Predicting Academic Performance through Data Mining: A Systematic Literature

Alfredo Daza, Carlos Guerra, Noemí Cervera, Erwin Burgos

Systems and Computer Engineering, Universidad Nacional del Santa, Nuevo Chimbote, Peru

Abstract **– The main objective of this work is to make a systematic review of the literature on the prediction of the academic performance of university students by applying data mining techniques. For this purpose, an exhaustive search was carried out and after the analysis of the documentation collected, aspects such as: methodology, attributes, selection algorithms, techniques, tools, and metrics were considered, which served as the basis for the elaboration of this document. The results of the study showed that the most used methodology is KDD(database knowledge extraction), the most important attribute to achieve prediction is CGPA(academic performance), the most commonly used variable selection algorithm is InfoGain-AttributeEval, among the most efficient techniques are Naïve Bayes, Neural Networks (MLP) and Decision Tree (J48), the most used tools for the development of the models is the Weka software and finally the metrics necessary to determine the effectiveness of the model were Precision and Recall.**

Keywords **– data mining, academic performance, academic performance in college students, prediction**

1. Introduction

In education, academic performance is an important factor for both students and academic institutions.

DOI: 10.18421/TEM112-57
https://doi.org/10.18421/TEM112-57

Corresponding author: Alfredo daza Vergaray,
Systems and Computer Engineering, Universidad Nacional del Santa, Nuevo Chimbote, Peru.
Email: adaza@uns.edu.pe

Received: 11 March 2022.
Revised: 11 May 2022.
Accepted: 17 May 2022.
Published: 27 May 2022.

Where student achievement is the measure that defines the quality of the teaching offered, that is why identifying the factors that affect such performance becomes a priority for institutions [16].

On the other hand, to know the level of academic performance of students, indicators are needed to define it, such as the GPA, grade point average, which is used as a tool to evaluate student performance [9].

In order to improve academic performance, the idea of making the prediction through Data Mining (DM) was born. In such a way that predicting performance is of great benefit to students and teachers, since they obtain the necessary information for decision-making with the aim of improving academic performance [18]

For the specific field of education, there is Educational Data Mining (EDM), which studies and develops various methods to analyze data in this environment [2]. The aim of researchers in this field is to discover useful knowledge, to help educational institutions to better manage their students or to help students to better manage their education and their results in order to improve their performance [1].

Currently, universities use many technological resources, such as learning management systems or information systems for students, allowing the generation of a large amount of academic data [7], and this makes the use of technologies such as EDM feasible.

Thus, one of the key areas of the EDM is the development of student performance prediction models, which allow predicting student performance in educational institutions [21].

After analyzing various literatures related to data mining, education, and studies on the factors that affect academic performance in university students, this study aims to answer the following question: What aspects are considered in the prediction of academic performance through the application of data mining? To solve this approach, a systematic review of the literature covering periods from 2015 to 2020 has been proposed.

The work is divided into five sections. The following section describes the methodology used for the systematic review of the literature and raises the

questions for research. Next, in section 3, the materials and methods. In section 4, result and discussion about the findings and finally in section 5, the conclusion is found.

2. Reading Review History

A set of research works have been carried out that allow systematic review of the literature on the prediction of the academic performance. Table 1. shows some studies that apply data mining techniques.

Table 1. Classification work to predict academic performance

Description	Reference
Investigation about information Systems Students' Study Performance Prediction Using Data Mining Approach	[16]
Work about predicting Academic Performance of Students in the UAE Using Data Mining Techniques	[19]
Review of predicting student performance in higher education institutions using decision tree analysis	[18]
Research on prediction Model for Classifying Students Based on Performance using Machine Learning Techniques	[2]
Work about educational Data Mining & Students' Performance Prediction	[1]
Study about predicting Critical Courses Affecting Students' Performance: A Case Study	[7]
Investigation about tracking Student Performance in Introductory Programming by Means of Machine Learning	[21]

3. Materials and Methods

For the development of the present systematic literature, the protocol was used (Kitchenham, 2004).

To contextualize the research related to the prediction of academic performance and to understand the degree of information about it, the questions have been posed based on two central points: a) Predictive models based on Data Mining and b) Prediction of Academic Performance. Likewise, the structure of the questions is given from three points of view: population, intervention and results.

2.1. Review Planning

2.1.1. Research questions

These questions are important to contextualize research related to the prediction of academic performance and understand the degree of information about it. That is why the questions were posed based on two central points: a) Predictive models based on Data Mining and b) Prediction of Academic Performance. Table 2. shows the structure of the question given from three points of view: Population, Intervention and Result.

Table 2. Structure of the question

Criterion	Details
Population	University students (academic performance)
Intervention	Prediction Methods/Techniques
Result	Prediction accuracy, Best prediction techniques

Therefore, the questions proposed for the present study were:

- Question 1 (Q1): What is the most commonly used methodology for applying data mining?
- Question 2 (Q2): What attributes were considered for the prediction of academic performance?
- Question 3 (Q3): Which variable selection algorithms are most commonly used?
- Question 4 (Q4): What were the techniques used in the prediction and which had better results in their accuracy?
- Question 5 (Q5): Which tools are the most concurrent for the development and testing of the predictive model?
- Question 6 (Q6): What metrics are used to determine the effectiveness of prediction techniques?

2.1.2. Search criteria

With the intention of finding and hosting relevant literature that can cover the questions posed above, the inclusion and exclusion criteria expressed in Table 3. were considered.

The articles published in the last 5 years were considered, that is, from March 2015 to March 2020, which is the moment in which the elaboration of this literature begins. The databases taken into account for the search were: IEEE Xplore, Science Direct, Springer Link, ResearchGate and Social Science Research Network, using terms such as Academic Performance, Data Mining Techniques, Educational Data Mining, Machine Learning and Academic Performance Prediction as keywords.

Table 3. Table of inclusions and exclusions

Inclusion	Documents considered:	Factors that influence academic performance.
		Prediction through data mining.
		Academic performance prediction model.
Exclusion	Documents not considered:	University students as the only study population.
		Published outside the search period range.

Based on the inclusions and exclusions considered, the process for searching for the articles is shown in:

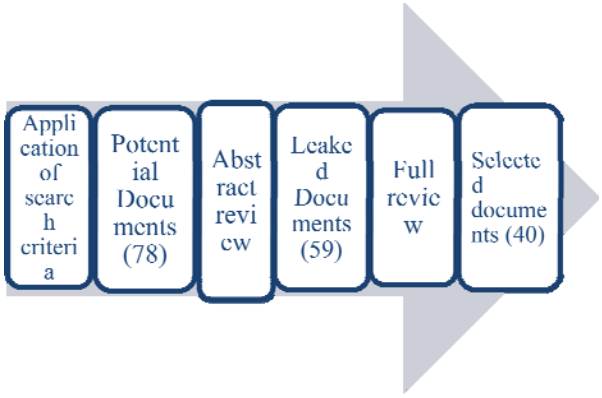

Figure 1. Process for the selection of articles

Taking into account the search period for the selection of the literature, the number of articles published in each year is shown in Figure 2.

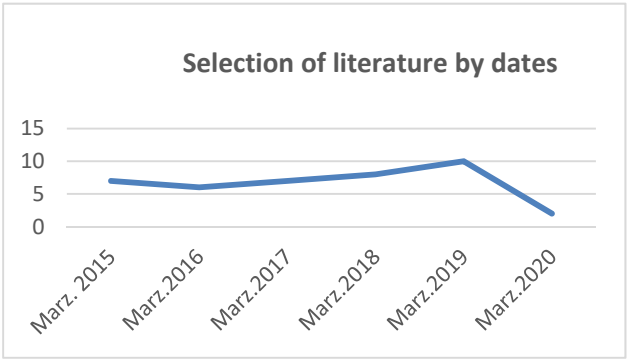

Figure 2. Publications collected by year

2.2. Conducting the Review

The initial collection of literature covered 78 articles, of which 38 were excluded for any of the following reasons:

- The central point of the research is student dropout.
- The investigation is not detailed, since the documentation has less than 5 sheets.
- They do not answer the questions initially raised.

Table 4. shows the articles that were selected of the indexed journals after having carried out the search:

Table 4. Potential, filtered and selected articles

Repository	Potential articles	Filtered articles	Selected articles	Percentage
Social Science Research Network	6	3	3	7.5%
Springer Link	11	8	2	5%
Science Direct	8	7	2	5%
Research Gate	16	12	8	20%
IEEE Explore	37	28	25	62.5%
Total	78	59	40	100%

4. Results and Discussion

Results

According to the selected documentation, it was possible to house 5 aspects related to the research that will answer the questions posed: methodologies, attributes/ factors, techniques, tools and metrics. That is why, to refer to the selected articles in each of the aspects, an identifier was assigned, as indicated in Table 5.

Table 5. Potential, filtered and selected articles

Identifier	Article	Identifier	Article
[1]	(Abu Saa, 2016)	[21]	(Khan, Al Sadiri, Ahmad, & Jabeur, 2019)
[2]	(Aggarwal, Mittal, & Bali, 2019)	[22]	(Kitchenham, 2004)
[3]	(Al Breiki, Zaki, & Mohamed, 2019)	[23]	(López, León, & González, 2015)
[4]	(Al-Barrak & Al-Razgan, 2016)	[24]	(Ma & Zhou, 2018)
[5]	(Almarabeh, 2017)	[25]	(Mahboob, Irfan, & Karamat, 2016)
[6]	(Alsalman, Khamees Abu Halemah, AlNagi, & Salameh, 2019)	[26]	(Merchan Rubiano & Duarte Garcia, 2015)

[7]	(Altujjar, Altamimi, Al-Turaiki, & Al-Razgan, 2016)	[27]	(Moine, Haedo, & Gordillo, 2011)
[8]	(Amornsinlaphachai, 2016)	[28]	(Mueen, Zafar, & Manzoor, 2016)
[9]	(Anzer, Tabaza, & Ali, 2018)	[29]	(Pang, Judd, O'Brien, & Ben-Avie, 2017)
[10]	(Azizah, Pujianto, & Nugraha, 2018)	[30]	(Puarungroj, Boonsirisumpun, Pongpatrakant, & Phromkhot, 2018)
[11]	(Bhutto, Farah, Ali, & Anwar, 2020)	[31]	(Punlumjeak & Rachburee, 2015)
[12]	(Canagareddy, Subarayadu, & Hurbungs, 2018)	[32]	(Rawat & Malhan, 2019)
[13]	(Chauhan, Shah, Karn, & Dalal, 2019)	[33]	(Rimadana, Kusumawardani, Santosa, & Erwianda, 2019)
[14]	(Chiheb, Boumahdi, Bouarfa, & Boukraa, 2017)	[34]	(Samuel, Hutapea, & Jonathan, 2019)
[15]	(Costa, Fonseca, Santana, de Arajo, & Rego, 2017)	[35]	(Segura & Loza, 2017)
[16]	(Gunawan, Hanes, & Catherine, 2019)	[36]	(Sivasakthi, 2017)
[17]	(Halde, Deshpande, & Mahajan, 2016)	[37]	(Toppireddy, Saini, & Hada, 2019)
[18]	(Hamoud, Hashim, & Awadh,	[38]	(Wang, Zhang, & Fu, 2018)
[19]	(Hamoud, Humandi, Awadh, & Hashim, 2017)	[39]	(Widyahastuti & Tjhin, 2017)
[20]	(Ketui, Wisomka, & Homjun, 2019)	[40]	(Widyaningsih, Fitriani, & Sarwinda, 2019)

3.1. Q1: What is the Most Commonly used Methodology to aApply Data Mining?

Methodologies are of great importance for the development of prediction models, since they not only reflect the phases of the process that must be followed, but also define the tasks to be performed and the way in which they should be carried out (Moine, Haedo & Gordillo, 2011). For the application of data mining there are several methodologies available and some are more used than others. As in this case, according to Table 6., it is observed that the most used one is the KDD. This happens because the phases that make it up are iterative and interactive. 26 out of 40 articles use this methodology, which makes it the most used when predicting academic performance.

Table 6. Data mining methodologies

Code	Methodology	Article
M01	KDD	[1], [2], [23], [32], [9], [19], [28], [17], [26], [23], [5],[33], [38], [11], [4], [20], [3], [25], [31], [32], [15], [34], [7], [37], [18], [31]
M02	CRISP-DM	[14], [6], [16]
M03	SEMMA	

3.2. Q2: What Attributes were Considered for the Prediction of Academic Performance?

As an answer to this question, we managed to identify 36 attributes housed in 3 categories: personal, academic and socio-economic. These categories were selected according to the concurrence of the attributes presented in the documentation, since in some cases such as the studies of [17], [27] or [36], the study is focused on a certain factor, whether psychological, economic, demographic, among others.

3.2.1. Personal factors

It is constituted by those characteristics related to the student's own behaviour. We identified 16 attributes, which represent the largest percentage of the total attributes. In Table 7., the concurrence of each attribute is also expressed, making it clear that the attributes Gender and Age have a great relevance in the prediction of academic performance. The variable Gender is used in most studies because it greatly influences academic performance, as in the articles by [3], [5], [6] and [11], where it is shown that this attribute is part of the set of most influential attributes in prediction. According to these studies, there is a greater possibility of approving if the gender is female, and this can be justified due to the learning method that women apply within their study period and this together with qualities of responsibility, make the teaching-learning process more effective in them.

Table 7. Attributes of the personal factor

Code	Attribute	Article
PF01	Gender	[1], [2], [30], [9], [21], [19], [8], [31], [26], [23],[34], [36], [25] , [38], [11], [29], [24], [6], [15] , [40]
PF02	Age	[19], [35], [25] , [23], [29], [24], [6] , [15], [40]
PF03	Marital status	[19], [26], [1], [15], [24], [6]
PF04	Country or City of Origin	[23], [1], [11], [29], [10]
PF05	Date of birth	[14]
PF06	Scholarship	[35], [29], [6]
PF07	Health Status	[24], [6], [40]
PF08	Place of Residence	[1], [34] , [29], [15], [40]
PF09	Address	[14], [24], [32]
PF10	Disability	[10]
PF11	Motivation	[40]
PF12	Employment Status	[19], [6], [40]
PF13	In a Romantic Relationship	[24]
PF14	Has Internet	[24]
PF15	Social class	[23], [26]
PF16	Transport	[1], [6], [40]

3.2.2. Academic factors

Table 8. is constituted by those attributes related to the development of learning during the student's period of study. In this factor, the attributes that were most frequently used are related to the admission test, the total credits and the high school average. In addition, the GPA is definitely the most used attribute and this is because it has a tangible value, which is directly related to the performance of students and that is why it is considered in 60% of the studies. On the other hand, the Total credit variable, as shown by [9] and [12] in their articles, is one of the attributes most correlated with the output variable (Pass or Fail).

Table 8. Attributes of the academic factor

Code	Attribute	Article
AF01	Admission Exam Score	[17], [23]
AF02	Total Credits	[18], [19], [23], [29], [10], [20]
AF03	CGPA	[21], [30], [18], [19], [35] , [8], [28] ,[17], [23], [5], [1], [36], [11] , [4],[14], [29] , [25], [31], [32] , [39], [40]
AF04	GPA High School	[35], [1], [36], [6]
AF05	Assistance	[21], [28], [5], [25], [39]
AF06	Faculty	[30], [6]
AF07	Participation	[28], [5], [11], [25]

3.2.3. Socio economic factors

Table 9. shows the characteristics related to the social environment of the student, with his way of facing and solving the economic expenses that arise during his period of university study. Among the most influential attributes are: the monthly family income, the education and occupation of parents.

Table 9. Attributes of the socio-economic factor

Code	Attribute	Article
AF01	Family Size	[1], [25], [6], [40]
AF02	Father's Work	[18], [1], [25], [32], [40]
AF03	Mother's Work	[2], [1], [24], [32], [40]
AF04	Father's Education	[1], [24], [32], [40]
AF05	Mother's Education	[1], [24], [40]
AF06	Number of friends	[2], [1]
AF07	Marital Status of Parents	[1], [24]
AF08	Monthly Family Income	[2], [8], [1], [15], [40]

3.3. *Q3: Which Variable Selection Algorithms are Most Commonly Used?*

The selection of variables is a stage that can improve or not the accuracy of the models, since the algorithms used determine the variables with the highest correlation to the class or output variable. By finding these attributes that have the greatest impact, the model can deliver better results that are closer to reality. That is why in several studies they apply this stage to their methodology. Table 10. shows those algorithms that are most used by researchers and within them, it is the InfoGain-AttributeEval the most used one.

Table 10. Methodologies for variable selection

Code	Methodology	Article
SV01	Chi - squared	[34]
SV02	CorrelationAttributeEval	[21], [18], [19]
SV03	GainRatio-AttributeEval	[11]
SV04	InfoGain-AttributeEval	[21], [15] , [24]
SV05	CFSSUBsetEval	[21]
SV06	mRmR	[31], [37]

3.4. *Q4: What Were the Techniques used in the Prediction and Which Had Better Results in their Accuracy?*

For the prediction of academic performance, data mining techniques can be employed in two ways, classification and regression. In relation to the analysis of the studies, most of them used the

techniques to classify a student, according to the output variable considered by each author. There are several algorithms to achieve this task and according to Table 11., the most used were Naive Bayes, Artificial Neural Networks, Decision Trees and Support Vector Machine. In addition, of the algorithms corresponding to the Decision Tree, J48 and Random Forest were the most used ones.

Table 11. Prediction techniques

Code	Technique	Article
TP01	One-R	[21]
TP02	C4.5	[8], [28], [23], [1], [10], [16], [34]
TP03	Naive Bayes (NB)	[2], [21], [19], [8], [28], [25], [5] , [12], [23], [5], [31], [32], [1], [33], [10], [36], [40]
TP04	Bayes Networks	[19], [5]
TP05	Decision trees	[35], [17], [33], [24], [31]
TP06	Neural Networks (MLPs)	[2], [21], [8], [28], [17], [32], [5], [31], [15], [33], [36], [6], [40]
TP07	Vector Support Machine (SVM)	[2], [13], [12], [15], [33], [29], [24]
TP08	J48	[2], [30], [18], [21], [32], [26], [12] , [5], [36], [4], [14], [6], [5] , [25]
TP09	JRip Rule	[21], [8]
TP10	Logistic regression (LR)	[2], [12], [11]
TP11	ID3	[8], [5], [1], [20]
TP12	Cart	[1]
TP13	RandomTree	[21], [18], [35], [20]
TP14	REPTree	[18], [21], [36]
TP15	K-Neighbors	[13], [8], [31], [32]
TP16	Minimum Sequential Optimization (SMO)	[21], [11], [36]
TP17	Random Forest (RF)	[2], [13], [21], [35], [12], [25], [33]
TP18	Logistic Classifier (LC)	[12]
TP19	Chi-square automatic interaction detection (CHAID)	[1]
TP20	Microsoft Decision Trees	[38]
TP21	Decision Tree Regression	[13]
TP22	Multiple Linear Regression	[13]
TP23	IBK	[21]
TP24	Decision Table	[21]
TP25	ZeroR	[21]
TP26	Linear Regression	[9], [38]
TP27	Gradient Boosted Trees	[35], [20]
TP28	Decision Stump	[35]
TP29	Bayesian Belief Network	[8]
TP30	PART	[26]
TP31	Decision Tree Weight-Based	[20]

To determine among the aforementioned prediction methods, which is the one that obtains the best accuracy, the summary of the results is shown in Table 12. The algorithms that had a better precision, are closely related to the size of the data that was used in the studies, so it can be said that the decision tree algorithms tend to have greater precision when the data set is smaller, quite the opposite, to the SVM, Naïve Bayes or MLP algorithms, which achieve greater accuracy when working with a large amount of data.

Table 12. Precision of techniques

Data size	Technique	Accuracy (%)	Article
131 students	Naive Bayes	75.50	[2]
	Neural Networks (MLPs)	92.30	
	Vector Support Machine (SVM)	70.99	
	J48	74.00	
	Logistic Regression (LR)	73.20	
	Random Forest (RF)	92.30	
4968 records	J48	84.62	[30]
50 students	Naïve Bayes	94.37	[39]

	BayesNet	92.21	
	JRip	97.84	
	J48	99.13	
161 students	J48	62.90	[18]
	Random Tree	59.70	
	REPTree	62.30	
161 students	Naïve Bayes	70.60	[19]
	Bayes Network	64.30	
1115445 records	Decision Tree	59.85	[35]
	Random Forest	52.98	
	Gradient Boosted Trees	67.41	
	Decision Stump	52.98	
	Random Tree	48.96	
150 students	Artificial Neural Network	99.99	[17]
	Decision Tree	93.35	
932 students	J48	82.41	[26]
	PART	66.33	
2000 records	Naive Bayes	94	[12]
	J48	98	
	Logistic Classifier	94	
1532 records	Naive Bayes	85	[23]
225 records	Naive Bayes	91.10	[5]
	Bayes Network	92	
	Neural Networks (MLPs)	90.40	
	J48	91.40	
	ID3	88.40	
17000 records	C4.5	63.80	[10]
	Naive Bayes	64.30	
161 students	Naive Bayes	86	[15]
	SVM	92	
	Neural Networks (MLPs)	88	
	J48	87	
785 records	C4.5	76.2	[16]
17875 records	Decision Tree	91.03	[20]
	ID3	89.66	
	Random Tree	84.14	
	Gradient Boosted Trees	92.41	
	Decision Tree Weight-Based	84.14	
60 students	Naive Bayes	86.30	[25]
	J48	93.90	
	Random Forest	100	
6882 records	Naive Bayes	80.09	[31]
	Decision Tree	87.94	
	Neural Networks (MLPs)	90.91	
	K-Nearest Neighbor	91.12	
1794 students	C4.5	83.05	[34]
300 students	Neural Networks (MLPs)	93.23	[36]
	Naive Bayes	84.46	
	SMO	90.03	
	J48	92.03	
	REPTree	91.03	
140 students	Naive Bayes	96	[40]
100 records	ID3	80	[7]
27 records	J48	85.18	[32]
	NB	81.48	
	IBK	88.88	
	ANN	88.88	
145 records	SMO	84.83	[3]
	SMOReg	96.98	
Not specified	Neural Networks (MLPs)	74	[13]
	Vector Support Machine (SVM)	74.3	
	Random Forest	84.6	
	K-Nearest Neighbor	87	
50 records	OneR	71	[21]
	Naive Bayes	84	
	J48	88	
	JRip	81	
	Random Tree	81	
	SMO	71	
	Random Forest	81	
	ZeroR	59	
182 records	Linear Regression	Not specified	[9]
474 students	C4.5	74.89	[8]
	Naive Bayes	72.10	
	Neural Networks (MLPs)	65.30	
	JRip	73.50	
	ID3	62.20	
	K-Nearest Neighbor	98.80	
Not	C4.5	79.20	[28]

specified	Naive Bayes	86	
	Neural Networks (MLPs)	82.70	
270 records	C4.5	35.19	[1]
	Naive Bayes	36.40	
	ID3	33.33	
	Cart	40	
	CHAID	34.07	
125 students	Naive Bayes	73.60	[33]
	Decision Tree	68	
	Neural Networks (MLPs)	75.20	
	SVM	80	
	Random Forest	77.80	
2936 students	Microsoft Decision Trees	95.10	[38]
500 records	Logistic Regression	73.40	[11]
	SMO	79.30	
236 students	J48	Not specified	[4]
Not specified	J48	83.33	[14]
350 students	SVM	85	[29]
649 students	Decision Tree	93	[24]
	SVM	97	
Not specified	Linear Regression	83	[37]
	CART	89	
	Random Forest	91.4	
	Quantile Random Forest	88.2	
	Rotated Regression Random Forest	85.2	
524 students	Neural Networks (MLPs)	97	[6]

3.5. Q5: Which Tools are the Most Concurrent for the Development and Testing of the Predictive Model?

There are several tools to develop data mining, and each of them presents different options for data management. Those most frequently used to carry out the development of a model for predicting academic performance are shown in Table 13. According to the results, the Weka tool is the one that has the greatest reception when developing data mining, since, of 40 studies analyzed, there were 22 that used this tool.

Table 13. Tools for data mining

Code	Tool	Article
H01	Weka	[2], [18], [21], [19], [28], [26], [12] , [23], [5], [34],[1], [11], [4], [14], [16], [6], [10], [15], [25], [32], [36], [39]
H02	Rapid Miner	[9], [35], [23], [1], [31]
H03	R Programming	[37], [29]
H04	Python	[33], [11], [29], [24]
H05	SQL Server 2008 Data Mining	[38]

3.6. Q6: What Metrics are Used to Determine the Effectiveness of Prediction Techniques?

Table 14. shows the metrics that are used to evaluate the performance of prediction models. The values obtained in each metric, clarify if the model that is being developed, achieve results that are similar to the data of reality. There are several metrics applied by the authors of the different studies, of which the one that has the greatest reception is the precision metric (Precision). This is because it helps to measure the quality of the predictive model. Other metrics that were also the most used are: Accuracy, Recall and F-mesure.

Table 14. Metrics for data mining

Code	Metric / Indicator	Article
I01	Accuracy	[16], [32], [20], [10], [40]
I02	Precision	[2], [30], [9], [18], [21], [35], [8], [28], [17], [23], [5],[1], [33] , [38], [11], [14], [29], [24], , [6] , [10], [16], [20], [25], [31],[34] , [36]
I03	Recall	[2], [21], [19], [8], [28], [11], [18], [29], [10], [16],[20], [25], [34]
I04	Kappa Statistician	[26]
I05	Percentage of correct ratings	[21], [26], [5], [14], [36]
I06	Incorrect ranking percentage	[21], [5], [14], [36]
I07	Mean Square Error (MSE)	[13], [17], [26], [12]
I08	correlation coefficient	[17], [39]
I09	Coefficient ratio	[12]
I10	Mean Absolute Error (MAE)	[13], [8], [26], [12], [39]

I11	Relative absolute error (RAE)	[12]
I12	Relative Square Error (CSR)	[12]
I13	True positive rate or sensitivity (TPR)	[17], [23]
I14	True negative index or specificity (TNR)	[28], [17], [23]
I15	Precisión equilibrada	[23]
I16	Root Mean Squared Error (RMSE)	[39]
I17	F-Measure	[2], [21], [8], [11], [20] , [25]

Discussion

Of the 40 studies studied about the prediction of the academic performance of university students applying data mining techniques, 26 authors use the KDD methodology, this highlights the importance of this methodology, being the most used when predicting academic performance. Likewise, with respect to the attributes were considered for the prediction of academic performance: Gender, CGPA and work of the father / mother are the most used factors and correspond to 20 to the personal factor, 21 to the academic factor and 5 to the socio-economic factor respectively. Although the total of the factors are multiple, their characteristics change from one to another. Thus, in relation to variable selection algorithms, 3 authors use both correlationAttributeEval and InfoGain-AttributeEval, since these models can offer better results that approach reality. On the other hand, with respect to the techniques used in the prediction and their accuracy, there are many algorithms for the achievement of each task, being that 17 use Naives Bayes and 14 used J48, showing that these are the most efficient when predicting academic performance. As for accuracy, 16 articles talk about Naives bayes, this being the most used for the ability to record large groups of data, with 94% accuracy. As well, the most used tool for the development and testing of the model is Weka, which was used by 22 of the researchers considered for the study, this can be justified due to the frequency and the greater reception it has when developing data mining. Finally, 27 researchers use the accuracy metric and 13 Recall, since they yielded better results when predicting performance and are the most effective when predicting.

5. Conclusion

The study carried out in this research allows us to conclude that the prediction of performance can be useful, mainly for students, but it can also be useful for teachers and the university itself, since they could improve their teaching methods and the students, their learning method.

It can also be seen that 21 of the researchers consider the CGPA (grade point average) as the key variable for prediction due to its direct relationship with academic performance, in addition to the fact that in 16 articles the Naive Bayes technique is used and 22 researchers use the Weka tool to develop, due to its frequency and interactivity.

Thus, with this study it can be shown that the prediction of academic performance is an issue of concern to educational entities, since researchers aim to cooperate to improve student learning.

There is therefore an important point in all areas orienting it to the elaboration of proposed models to improve prediction, and that researchers have a basis for research on this topic, thus allowing to have an overview of prediction of academic performance with data mining.

References

[1]. Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5), 212-220. doi:10.14569/IJACSA.2016.070531

[2]. Aggarwal, D., Mittal, S., & Bali, V. (2019). Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques. *International Journal of Recent Technology and Engineering*, *8*, 497-503. doi:10.35940/ijrte.B1093.0782S719

[3]. Al Breiki, B., Zaki, N., & Mohamed, E. A. (2019, November). Using educational data mining techniques to predict student performance. In *2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA)* (pp. 1-5). IEEE. doi:10.1109/ICECTA48151.2019.8959676

[4]. Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. *International journal of information and education technology*, *6*(7), 528. doi:10.7763/IJIET.2016.V6.745

[5]. Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science*, *9*(8), 9. doi:10.5815/IJMECS.2017.08.02

[6]. Alsalman, Y. S., Halemah, N. K. A., AlNagi, E. S., & Salameh, W. (2019, June). Using decision tree and artificial neural network to predict students academic performance. In *2019 10th International Conference on Information and Communication Systems (ICICS)* (pp. 104-109). IEEE. doi:10.1109/IACS.2019.8809106

[7]. Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting critical courses affecting students performance: a case study. *Procedia Computer Science*, *82*, 65-71. doi:10.1016/j.procs.2016.04.010

[8]. Amornsinlaphachai, P. (2016, February). Efficiency of data mining models to predict academic performance and a cooperative learning model. In *2016 8th International Conference on Knowledge and Smart Technology (KST)* (pp. 66-71). IEEE. doi:10.1109/KST.2016.7440483

[9]. Anzer, A., Tabaza, H. A., & Ali, J. (2018, June). Predicting academic performance of students in uae using data mining techniques. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)* (pp. 179-183). IEEE. doi:10.1109/ICACCE.2018.8458053

[10]. Azizah, E. N., Pujianto, U., & Nugraha, E. (2018, October). Comparative performance between C4. 5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment. In *2018 4th International Conference on Education and Technology (ICET)* (pp. 18-22). IEEE. doi:10.1109/ICEAT.2018.8693928

[11]. Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020, February). Predicting students' academic performance through supervised machine learning. In *2020 International Conference on Information Science and Communication Technology (ICISCT)* (pp. 1-6). IEEE. doi:10.1109/ICISCT49550.2020.9080033

[12]. Canagareddy, D., Subarayadu, K., & Hurbungs, V. (2018, November). A machine learning model to predict the performance of university students. In *International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering* (pp. 313-322). Springer, Cham.

[13]. Chauhan, N., Shah, K., Karn, D., & Dalal, J. (2019, April). Prediction of student's performance using machine learning. In *2nd International Conference on Advances in Science & Technology (ICAST)*.

[14]. Chiheb, F., Boumahdi, F., Bouarfa, H., & Boukraa, D. (2017, December). Predicting students performance using decision trees: Case of an Algerian University. In *2017 International Conference on Mathematics and Information Technology (ICMIT)* (pp. 113-121). IEEE.

[15]. Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in human behavior*, *73*, 247-256. doi:10.1016/j.chb.2017.01.047

[16]. Gunawan, Hanes, & Catherine. (2019). Information Systems Students' Study Performance Prediction Using Data Mining Approach. *2019 Fourth International Conference on Informatics and Computing (ICIC)*. doi:10.1109/ICIC47613.2019.8985718

[17]. Halde, R. R., Deshpande, A., & Mahajan, A. (2016, May). Psychology assisted prediction of academic performance using machine learning. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 431-435). IEEE. doi:10.1109/RTEICT.2016.7807857

[18]. Hamoud, A., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*, 26-31. doi:10.9781/ijimai.2018.02.004

[19]. Hamoud, A., Humadi, A., Awadh, W. A., & Hashim, A. S. (2017). Students' success prediction based on Bayes algorithms. *International Journal of Computer Applications*, *178*(7), 6-12. doi:10.5120/ijca2017915506

[20]. Ketui, N., Wisomka, W., & Homjun, K. (2019). Using classification data mining techniques for students performance prediction. In *2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)* (pp. 359-363). IEEE. doi:10.1109/ECTI-NCON.2019.8692227

[21]. Khan, I., Al Sadiri, A., Ahmad, A. R., & Jabeur, N. (2019, January). Tracking student performance in introductory programming by means of machine learning. In *2019 4th mec international conference on big data and smart city (icbdsc)* (pp. 1-6). IEEE. doi:10.1109/ICBDSC.2019.8645608

[22]. Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1-26.

[23]. Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de tecnologias del Aprendizaje*, *10*(3), 119-125.

[24]. Ma, X., & Zhou, Z. (2018, January). Student pass rates prediction using optimized support vector machine and decision tree. In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)* (pp. 209-215). IEEE. doi:10.1109/CCWC.2018.8301756

[25]. Mahboob, T., Irfan, S., & Karamat, A. (2016, December). A machine learning approach for student assessment in E-learning using Quinlan's C4. 5, Naive Bayes and Random Forest algorithms. In *2016 19th International Multi-Topic Conference (INMIC)* (pp. 1-8). IEEE. doi:10.1109/INMIC.2016.7840094

[26]. Rubiano, S. M. M., & Garcia, J. A. D. (2015, October). Formulation of a predictive model for academic performance based on students' academic and demographic data. In *2015 IEEE Frontiers in Education Conference (FIE)* (pp. 1-7). IEEE. doi:10.1109/FIE.2015.7344047

[27]. Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. In *XIII Workshop de Investigadores en Ciencias de la Computación*.

[28]. Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science*, *8*(11), 36. doi:10.5815/ijmecs.2016.11.05

[29]. Pang, Y., Judd, N., O'Brien, J., & Ben-Avie, M. (2017, October). Predicting students' graduation outcomes through support vector machines. In *2017 IEEE Frontiers in Education Conference (FIE)* (pp. 1-8). IEEE. doi:10.1109/FIE.2017.8190666

[30]. Puarungroj, W., Boonsirisumpun, N., Pongpatrakant, P., & Phromkhot, S. (2018, January). Application of data mining techniques for predicting student success in English exit exam. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication* (pp. 1-6). doi:10.1145/3164541.3164638

[31]. Punlumjeak, W., & Rachburee, N. (2015, October). A comparative study of feature selection techniques for classify student performance. In *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 425-429). IEEE. doi:10.1109/ICITEED.2015.7408984

[32]. Rawat, K. S., & Malhan, I. V. (2019). A hybrid classification method based on machine learning classifiers to predict performance in educational data mining. In *Proceedings of 2nd International Conference on Communication, Computing and Networking* (pp. 677-684). Springer, Singapore. doi:10.1007/978-981-13-1217-5_67

[33]. Rimadana, M. R., Kusumawardani, S. S., Santosa, P. I., & Erwianda, M. S. F. (2019, December). Predicting student academic performance using machine learning and time management skill data. In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 511-515). IEEE.#

[34]. Samuel, Y. T., Hutapea, J. J., & Jonathan, B. (2019, July). Predicting the Timeliness of Student Graduation Using Decision Tree C4. 5 Algorithm in Universitas Advent Indonesia. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)* (pp. 276-280). IEEE. doi:10.1109/ICTS.2019.8850948

[35]. Segura-Morales, M., & Loza-Aguirre, E. (2017, December). Using decision trees for predicting academic performance based on socio-economic factors. In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1132-1136). IEEE.

[36]. Sivasakthi, M. (2017, November). Classification and prediction based data mining algorithms to predict students' introductory programming performance. In *2017 International Conference on Inventive Computing and Informatics (ICICI)* (pp. 346-350). IEEE.

[37]. Toppireddy, H. K. R., Saini, B., & Hada, P. S. (2019, February). Academic Enhancement System using EDM Approach. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)* (pp. 1-10). IEEE. doi:10.1109/ICACCP.2019.8882954

[38]. Wang, G. H., Zhang, J., & Fu, G. S. (2018, December). Predicting student behaviors and performance in online learning using decision tree. In *2018 Seventh International Conference of Educational Innovation through Technology (EITT)* (pp. 214-219). IEEE.

[39]. Widyahastuti, F., & Tjhin, V. U. (2017, July). Predicting students performance in final examination using linear regression and multilayer perceptron. In *2017 10th International Conference on Human System Interactions (HSI)* (pp. 188-192). IEEE. doi:10.1109/HSI.2017.8005026

[40]. Widyaningsih, Y., Fitriani, N., & Sarwinda, D. (2019, July). A Semi-Supervised Learning Approach for Predicting Student's Performance: First-Year Students Case Study. In *2019 12th International Conference on Information & Communication Technology and System (ICTS)* (pp. 291-295). IEEE. doi:10.1109/ICTS.2019.8850950