Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/121228

Modeling and Classification of Student Performance Based on a Machine Learning Model

Abdulellah A. Alsulaimani

Department of Educational Technologies, Faculty of Education, King Abdulaziz University, Jeddah, Saudi Arabia

Received: 10 Sep. 2023, Revised: 25 Oct. 2023, Accepted: 28 Nov. 2023
Published online: 1 Dec. 2023.

Abstract: The increasing popularity of Machine Learning in the education business can be attributed to its capacity to enhance several aspects of the educational system. The objective of the present study is to construct a prediction model utilizing Machine Learning techniques in order to forecast students' academic performance. In the contemporary competitive landscape, academic institutions are compelled to engage in the proactive task of predicting students' academic performance, categorizing them based on their individual talents, and implementing strategies to enhance their success in examinations. In order to identify students who may require early intervention and support, educational institutions must have the capacity to analyze student learning behavior through the application of predictive models for student achievement. The present study utilized a sample of 1087 students enrolled at King Abdulaziz University in Saudi Arabia to make predictions about student scores by employing the GMM model. The results indicated that the mean score achieved by students enrolled in this particular course varied between 14 and 93. The findings also indicate that the optimal model for predicting students' academic achievement is the mixture model with four components and varying variances.

Keywords: Machine Learning; Modeling; classification; Gaussian mixtures model, prediction.

1 Introduction

The study examines age, sex, obesity, average family income, family size, father and mother education, marital status, and school characteristics including gender, academic level, and others [1] stress, lifestyle, and academic performance. Random forests, gradient boosting, stacking, and artificial neural networks were used. Gradient boosting, randomization, ANN, and logistic regression followed stacking as the best algorithm. Lifestyle greatly impacts academic success. This study predicts undergraduate Chinese university students' GPAs using socioeconomic background and admission exam results using ANN [2]. In the first stage, statistical assessments of students' background information showed that GPAs improved annually, female students had better grades than male students, rural and urban students performed similarly, and non-repeating students performed substantially better than repeat students. GPA matches the admission exam English test best. Mothers affect academic success more than fathers. Predicting final test scores for undergraduates based on their midterm results is the focus of this study [3], which uses machine learning techniques. Various methods for predicting students' final test scores, including Random Forests, Nearest Neighbors, Support Vector Machines, Logistic Regression, Naive Bayes, and Nearest Neighbors, were evaluated. Looking into two parts. Grades in Achievement are a Measure of Academic Success. Quantitatively evaluate various machine learning approaches. 75%-90% precision. The study [4] looks for a prediction tool that can assist at-risk pupils to succeed by identifying them early. This forecasting approach might increase student achievement and decrease dropout rates. To reduce failure and improve performance, it can assist in identifying students who require additional support. We put decision trees, logistic regression, support vector machines, random forest classifiers, and K-nearest neighbors to the test. 91.7% accuracy and an R2 of 0.977 in regression gave the Random Forest Classifier the highest overall classification score, according to the experiments. Eight models for predicting. In this research [5] decision tree, naive Bayesian, SVM, and neural network approaches were all used in this study. SVMs fared better than other machine learning techniques. The support vector machine technique can be used to develop a prediction model to foretell learners' performance after achieving 78.75% correct categorization. The program forecasts student achievement and recognizes pupils who are at danger. In this paper [6] four categorization models were created to predict the performance of Al-Muthanna University, College Of Humanities computer science students. Fully linked feed-forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression were employed. ROC index performance and classification accuracy were used to compare models. ANN model gets the highest ROC index (0.807) and accuracy (77.04). The decision tree model shows that not all qualities are classified. Computer Grades-Course1, Accommodation, Computer Study Interest, Educational Environment Decision tree models use satisfaction and residency. In this paper [7] research seeks to accurately predict student performance to improve academic outcomes.

*Corresponding author e-mail: aalsulaimani@kau.edu.sa

Educational data mining can help give high-quality education. Accurate student learning estimation is one strategy to improve higher education quality. Many mining-based prediction methods exist. Existing policies have struggled to meet the education framework's higher and master training requirements. This paper reviews current models and proposes a new model to predict student achievement. This research addresses the problems and opportunities of quality education in higher education institutions and proposes a methodology for increasing education quality. In [8] this study looks at at-risk identification and academic achievement prediction using deep learning. This study uses deep neural networks (DNN), decision trees, random forests, gradient boosting, logistic regression, support vector classifiers, and K-nearest neighbors to forecast students' future academic success based on their first-year grades. Paper [9]. In order to forecast the outcomes of pupils, the study analyzed various classification techniques, including the Artificial Immune Recognition System v2.0 and AdaBoost. The study's highest categorization accuracy, 95.34%, was generated using deep learning techniques. A statistical choice was made to choose the optimum classification techniques by calculating the Precision, Recall, F-Score, Accuracy, and Kappa Statistics Performance. The 10140 student records in the dataset were used in this investigation. [10] predicts course performance. Data mining uncovers patterns in large volumes of data. These patterns may aid analysis and prediction. Education data mining includes data mining applications in education. These apps analyze student-teacher data. Analyses can classify or predict. Nave Bayes, ID3, C4.5, and SVM are studied. Experiment uses UCI machinery student performance data. Algorithm accuracy and error rate are assessed. This research [11] suggests utilizing a statistic that has never been used in the K-means technique to optimize the clustering-number determination. The K-means algorithm's grouping effect is then evaluated by discriminant analysis. This [12] essay tries to discuss current developments in estimating students' academic performance. We outline the metrics for measuring educational achievement and point out the advantages and disadvantages of the most popular data processing (DM) tools and techniques now in use. Additionally, we provide a current evaluation of the EDM research that has been published in recent years with a focus on predicting academic achievement in educational activity. .In this study [13] the Naïve , Bayes, K Star, IBK, and Nearest Neighbor (KNN) classifiers that can run incrementally have been contrasted. Applying the closest neighbor algorithm to the utilized Student Evaluation data-set shows that it performs more accurately than other algorithms. In this study [14] a group of students majoring in computer science at several undergraduate colleges in Kolkata are first given a set of characteristics. Since there are a lot of qualities, feature selection algorithms are used on the data set to get rid of some of them. Then, five classes of Machine Learning Algorithm (MLA) are applied to this data set, and it is discovered that the decision tree class of algorithms produced the best results.

2 Gaussian mixture model

Without specifying a data set for the sub-population that contains a single observation, a mixture model, which is probabilistic, simulates the existence of sub-populations within the overall population [15]. The mixture distribution, which describes the probability distribution of observations throughout the entire population, is met by this model [16]. The issue of determining the characteristics of the total population from the characteristics of the sub-populations is connected to "mixture distributions" problems. However, "mixture models" are used to provide observations on the pooled population without knowing the identification of the sub-populations and to draw statistical conclusions about their characteristics. [17]. Some mixture model execution techniques include stages that draw attention to (or give more weight to) hypothesized sub-population identities for particular observations [18]. They are regarded as types of unsupervised learning or clustering techniques in this situation [19].

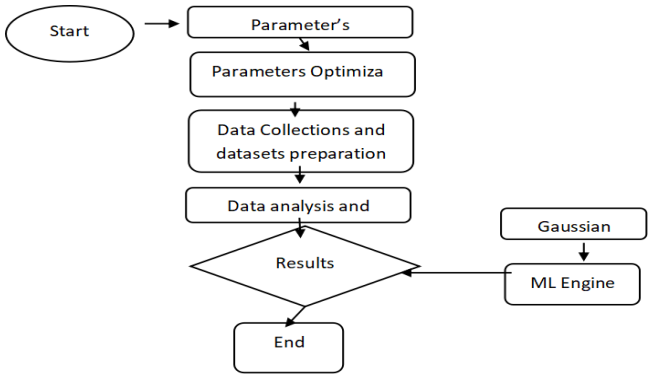

Fig. 1: Flowchart of the model algorithm which started with parameter's identifications and ended with achieved results.

Model Selection

The issue of selecting a single model from a group of potential models is known as model selection It is usual practice to select a model based on its performance on a hold-out test data-set or to calculate model performance by resampling [20].

It is common practice to utilize AIC and BIC as model selection criterion [21] AIC and BIC stand for Akaike's Information Criteria and Bayesian Information Criteria, respectively. The AIC and BIC are calculated using the equations below [22]:

$$AIC = -2Ln(L) + 2K \quad (1)$$

$$BIC = -2Ln(L) + 2Ln(LK) \quad = \quad (2)$$

where k is the number of estimated parameters, N is the number of recorded measurements, and L is the likelihood value. An improved fit is indicated by a lower AIC or BIC value [23].

3 Numerical results

The performance of the students that available historical in King Abdul Aziz University.

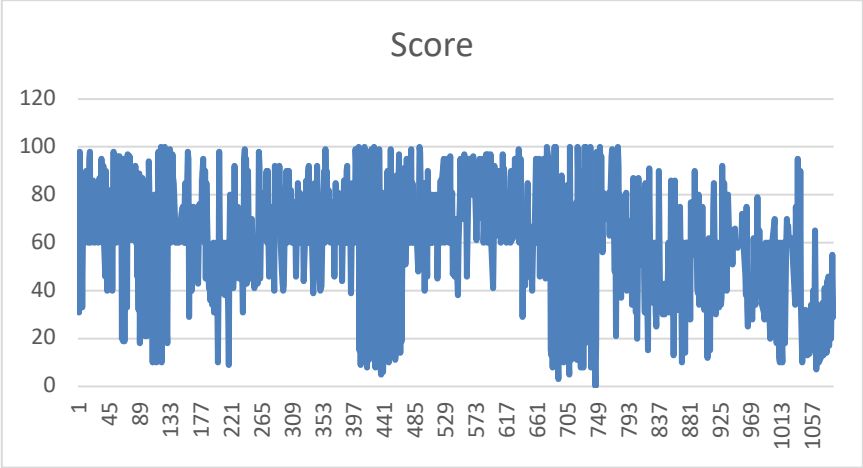

Fig. 2: Shows the distribution for the score of students

Table 1: The descriptive statistics for the score of students

Var.	Mini.	Maxi.	S.D	Mean
The Score	0.000	100.000	23.680	60.684

Table 2: Evolution of the AIC for each mode

Model/Number of classes	**2**	**3**	**4**	**5**
V	-9860.178081	-9794.981	-9729.850	
E	-9971.720	-9897.766731	-9860.387	-9864.387

Table 3: Different four components and its proportions.

Class	1	2	3	4
Proportions	0.096	0.145	0.627	0.133

Table 4: The mean by the four components

Class	1	2	3	4
Mean	14.000	36.919	66.250	93.988

Table 5: The variance by the four components

Class	1	2	3	4
Variance	25.777	33.171	149.131	20.984

Table 6: The model selection under selected criterion. The smallest one criterion is NEC which indicate that there is a clustering structure in the data

BIC	AIC	ICL
-9784.753	-9729.850	-10214.587
Log-likelihood	NEC	Entropy
-4853.925	1.680	214.917

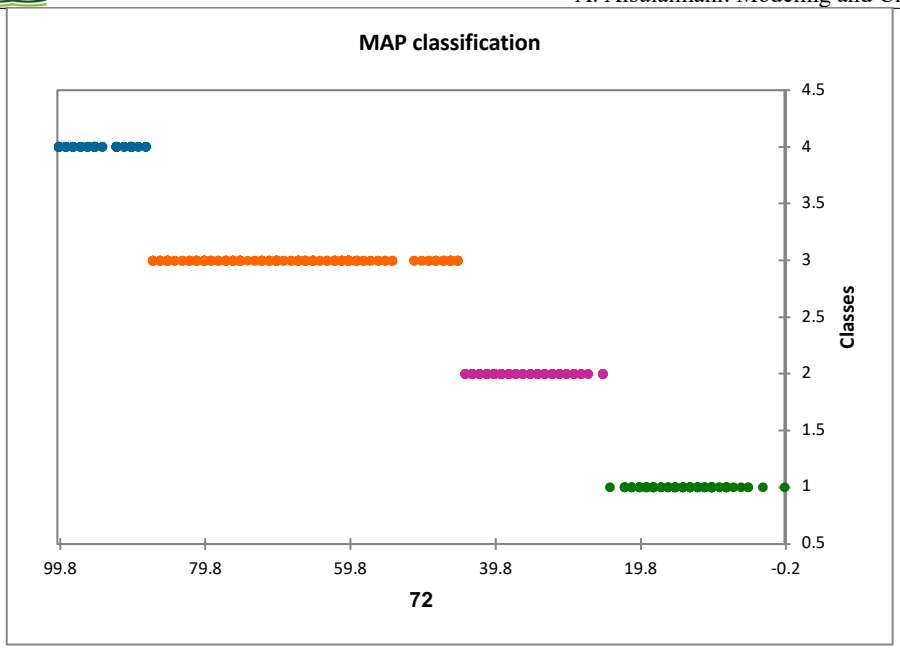

Fig. 3: No assignment exists in classes 4 by the MAP classification.

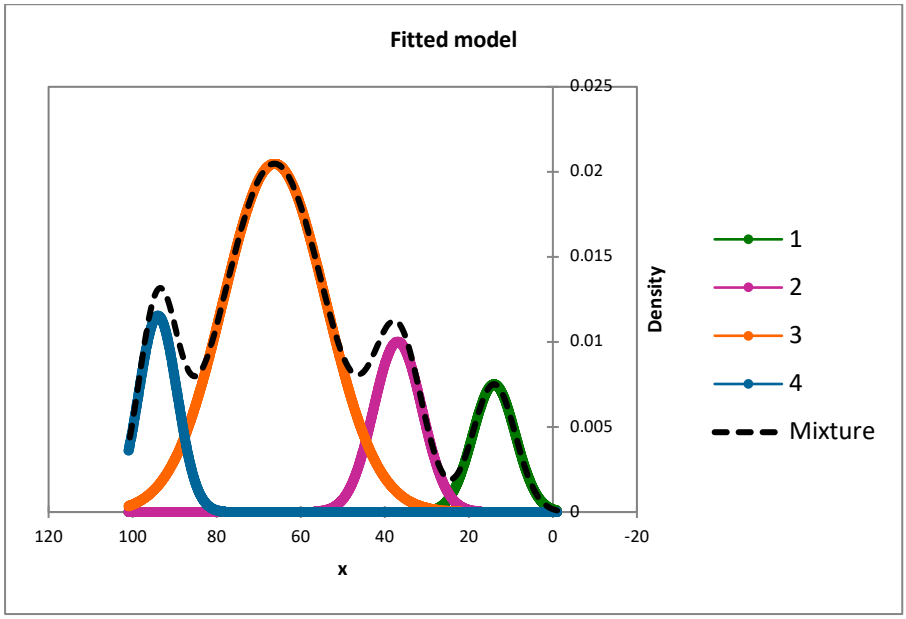

Fig. 4: Shows the mixture model

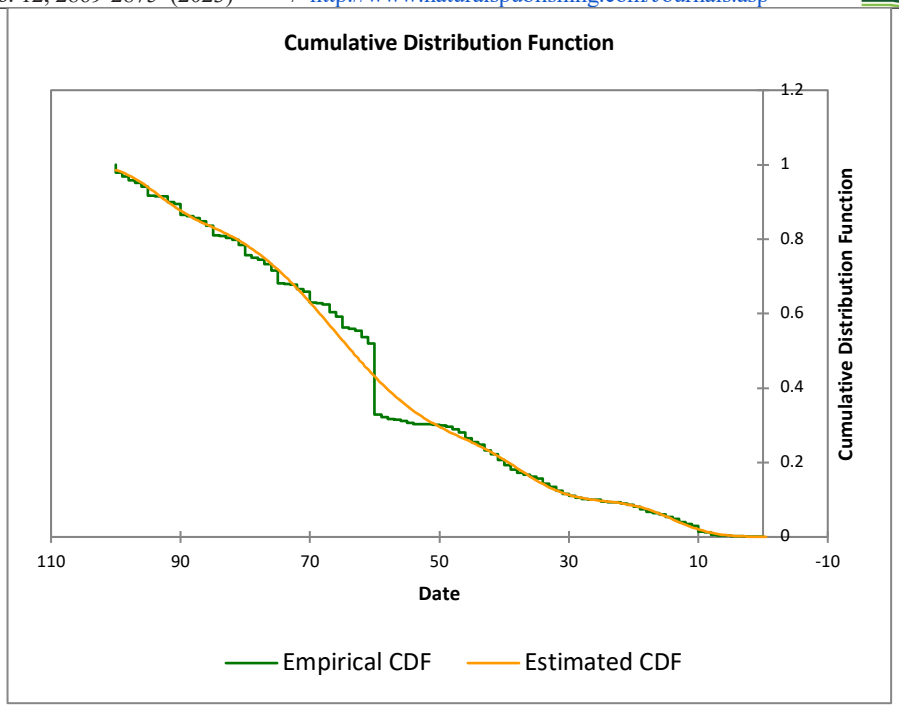

Fig. 5: The estimated CDF and empirical CDF are very close under the mixture model, which satisfies the accuracy of the estimation

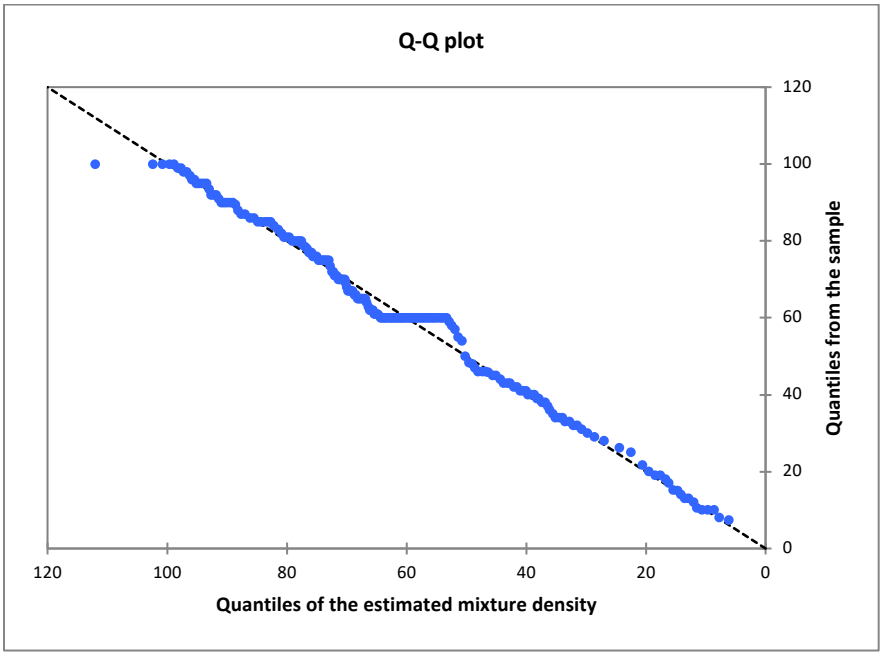

Fig. 6: Shows the quintiles of the estimated mixture density.

4 Conclusion

In this study, we classified student performance using the model of Gaussian mixtures and discussed the enormous challenge of forecasting student performance. The V (Variable variance) with 4 component(s) is the best mixture model, as determined by the AIC criterion. The EM algorithm failed to converge. The Bayesian information criteria recommend increasing the maximum number of iterations (see Table 2), and it shows that the mixture model with four components and different variance is the optimum model. The proportions of the four components ranged from 0.096 to 0.133 (see Table 3). Four components' means ranged from 14 to 93 (see Table 4). and the four components' variances ranged from 20 to 1 (see Table 5). Finally, it must be pointed out that the implementation of such a mechanism to predict the

performance of students is extremely useful.

Conflicts of Interest Statement

The authors declare no conflict of interest.

Reference

[1] Rajendran, S., Chamundeswari, S., & Sinha, A. A. (2022). Predicting the academic performance of middle-and high-school students using machine learning algorithms. Social Sciences & Humanities Open, 6(1), 100357 (2022).

[2] Lau, E. T., Sun, L., & Yang, Q. (2019). Modeling, prediction, and classification of student academic performance using artificial neural networks. SN Applied Sciences, 1, 1-10., (2019).

[3] A. Cannav`o, C.D., L. Morra, and F. Lamberti, (2019). Immersive virtual reality-based interfaces for character animation. IEEE Access., 2019. 7: p. 125463–125480 (2019).

[4] Costa-Mendes, R., et al.,(2021). A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach. Education and Information Technologies,26(2),1527-1547., (2021).

[5] Sarmonpal, S. (2018). Learning analytics from research to practice: a content analysis to assess information quality on product websites (Doctoral dissertation, Pepperdine University),. (2018).

[6] Buniyamin, N., U. bin Mat, and P.M. Arshad. (2015). Educational data mining for prediction and classification of engineering students' achievement. in 2015 IEEE 7th International Conference on Engineering Education (ICEED). (2015).

[7] Viswanathan, M.S., (2021). Study Of Students' Performance Prediction Models Using Machine Learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(2): p. 3085-3091. (2021).

[8] Nabil, A., M. Seyam, and A. Abou-Elfetouh, (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. IEEE Access, 9: p. 140731-140746. (2021).

[9] Hussain, S., et al., (2019). Prediction model on student performance based on internal assessment using deep learning. International Journal of Emerging Technologies in Learning, 14(8). (2019).

[10] Musso, M.F., C.F.R. Hernández, and E.C. (2020). Cascallar, Predicting key educational outcomes in academic trajectories: a machine-learning approach. Higher Education, 80: p. 875-894. (2020).

[11] Kardaş, K. and A. Güvenir, (2020). Analysis of the effects of Quizzes, homework, and projects on final exams with different machine learning techniques. EMO Journal of Scientific, 10(1): p. 22-29. (2020).

[12] Neha, K., A, (2021). study on the prediction of student academic performance based on expert systems. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(7): p. 1483-1488. (2021).

[13] Kulkarni, P. and R. Ade, (2014). Prediction of student's performance based on incremental learning. International Journal of Computer Applications, 99(14): p. 10-16. (2014).

[14] Acharya, A. and D. Sinha, (2014). Early prediction of students' performance using machine learning techniques. International Journal of Computer Applications, 107(1): p. 37-43. (2014).

[15] Saraiva, E.F., et al., (2019). An Integrated Approach for Making Inference on the Number of Clusters in a Mixture Model. Entropy, 21(11): p. 1063. (2019).

[16] Carlsson, K.C., et al., (2009). Modeling subpopulations with the $ MIXTURE subroutine in NONMEM: finding the individual probability of belonging to a subpopulation for the use in model analysis and improved decision making. The AAPS journal, 11: p. 148-154. (2009).

[17] Arshad, U., et al., (2019). Development of visual predictive checks accounting for multimodal parameter distributions in mixture models. Journal of Pharmacokinetics and Pharmacodynamics, 46: p. 241-250. (2019).

[18] Pallathadka, H., et al., (2023). Classification and prediction of student performance data using various machine learning algorithms. Materials today: proceedings, (2023). 80: p. 3782-3785.

[19] Gök, M.,(2017). Predicting academic achievement with machine learning methods. Gazi University Journal of Science Part c: Design and Technology, 5(3): p. 139-148.(2017).

[20] Maydeu-Olivares, A. and C. Garcia-Forero, (2010). Goodness-of-fit testing. International encyclopedia of education, (2010). 7(1): p. 190-196. (2010).

[21] Murphy, K.P., (2012) Machine learning: a probabilistic perspective. MIT press.(2012)

[22] Forster, M. and E. Sober, (2011).AIC scores as evidence: A Bayesian interpretation, in Philosophy of statistics., Elsevier. p. 535-549. (2011)

[23] Acquah, H.D.G.,(2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship.(2010).