International Journal of Advanced Multidisciplinary Research and Studies

Prediction and Evaluation of Students' Performance in E-Learning Using Data Mining Algorithm

[1] Noor Sami Razzaq Najjar, [2] Karrar Al-Jammali

[1] Department of Computer Science, The General Directorate of Education, the Ministry of Education, Najaf, Iraq
[2] Faculty of Pharmacy, University of Kufa, Najaf, Iraq

Corresponding Author: **Noor Sami Razzaq Najjar**

Abstract

Purpose: The proposed work aims to analyze the used and most efficient machine learning techniques in both online and offline education contexts, for different objectives, this research used various machine learning approaches to predict student performance. The educational data is analyzed based on performance among all compared algorithms on the dataset.

Methods: We used the algorithms k-NN, SVM, LR and ANN for student performance analysis depending on various input features, students from secondary schools and an intermediate school in Iraq were used to compile the performance dataset for the students, the dataset includes attributes suggested about the students, our data set includes 800 instances, with 8 attributes. We used a variety of classification algorithms on datasets of student performance to analyses the dataset and improve the generic reliability of the algorithms to find classifiers with higher accuracy. So, we calculate the algorithm's accuracy, that the accuracy of all implemented classification methods is training and testing calculated by using split-validation. The student's performance dataset was used to train and after that test a variety of classification methods.

Results: ANN classification algorithm's results are very encouraging when compared to other methods for classification. Neural Networks Classification Algorithm,

when we experimented with the whole 8 attributes, this algorithm showed 100% as the highest accuracy, and when implementation The logistic regression Classification Algorithm, when we experimented with the whole 8 attributes, this algorithm showed 99.71% as the highest accuracy and when implementing the K Nearest Neighbor KNN Classification Algorithm, when experimented with the whole 8 attributes, this algorithm showed 99.43%, but when we implementation SVM Classification Algorithm, when we experimented with the whole 8 attributes, this algorithm showed 99.17% as the Lower Accuracy compares with other algorithms.

Conclusions: Using E-learning classification algorithms for a student performance dataset using data mining methods, data mining algorithms have been applied to the student performance data set in E-learning.

ANN classification algorithm's results are very encouraging when compared to other methods for classification, when we apply the training data ANN classifier the result of accuracy is higher, While in SVM little lasting accuracy.

We can conclude that ANN is a more efficient model than others for classifying student performance across these multiple attributes.

The best important factors that affect student effort have been made to improve the quality of E-learning.

1. Introduction

The development of information technology across a number of industries has resulted in the storage of huge amounts of data in a variety of formats, including records, files, papers, videos, photos, sound, and scientific data.

In order to improve decision-making, it is often necessary to properly extract information from huge data sets using the data collected from various applications [1].

The education system has changed from a traditional offline mode to an online mode, known as E-Learning (EL), as a result of extensive usage of Internet technology, this has provided scholars with a new topic for research [1].

The knowledge and performance of students were typically improved by studying in face-to-face settings. However, during the COVID-19 pandemic, all educational facilities were shut down and moved to an online environment, which increased the significance of the e-learning environment [2].

The aim of E-Learning is an actual evaluation of a student's performance and the prediction of student performance will be

useful to teachers/course coordinators in the early stages of the course in determining who requires attention and help [3]. The main aim of my research is to use data mining technology and methodologies to study student's performance in E-Learning.

Data mining provides many methods for studying student performance, as well as numerous methodologies for data classification [4].

2. Data Mining Definition
The term "data mining" describes the process of obtaining information from huge datasets. Data mining techniques are applied to these large amounts of data to find hidden links and patterns that may be used to inform decision-making.

Data mining is part of the knowledge discovery process, it is a sequence of steps identified in extracting knowledge from data [5]. That is shown in Fig 1.
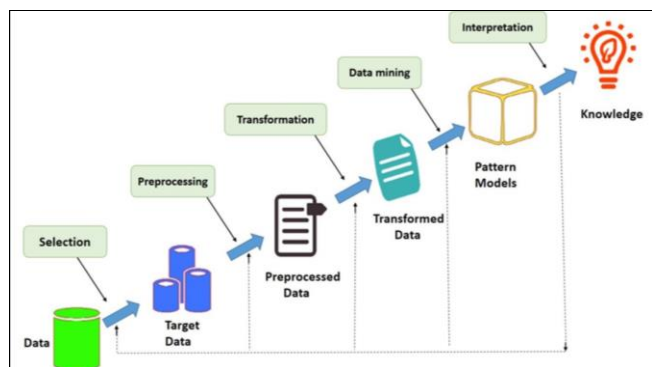

Fig 1: The extracting knowledge from data

3. Classification Techniques
Some of the intelligence techniques used in learning analytics to predict student success include supervised learning, data mining, unsupervised learning, and statistical approaches. The algorithms and techniques are various like Clustering, Classification, Regression, Artificial Intelligence, Neural Networks, Genetic Algorithms, the Nearest Neighbor method, Association Rules, Decision Trees, etc.

They are used to extract knowledge from databases. These data mining methodologies and strategies must be briefly discussed in order to be understood [5]. To anticipate the student's performance, we chose four data mining techniques:

K Nearest Neighbor (K-NN), logistic regression (LR), Support Vector Machine (SVM) and Artificial neural networks (ANN).

3.1 Artificial Neural Network (ANN):
This algorithm works simulates the behavior of the human brain. It consists of a group of neurons connected through connections. The effect of one node over another is determined by the weight allocated to each link. It processes input signals and sends them to other neurons.

ANN algorithm consists of input and output layer as well as a number of several, the neurons at each layer connected with another layer [6].

3.2 K Nearest Neighbor (K-NN) Classification:
KNN is a more widely used strategy that uses k nearest neighbors to identify the class. K-NN is a non-parametric machine learning technique, and because the best-suited

data distribution is unknown, KNN predicts straight from the training dataset. As a result, it is the most straightforward and appropriate technique for forecasting [7].

3.3 Support Vector Mechanism (SVM):
SVM is a powerful classification tool in machine learning. It is one of several ML approaches and, when compared to other ML methods, it is particularly effective at recognizing patterns in complicated datasets [8].

3.4 The Logistic Regression (LR):
The logistic regression equation is one of the modern non-parametric methods developed to predict qualitative variables (nominal or ordinal). It is considered an alternate test for linear regression equations, both simple and numerous [9].

This equation depends on the concept of the model in terms of the possibility of testing the set overall layout effect of the variables on one qualitative, the dependent variable, and its use of the goodness of fit modelling concepts [10].

In probability science logistic regression is a model used to predict the probability of an event occurring with its fit, for data on a logistic, logistic regression uses several (independent) predictor variables which can be relative, categorical, nominal, or ordinal versus a single function predictor variable [11].

4. Methodology
This section provides details on the data set and the methods used for performing experimental analysis and the results of implementation.

We used the algorithms k-NN, SVM, LR and ANN for student performance analysis depending on various input features.

Students from secondary schools and an intermediate school in Iraq were used to compile the performance dataset for the students.

The dataset includes attributes suggested about the students; our data set includes 800 instances, with 8 attributes. Table 1 below shows the details for this dataset.

Table 1: Attributes of Dataset

A	Age
B	sex (M=0, F=1)
C	School of student (Middle=0, secondary=1)
D	Number of hour study (0,1,2)
E	Current health status(0=healthy,1=sick)
F	Internet at home(0=yes,1=no)
Class	Grade(0=Pass,1=Fail)

Table 2: Sample of Dataset

A	B	C	D	E	F	G	CLASS
1	1	2	2	0	0	0	pass
1	2	2	0	0	0	0	pass
1	2	1	0	0	0	0	pass
0	2	0	0	0	0	0	pass
1	1	1	0	0	0	0	pass
0	1	2	0	0	0	0	pass
0	2	1	1	0	0	0	pass
0	2	1	2	0	0	0	pass
1	1	1	0	0	0	0	pass
1	1	0	0	0	0	0	pass
0	2	1	0	0	0	0	pass
1	2	0	0	0	0	0	pass
0	1	2	0	0	0	0	pass

1	1	0	0	0	0	0	pass
1	2	1	0	0	0	0	pass
1	1	2	0	0	0	0	pass
1	1	1	2	1	1	1	fail

5. The Results and Discussion
5.1 Applying Classification Algorithms
We used the Rapid Miner Framework for doing the processes, which is free to use and comes with split - Validation for training and testing data for a number of classification algorithms, after that, we used the pre-process function to import the dataset and apply classification algorithms., Fig 2 Shows Framework to the steps for E-Learning Student Academic Performance Prediction.
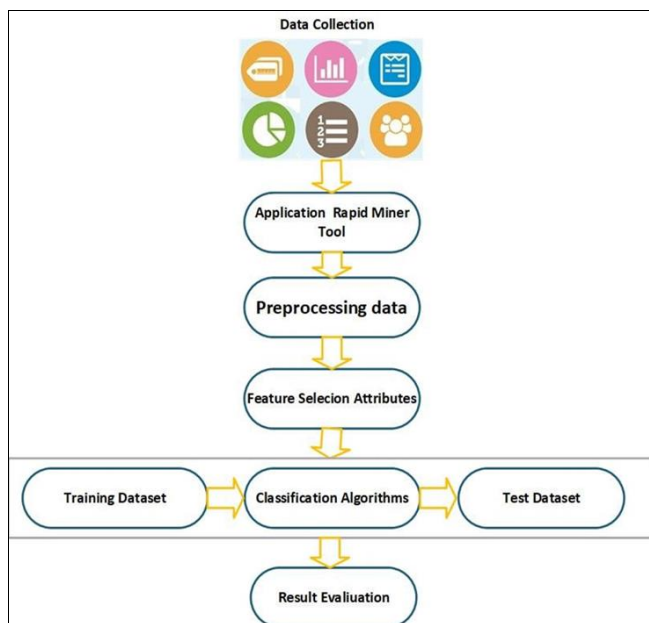

Fig 2: Framework for E-Learning Student Academic Performance Prediction

5.2 Classification Technique Performance
We used a variety of classification algorithms on datasets of student performance to analyses the dataset and improve the generic reliability of the algorithms to find classifiers with higher accuracy. So, we calculate the algorithm's accuracy, that the accuracy of all implemented classification methods is training and testing calculated by using split-validation. The student's performance dataset was used to train and after that test a variety of classification methods. The many classification algorithms that are used like (KNN, SVM, LR and ANN algorithm).

The accuracy of classification algorithms with parameter is as shown in table 4. In the implementation Neural Networks Classification Algorithm, when we experimented with the whole 8 attributes, this algorithm showed 100% as the highest accuracy, and when implementation the logistic regression Classification Algorithm, when we experimented with the whole 8 attributes, this algorithm showed 99.71% as the highest accuracy and when implementing the K Nearest Neighbor KNN Classification Algorithm, when experimented with the whole 8 attributes, this algorithm showed 99.43%, but when we implementation SVM Classification Algorithm, when we experimented with the whole 8 attributes, this algorithm showed 99.17% as the Lower Accuracy compares with other algorithms. Table 3 shows the accuracy of the prediction for student performance using a classification algorithm.

Table 3: Attributes of Dataset

Model	Accuracy result
ANN	100%
SVM	99.17%
KNN	99.43%
LR	99.71%

5.3 Analyzing the Results
ANN classification algorithm's results are very encouraging when compared to other methods for classification. There is a small difference in accuracy, when we apply the training data ANN classifier the result of accuracy is higher, While in SVM little lasting accuracy.

The following table 3 shows that ANN is higher accurate than SVM and LR and KNN classifiers the result between SVM and LR in accuracy reflex that the two model-based the technique in training. The following table 3 shows that ANN is more accurate than the SVM and LR and KNN classifiers, this happens because of the instances used to train the ANN model and emphasizes that the ANN model should use with a large dataset in deep learning to produce the most efficient results.

On the other hand, the ANN classifier produces the best results depending on the experiment in Table 3. By comparing the two tables, we can conclude that ANN is a more efficient model than others for classifying student performance across these multiple attributes.

Fig 3: Shows the accuracy of the prediction of student performance using a classification technique.
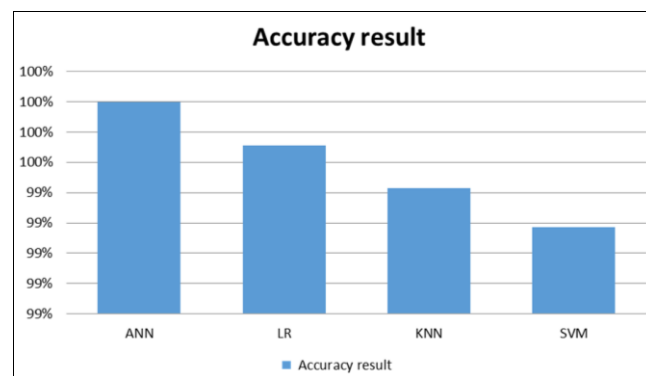

Fig 3: Accuracy of the prediction of student performance

6. Conclusion
As information technology has improved, educational institutions have created courses on using E-learning and helping students to obtain the best education possible.

In my research, we built E-learning classification algorithms for a student performance dataset using data mining methods, the suggested model has 800 instances. Data mining algorithms have been applied to the student performance data set in E-learning.

We use questionnaires to collect data and we used the Rapid Miner for the analysis of this dataset.

ANN classification algorithm's results are very encouraging when compared to other methods for classification. there is a small difference in accuracy, when we apply the training data ANN classifier the result of accuracy is higher, While in SVM little lasting accuracy.

We can conclude that ANN is a more efficient model than others for classifying student performance across these multiple attributes.

The best important factors that affect student effort have been made to improve the quality of E-learning.

E-learning must be combined with traditional education, as a result of which professional skills will be developed and training quality will be improved. This should increase the level of knowledge of students in this sector, based on their desires to provide a suitable environment for the development and expansion of e-learning, designing training programs, such as workshops, and the capacity to use e-learning as an effective training tool to increasing quality and providing services.

COI statement:
Conflict of Interest: We declare that we have no conflict of interest.
Ethical Standards: This type of study formal consent is not required.
Informed Consent: Informed consent was obtained from all individual participants included in the study.
Author Contributions: All authors contributed to the research and project.
Funding: There is no Funding (Self-financing).

7. References

1. Mannila, Heikki. Data mining: Machine learning, statistics, and databases. In Proceedings of 8th International Conference on Scientific and Statistical Data Base Management, IEEE, 1996, 1-9. Doi: https://doi.org/10.1109/ssdm.1996.505910
2. Hellas, Arto, Petri Ihantola, Andrew Petersen, Vangel V, Ajanovski, *et al*. Predicting academic performance: A systematic literature review. In Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education, 2018, 173-198. Doi: https://doi.org/10.1145/3293881.3295783
3. Noble, William Stafford. 3 Support Vector Machine Machine Applications in Computational Biology. Kernel methods in computational biology. 2004; 71. Doi: https://doi.org/10.7551/mitpress/4057.003.0005
4. Brijesh Kumar, Saurabh, Sourabh P. Mining educational data to analyze students performance. IJACSA 2. 2011; 6.
5. Coomans, Danny, Désiré Luc Massart. Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. Analytica Chimica Acta. 1982; 136:14-27. Doi: https://doi.org/10.1016/s0003-2670(01)95359-0
6. Al-Jammali, Karrar. Prediction of Heart Diseases Using Data Mining Algorithms. Informatica. 2023; 47(5). Doi: https://doi.org/10.31449/inf.v47i5.4467
7. Najjar, Noor. Analyzing Data Mining Statistical Models of Bio Medical. Analyzing Data Mining Statistical Models of Bio Medical (elte.hu), 2018.
8. Abraham, Ajith. Artificial neural networks. Handbook of measuring system design, 2005.
9. Mehrotra, Kishan, Chilukuri K, Mohan, Sanjay Ranka. Elements of artificial neural networks. MIT press, 1997.
10. Park, Hyeoun-Ae. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing. 2013; 43(2):154-164. Doi: https://doi.org/10.4040/jkan.2013.43.2.154
11. Mthimunye, Katlego, Daniels FM. Predictors of academic performance, success and retention amongst undergraduate nursing students: A systematic review. South African Journal of Higher Education. 2019; 33(1):201-220. Doi: https://doi.org/10.20853/33-1-2631