal/Boolean	{Jordan, others}
Nationality	Student nationality.	Nominal/Boolean	{Jordanian, non-Jordanian}
Tuition fees	Indicates whether the student pays his/her tuition fees.	Nominal/Boolean	{Yes, No}
Grade	Student final course grade **(the class label).**	Nominal	{A, B, C, D and F}

Table 2: The information gain results for Data Mining dataset

Feature	Rank
Mid Exam Mark	0.83067
GPA	0.63641
Project	0.51957
Taken Previously	0.4566
Warned	0.1656
Gender	0.07958
Secondary Certificate Branch	0.0715
Honor List	0.06519
Admission Type	0.06075
Semester	0.03774
Major	0.02693
International/Local Secondary	0.02205
Birth Place	0.01966
Nationality	0.00583
Tuition fees	0.00511
Secondary Average	0
Taken Hours	0
Semester Hours	0

Table 3: The information gain results for Multimedia dataset

Feature	Rank
Mid Exam Mark	0.664021
GPA	0.395574
Project	0.282151
Honor List	0.125627
Taken Previously	0.100109
Secondary Average	0.092659
Warned	0.091724
Secondary Certificate Branch	0.033897
Major	0.029274
Admission Type	0.026564
Semester	0.01346
Gender	0.009368
Birth Place	0.009317
Tuition fees	0.004704
International/Local Secondary	0.002851
Nationality	0.00092
Semester Hours	0
Taken Hours	0

Table 4: The information gain results for System Analysis dataset

Feature	Rank
GPA	0.63588
Second Exam Mark	0.61793
First Exam Mark	0.56284
Taken Previously	0.43428
Warned	0.12264
Honor List	0.10576
Secondary Average	0.09844
Semester	0.05052
Admission Type	0.04323
Secondary Certificate branch	0.03266
Major	0.03125
Birth Place	0.01368
Gender	0.00715
International/Local Secondary	0.00671
Tuition fees	0.0045
Nationality	0.00409
Taken Hours	0
Project	0
Semester Hours	0

In the context of the Multimedia course dataset, Table 6 presents the obtained results when using the two classification algorithms coupled with the three alternative data normalization techniques. Again, “No normalization” was also considered when generating classification models. The same as the case of the Data Mining course dataset, the adopted normalization technique has a significant effect on classification effectiveness. However, for the Multimedia course dataset Min-Max normalization generated the best AUC result when the ANN was used to generate the classification model, while Z-score normalization produced the best AUC result when the SVM was used to generate the classification model. With respect to the System Analysis and design dataset, Table 7 presents the obtained results when using the two classification algorithms coupled with the three alternative data normalization techniques. From the table, it can be noted that Z-score normalization outperforms other normalization techniques and also no normalization when the ANN classification algorithm was adopted to generate the classification models. While the AUC results obtained when using the SVM classifier were the same for No Normalization, Z-score and Min-Max normalization.

Table 5: Average accuracy and AUC values obtained using Data Mining course dataset

Data Mining Course Dataset								
Pre-Processing Technique	No Normalization		Decimal		Zscore		MinMax	
Classifier	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Artificial Neural Networks (MultilayerPerceptron)	39.24%	0.49	56.12%	0.78	62.45%	**0.84**	60.34%	0.82
Support Vector Machines (SMO)	67.09%	0.85	49.79%	0.69	67.51%	**0.86**	65.82%	0.84

Table 6: Average accuracy and AUC values obtained using Multimedia course dataset

Multimedia Course dataset								
Pre-Processing Technique	No Normalization		Decimal		Zscore		MinMax	
Classifier	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Artificial Neural Networks (MultilayerPerceptron)	38.33%	0.50	57.60%	0.79	38.33%	0.50	58.03%	**0.82**
Support Vector Machines (SMO)	69.16%	0.86	48.39%	0.66	69.16%	**0.86**	65.52%	0.84

Table 7: Average accuracy and AUC values obtained using System Analysis and Design course dataset

System Analysis and Design dataset								
Pre-Processing Technique	No Normalization		Decimal		Zscore		MinMax	
Classifier	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Artificial Neural Networks (MultilayerPerceptron)	39.56%	0.50	59.81%	0.82	66.46%	**0.84**	60.13%	0.83
Support Vector Machines (SMO)	70.57%	**0.86**	57.59%	0.71	69.94%	**0.86**	70.89%	**0.86**

Considering the results presented in Tables 5, 6 and 7, student performance could be successfully predicted using classification models coupled with appropriate preprocessing technique, particularly normalization techniques. More specifically, the SVM classifier coupled with the Z-score normalization generated the highest AUC result (0.86) with respect to the three considered datasets (Data Mining Course, Multimedia Course and System Analysis and Design Course).

6. CONCLUSION AND FUTURE WORK

In this paper, generating a student performance prediction model and investigating the impact of three data normalization techniques on its effectiveness have been considered. The performance prediction model has been generated using ANN and SVM classification algorithms for three different datasets representing three different computer science courses. The reasons behind using three alternative courses were to examine the possibility of predicting student performance for any course with acceptable classification effectiveness, and to determine the common dominant features that affect predicting performance. The experimental results presented earlier in this paper indicated that the effectiveness of the performance prediction model significantly affected by the adopted data normalization technique. Strikingly, the Z-score normalization technique outperformed Min-Max and Decimal data normalization techniques. Regarding the key features that affect predicting performance, it was found that exams marks and GPA were the most significant features. In addition, it was noted that course project mark could be considered as one of the key features that affect predicting performance only if it is assigned with a relatively high weight. Overall, SVM classifier coupled with Z-score normalization was the most effective aggregation with respect to the three considered datasets (Data Mining Course, Multimedia Course and System Analysis and Design Course). As future work, the authors plan to generate different variations of student

performance prediction model by adopting completely different labeling approaches.

ACKNOWLEDGEMENT

The authors would like to thank the Hashemite University for the continuous support.

REFERENCES

[1] A. Shahiri, W. Husain and N. A. Rashid, **A review on predicting student's performance using data mining techniques**, *Procedia Computer Science*, vol. 72, pp. 414-422, 2015. https://doi.org/10.1016/j.procs.2015.12.157

[2] J. Demsar, **Statistical comparisons of classifiers over multiple data sets**, *Journal of Machine Learning Research*, pp. 1-30, 2006.

[3] J. Han, M. Kamber and J. Pei, ***Data Mining: Concepts and Techniques***, Morgan Kaufmann, 2011.

[4] U. Bin Mat, N. Buniyamin, . P. Arsad and R. Kassim, **An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention**, in Proc. *2013 IEEE 5th Conference on Engineering Education (ICEED)*, 2013.

[5] O. Adejo and T. Cannolly, **Predicting student academic performance using multi-model heterogeneous ensemble approach**, *Journal of Applied Research in Higher Education*, vol. 10, no. 1, pp. 61-75, 2017. https://doi.org/10.1108/JARHE-09-2017-0113

[6] R. Asif, A. Merceron, S. Ali and N. Haider, **Analyzing undergraduate students' performance using educational data mining**, *Computers & Education*, vol. 113, pp. 177-194, 2017.

[7] E. Costa, B. Fonseca, M. Santana, F. de Araujo and J. Rego, **Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses**, *Computers in Human Behavior*, vol. 73, pp. 247 - 256, 2017.

[8] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho and G. Cordeiro Galv~ao van Erven, **Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil**, *Journal of Business Research*, pp. 335-343, 2019.

[9] B. Francis and S. Babu, **Predicting academic performance of students using a hybrid data mining approach**, *J. Med. Syst.*, vol. 43, no. 6, pp. 1-15, 2019. https://doi.org/10.1007/s10916-019-1295-4

[10] N. Yassein, R. Helali and S. Mohomad, **Predicting student academic performance in ksa using data mining techniques**, *Journal of Information Technology & Software Engineering*, vol. 7, no. 5, 2017.

[11] A. O. Gamao and B. D. Gerardo, P**rediction-Based Model for Student Dropouts using Modified Mutated Firefly Algorithm**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 6, pp. 3461-3469, 2019. https://doi.org/10.30534/ijatcse/2019/122862019

[12] J. S. Gil, A. J. P. Delima and R. N. Vilchez, **Predicting Students' Dropout Indicators in Public School using Data Mining Approaches**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 774-778, 2020. https://doi.org/10.30534/ijatcse/2020/110912020

[13] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson and D. Murray, **Identifying key factors of student academic performance by subgroup discovery**, *International Journal of Data Science and Analytics*, vol. 7, no. 3, pp. 227-245, 2019.

[14] S. P. S. Kumar, S. Kumar and S. Pal, **Data mining: A prediction for performance improvement of engineering students using classification**, *World of Computer Science and Information Technology Journal*, vol. 12, pp. 51-56, 2012.

[15] A. Mueen, B. Zafar and U. Manzoor, **Modeling and predicting students ' academic performance using data mining techniques**, *I.J. Modern Education and Computer Science*, vol. 11, pp. 36-42, 2016.

[16] S. Natek and M. Zwilling, **Student data mining solution-knowledge management system related to higher education institutions**, *Expert Syst. Appl.*, vol. 41, pp. 6400-6407, 2014. https://doi.org/10.1016/j.eswa.2014.04.024

[17] L. Abu Zohair, **Prediction of student's performance by modelling small dataset size**, *International Journal of Educational Technology in Higher Education*, vol. 16, 2019.

[18] A. Shanthini, G. Vinodhini and R. Chandrasekaran, **Predicting student's academic performance in the university using meta decision tree classifiers**, *JCS*, vol. 14, no. 5, pp. 654-662, 2018.

[19] A. Ahmad and I. Elaraby, **Data mining: A prediction for student's performance using classification method**, *World Journal of Computer Application and Technology*, vol. 2, no. 2, pp. 43-47, 2014.

[20] D. Kabakchieva**, Predicting student performance by using data mining methods for classification,** *Cybernetics and Information Technologies*, vol. 13, no. 1, pp. 61-72, 2013.

[21] S. Crone, S. Lessmann and R. Stahlbock, **The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing**, *European Journal of Operational Research*, pp. 781-800, 2006. https://doi.org/10.1016/j.ejor.2005.07.023

[22] J. T. and S. A., **Impact of preprocessing for diagnosis of diabetes mellitus using artificial neural networks**, in *Proc. Second International Conference on Machine*

Learning and Computing, 2010.

[23] B. Singh, K. Verma and A. Thoke, **Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification,** ***International Journal of Computer Applications***, vol. 116, no. 19, pp. 11-15, 2015.

[24] H.C. Huang and L.X. Qin, **Empirical evaluation of data normalization methods for molecular classification**, *PeerJ*, vol. 6, p. e4584, 2018. https://doi.org/10.7717/peerj.4584

[25] A. Alizadeh Naeini, M. Babadi, S. Homayouni, **Assessment of normalization techniques on the accuracy of hyperspectral data clustering**, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-4/W4, pp. 27-30, 2017.

[26] O. Rozenstein, T. P. Kagan, C. Salbach and A. Karnieli, **Comparing the effect of preprocessing transformations on methods of land-use classification derived from spectral soil measurements**, *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, pp. 1-12, 2014.

[27] H. Mark, F. Eibe, H. Geoffrey, P. Bernhard, R. Peter and W. Ian, **The weka data mining software: An update**, *SIGKDD Exploration*, vol. 11, no. 1, pp. 10-18, 2009. https://doi.org/10.1145/1656274.1656278