Journal of Education for Pure Science- University of Thi-Qar
Vol.12, No2 (December, 2022)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.12.02.22

Enhanced Machine Learning model for Student Academic Performance Prediction using Genetic algorithm and compare with CNN model

Shaymaa Kaseb Layus

Department of Cmputer Science , College of Eduction for Pure Science, University of Thi-Qar

Received 14/8/2022 Accepted 26/9/2022 Published 01/12/2022

Abstract:

In this paper, a comparison between various machine learning algorithms for prediction of student academic performance has been achieved. Students' data is being acquired many resources such as e-learning resources, online and virtual courses and institutional technology. Many researches might use the collected data to find out and comprehend students' behaviors toward learning. The obtained data need to be preprocessed then used to build prediction model. The achieved models are Support Vector Machine (SVM), Logistic Regression (LR), Artificial Neural Network (ANN) with accuracies 99.99 %, 100 % and 99.99 % respectively. Genetic algorithm was used to adapt parameters of the algorithms to give best accuracy. OULAD dataset was used for validation. Results showed that these algorithms are sufficient for prediction of student academic performance and CNN is not needed for this purpose.

Keywords: Prediction, Machine Learning, CNN, SVM, LR, ANN, Genetic algorithm.

1- Introduction:

Education is an important issue in all societies; it plays an important role in Societies Evolution. With the wide evolution in everyday life techniques, distraction tools are increasing and this will affect education negatively. Machine learning (ML) techniques with their tools allow a high-level extraction of useful data from pure raw data, offer interesting possibilities for the education domain [1]. In General, estimating the academic performance of students has been an essential topic in various academic fields. Using the outputs of an estimation model, the teacher can take useful data to enhance student learning in general, especially for students who have low studying performance [2, 3].

Many studies have used ML techniques to enhance education quality and improve management of school resource. Hence, estimating student academic performance is challenging because there are diverse factors that might affect it (family, personal, friends, psychological, socio-economic and different environmental factors [4, 5]. Deep learning is a part of machine learning that is used basically with prediction and classification problems and with images data types basically.

The aim of this paper is to apply many ML methods to estimate student performance and analysis them and discuss how to implement deep learning in this type of problems, then comparing all methods to determine which method will give high prediction accuracy. OULAD dataset is used for evaluation the implemented methods:

- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)
- Logistic Regression (LR)
- Convolution Neural Network (CNN).

2- Related Works:

In [6], authors analyzed old students' data that contained their personal information such as family background, demographic allocation, age and how they valuate study. They used machine learning to validate this data. They used neural network, linear regression alone, deep learning with linear regression. Data was split into test and train. The accuracy was calculated depending on mean average error (MAE), authors did not optimize their method to get best parameters of proposed algorithms.

Authors in [7] developed a new technique to deal with prediction of student performance with a problem of type Multi-Input Multi-Output (MIMO). The developed method MIMO works to estimate the future achievement of any student. Their new method depends on Adaptive Neuro-Fuzzy Inference System (ANFIS); it is called MANFIS (Multi ANFIS); it uses multiple parameter groups with adaptive learning for these parameters. It searches for best value for the set of parameters, the results were with high accuracy of predication. They worked on strategy that uses two types of learning local and global. For global type, a random parameter group is trained sequentially from first record to the last one. Each train results in new values for the group of parameters. Global training will process all resulted groups by running training many times and the last try will result a meaningful value for this group of parameters. For local type, two algorithms were used Particle Swarm Optimization and gradient descent in a hybrid method. And the result is that each record will have its own group. The proposed is validated against using many datasets such as dataset from UCI (University of California Irvine) and different datasets that were used in KDD Cup. The results proved that the proposed method did better than other algorithms in term of accuracy.

In [8], authors made an effective survey to study more than 60 articles about student performance. They focused on three points, a) how to estimate learning outcomes, b) Analytics models that were achieved to predict performance of students and c) parameters that have a major effect on student achievements. They have reviewed these studies. The main results are that machine learning techniques such as prediction and regression models were used mostly to predict student performance. They ended the research by highlighting some major research challenges and give significant recommendations to motivate future works in this field.

Authors in [9] performed analysis to identify the impact of many features such as student social activities, student background and student coursework achievement. These features are used to estimate student academic performance. Supervised learning methods such as Random Forest, Naïve Bayesian, Decision Tree J48 and Multilayer Perceptron were employed in estimating the performance of students for Math subject in the school. The prediction was done on multi-levels classification on final grade. The results showed that the two features social activities and student background have a great impact on estimating performance of student on 2-level classification. The authors developed this model to help in improving performance of students. Authors did work with deep learning and not use optimization technique to optimize their method to get best results.

In [10], authors created various models for estimating performance of students, they worked on data of students from Australian university. The dataset many features such as details of student enrolment and student activity details which are created by a management system. The above details contain information about:

socio-demographic

university admission basis

attendance (full or part time).

Authors considered student heterogeneity in constructing the predictive models. The results showed the effectively of both features enrolment and course activity; they help in finding students precisely with bad performance.

Authors in [11] developed a new prediction algorithm for evaluating student's performance in academia depending on clustering and classification; it has been validated on data in real time; the used dataset is consisted of different academic university from, India. The validation proved that usage of algorithm of classification and clustering resulted high accuracy.

In our work, we are going to use and build 3 algorithms SVM, ANN, LR with tuning their parameters using genetic algorithm and comparing their results with other studies. We also will find the effectively of using traditional ML algorithms (SVM, ANN, LR) subjected to deep learning CNN. Related works did not use optimization technique to optimize their methods.

3- The developed algorithm

In this section, details of proposed algorithm are explained. It starts with basic **operations of dataset cleaning and then a brief description of Genetic Algorithm is** introduced. The basic developed algorithms are explained with details.

3-1- Dataset Cleaning

The first stage of our is data cleaning as follows:

- The column with string data type needs to be converted to numeric.
- The columns with null values need to be handled.
- The columns with nan values need to be handled (replaced with -1).
- Unified the tables into one table with primary column since there are many tables for the used dataset

Conversion to numeric type is done after replacing null and nan values with the string "-1". The conversion process has two types:

- Convert the row string to number such as "-1" to -1
- Indexing string columns to valid number such as the column with gender features "Female" become 0, "Male": become 1.

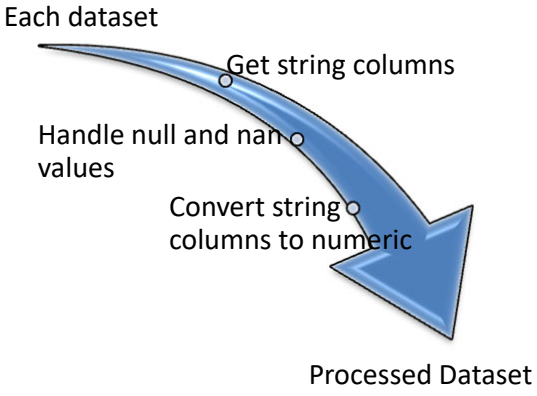

Figure 1: Dataset cleaning

Each processed dataset has a primary key, the last step is to gather all processed datasets into one dataset.

3-2- Genetic Algorithm:

Depending on [12]. Genetic algorithm (GA) is an algorithm that uses the concept of natural selection; natural selection that depends on evolutionary algorithms (EA). Genetic algorithms are used to find out

solutions for complex problems that need optimization relying on biologically inspired operators such as mutation, crossover and selection. Genetic Algorithms produce solutions with high quality for optimization issues. To get more details about GA, see [13]

The flowchart of GA process is as follows:

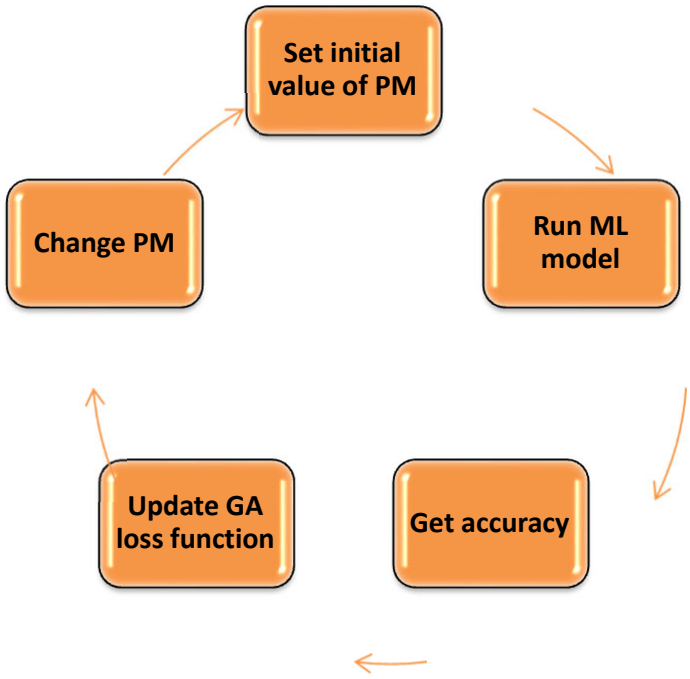

Figure 2: GA Process

Suppose the parameter is PM, we start by initial value and after update GA loss function, PM must be changed. The loop ends after a specific number of iterations.

3-3- SVM Model

Support vector machines (SVM) is a famous and well-known supervised learning method; it is employed mainly for prediction, regression and classification. There are positive points of SVM: it works with in high dimensional data. The main contribution here is that the selection of the term tolerance of SVM algorithm was done using Genetic algorithm method

3-4- LR Model

Logistic regression is similar to linear regression, logistic regression is used to find and predict the relation between many variables (dependent and independent); the relation is a function between input variables and output variable. It finds the relation based on math model and best relation has minimum error between true output and predicted output. The math model has many parameters that can have

default values or can be tuned. GA is used to find best values of Tolerance and Max iteration parameters of LR model [14].

3-5- ANN Model

The used ANN is Multilayer perceptron (MLP) which is feed forward neural network. It has input layer, output layer and hidden layer. The perceptron is a simple neuron model that takes input signals (patterns) coded as (real) input vectors $x = (x_1, x_2, \ldots, x_{n+1})$ through the associated (real) vector of synaptic weights $w = (w_1, w_2, \ldots, w_{n+1})$. The output o is determined by:

$$o = f(net) = f(w.x)$$

$$o = f\left(\sum_{j=1}^{n+1} w_j . x_j\right)$$

Biases to each neuron is added also. The equation is applied to each layer input to get layer output. The number of layers is a basic variable that can has in direct impact on the final output. GA find the optimal number of neurons in the three layers [16].

The next flowchart shows the general structure of our proposed method

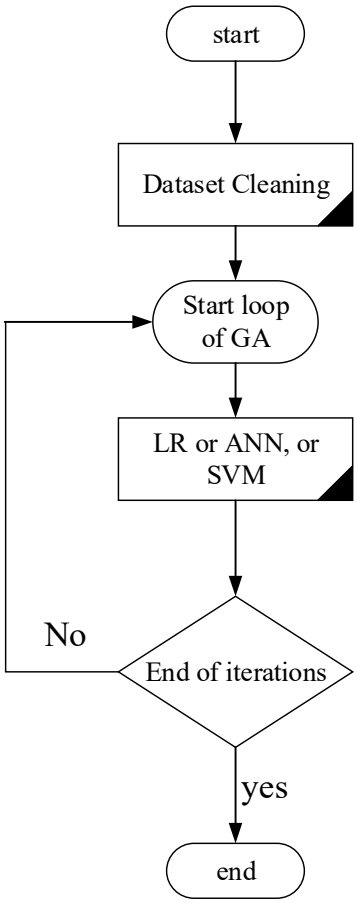

Figure 3: general flowchart of the proposed method

4- Implementation:

The implementation is done using python programming language with Anaconda software. Many libraries were used in the program such as machine learning libraries. Three methods were implemented and results were compared with state of art.

4-1-OULAD dataset:

OULAD dataset has details about students, their activities and subjects. The information is related to Virtual Learning Environment (VLE); there are information about 7 subjects. Data was collected for 9 months (from February to October). The dataset contains many tables that are existed in form of csv format1 [17]. The dataset contains sub datasets in form of csv files, we used the following tables:

- assessments.csv: it has details about homework with presentations. Each presentation has a unique number of assessments with final exam.

- studentInfo.csv: This file has demographic information about the students together with their results.
- studentAssessment.csv: it has students' grades of subjects, materials and homework. It contains no results in case when student did not do any of them.
- studentVle.csv: The studentVle.csv has details about all students' activities and how they interact with the subjects in the VLE.

4-2- Cleaning dataset

First stage is to clean datasets and get final file that summarized all sub datasets into one final dataset.

```
>>> studentAssessment
        id_assessment  id_student  date_submitted  is_banked  score
0                1752       11391              18          0   78.0
1                1752       28400              22          0   70.0
2                1752       31604              17          0   72.0
3                1752       32885              26          0   69.0
4                1752       38053              19          0   79.0
...               ...         ...             ...        ...    ...
173907          37443      527538             227          0   60.0
173908          37443      534672             229          0  100.0
173909          37443      546286             215          0   80.0
173910          37443      546724             230          0  100.0
173911          37443      558486             224          0   80.0
```

Figure 4: each student with score for assessment in specific date

The main table is student info that contains the final result of each student in the modules

```
>>> studentInfo
       id_student      highest_education  studied_credits  num_of_prev_attempts final_result disability
0           11391       HE Qualification              240                     0         Pass          N
1           28400       HE Qualification               60                     0         Pass          N
2           30268  A Level or Equivalent               60                     0    Withdrawn          Y
3           31604  A Level or Equivalent               60                     0         Pass          N
4           32885     Lower Than A Level               60                     0         Pass          N
...           ...                    ...              ...                   ...          ...        ...
32588     2640965     Lower Than A Level               30                     0         Fail          N
32589     2645731     Lower Than A Level               30                     0  Distinction          N
32590     2648187  A Level or Equivalent               30                     0         Pass          Y
32591     2679821     Lower Than A Level               30                     0    Withdrawn          N
32592     2684003       HE Qualification               30                     0  Distinction          N
```

The final result was indexed as follows:

Table 1: indexing final result

Final result	Indexing
Pass	3
Withdraw	2
Distinction	4

Fail	1
Nan	-1

the indexing is important since ML algorithms only work with numeric values. So, the index 3 when it is found in the data means that this row for a student with pass result. The value 2 means the state of student is Withdraw and so forth.

Table 2: indexing highest education

Final result	Indexing
HE Qualification	4
A level or Equivalent	3
Lower Than a Level	2
Post Graduate Qualification	5
No Formal quals	1
Nan	-1

	A	B	C	D	E	F	G	H	I	J
1	date_submitted	disability	final_result	highest_education	id_student	num_of_prev_attempts	score	studied_credits	sum_of_sum_click	procastinate
2	18	1	3	4	11391	0	78	240	1435	0
3	22	1	3	4	28400	0	70	60	0	0
4	17	2	2	3	30268	0	72	60	0	0
5	26	1	3	3	31604	0	69	60	0	0
6	19	1	3	2	32885	0	79	60	0	0
7	20	1	3	3	38053	0	70	60	0	0
8	18	1	3	4	45462	0	72	60	0	0
9	19	1	3	3	45642	0	72	120	0	0
10	9	1	3	3	52130	0	71	90	0	0
11	18	1	3	5	53025	0	68	60	0	0
12	19	1	3	2	57506	0	73	60	0	0
13	18	1	3	3	58873	0	67	60	0	0
14	17	1	3	2	59185	0	73	60	0	0
15	19	1	3	4	62155	0	83	60	0	0
16	17	1	3	2	63400	0	66	60	281	0
17	19	1	2	3	65002	0	59	60	0	0
18	19	1	3	3	70464	0	82	60	0	0
19	22	1	3	4	71361	0	60	60	0	0
20	18	1	1	3	74372	0	67	150	2158	1
21	18	1	3	3	75091	0	73	60	0	0
22	21	1	3	3	77367	0	75	60	0	0
23	17	1	3	4	91265	0	74	60	0	0

Figure 5: final table after cleaning

4-3- Classifiers results

The confusion matrix for SVM

Table 3: SVM Confusion Matrix

	precision	*recall*	*f1-score*	*support*
0	***1.00***	***1.00***	***1.00***	***96729***
1	***1.00***	***1.00***	***1.00***	***3075***
accuracy			***1.00***	***99804***
macro avg	***1.00***	***1.00***	***1.00***	***99804***
weighted avg	***1.00***	***1.00***	***1.00***	***99804***

The confusion matrix for LR and ANN is the same for SVM. ANN loss curve is shown below

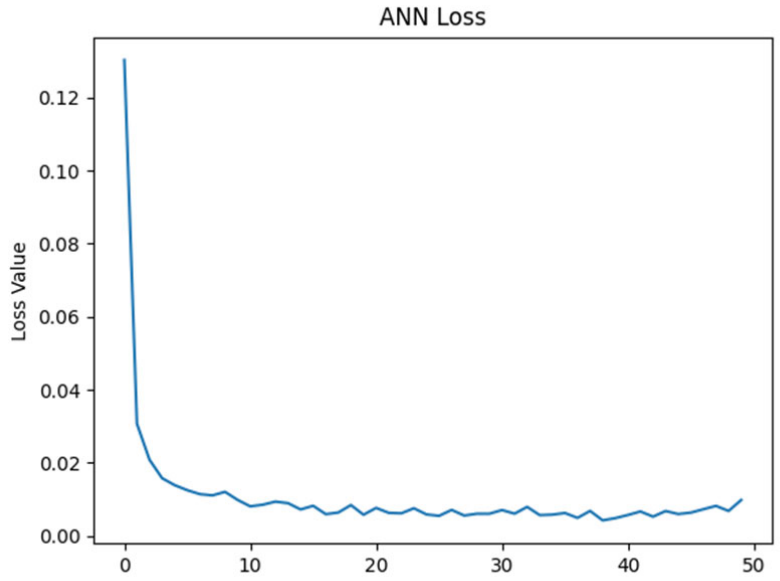

Figure 6: ANN loss curve

The figure above showed that when number of iterations is increased to reach 50 iteration the loss function value will change around a stable and minimum value about 0.01, at first iteration the loss value has a great value. When curve goes from large value to small value, this means that GA is finding best parameters the minimize loss function.

Table 4: Results summarization

Classifier	SVM	LR	ANN
Best value	Tolerance=0.00005	Tolerance=0.00009 Max iteration = 1000	Three layers With neurons (10,10,5)
Accuracy	99.99 %	100.0 %	99.99%

The results above show that GA founded best values of terms Tolerance and number of layers and neurons. With these best values, the results of predictions outperformed other researches as shown in Table 5 where Table 5 contains a comparing the results with state of arts (all studies worked on OULAD dataset)

Table 5: Results comparison

Study	Techniques	Accuracy
[16]	Decision tree, random forest, extreme gradient boosting, multilayer perceptron	78.2 %
[18]	CNN and Long Short-Term Memory	61 %
[19]	Decision Tree	83.14 %
[20]	Naïve Bayes	63.8 %
[21]	Hybrid 2D CNN	88 %
Ours	SVM LR ANN	99.99 % 100.0 % 99.99 %

The results above showed that normal machine learning with parameters tuning will give high accuracy. Therefor, deep learning will not necessarily give better results than traditional ML algorithms.

5- Conclusion:

This paper dealt with an important problem for academic performance prediction of students. The main contribution is using Genetic Algorithm to find optimal values of ML model parameters. The results showed that traditional ML algorithms are better than CNN in this field of study. Moreover SVM, LR, ANN with optimization techniques are able to outperform the deep learning models for the same problems since the three algorithms have performance more than 99.9%.

References:

1 Satyanarayana, A. and Nuckowski, M., 2016. Data mining using ensemble classifiers for improved prediction of student academic performance.

2 Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers & Education, 61, 133-145.

3 Cohen, L., Manion, L., & Morrison, K. (2002). Research methods in education. Routledge.

4 Ware, W. B., & Galassi, J. P. (2006). Using correlational and prediction data to enhance student achievement in K-12 schools: A practical application for school counselors. Professional School Counseling, 344-356.

5 Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:1203.3832.

6 Oyedeji, A.O., Salami, A.M., Folorunsho, O. and Abolade, O.R., 2020. Analysis and prediction of student academic performance using machine learning. JITCE (Journal of Information Technology and Computer Engineering), 4(01), pp.10-15.

7 Son, L.H. and Fujita, H., 2019. Neural-fuzzy with representative sets for prediction of student performance. Applied Intelligence, 49(1), pp.172-187.

8 Namoun, A. and Alshanqiti, A., 2020. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), p.237.

9 Kiu, C.C., 2018, October. Data mining analysis on student's academic performance through exploration of student's background and social activities. In 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA) (pp. 1-5). IEEE.

10 Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D.J. and Long, Q., 2018. Predicting academic performance by considering student heterogeneity. Knowledge-Based Systems, 161, pp.134-146.

11 Francis, B.K. and Babu, S.S., 2019. Predicting academic performance of students using a hybrid data mining approach. Journal of medical systems, 43(6), pp.1-15.

12 Mitchell, M., 1998. An introduction to genetic algorithms. MIT press.

13 Alam, T., Qamar, S., Dixit, A. and Benaida, M., 2020. Genetic algorithm: Reviews, implementations, and applications. arXiv preprint arXiv:2007.12673.

14 Stoltzfus, J. C. (2011). Logistic regression: a brief primer. Academic emergency medicine, 18(10), 1099-1104.

15 Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling, 160(3), 249-264.

16 Tino, P., Benuskova, L., & Sperduti, A. (2015). Artificial neural network models. In Springer Handbook of Computational Intelligence (pp. 455-471). Springer, Berlin, Heidelberg.

17 Jha, N. I., Ghergulescu, I., & Moldovan, A. N. (2019, May). OULAD MOOC Dropout and Result Prediction using Ensemble, Deep Learning and Regression Techniques. In CSEDU (2) (pp. 154-164).

18 Song, X.; Li, J.; Sun, S.; Yin, H.; Dawson, P.; Doss, R.R.M. SEPN: A Sequential Engagement Based Academic Performance Prediction Model. IEEE Intell. Syst. 2021, 36, 46–53

19 Rizvi, S.; Rienties, B.; Khoja, S.A. The role of demographics in online learning; A decision tree-based approach. Comput. Educ. 2019, 137, 32–47

20 Azizah, E.N.; Pujianto, U.; Nugraha, E. Comparative performance between C4. 5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment. In Proceedings of the 4th International Conference on Education and Technology (ICET), Malang, Indonesia, 26–28 October 2018; pp. 18–22

21 Poudyal, S., Mohammadi-Aragh, M.J. and Ball, J.E., 2022. Prediction of Student Academic Performance Using a Hybrid 2D CNN Model. Electronics, 11(7), p.1005.