I. J. Education and Management Engineering, **2020, 6, 27-33**
Published Online December 2020 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2020.06.04

Predicting Students' Academic Performance in Educational Data Mining Based on Deep Learning Using TensorFlow

Mussa S. Abubakari [*], **Fatchul Arifin**
Department of Electronics & Informatics Engineering Education, Postgraduate Program, Universitas Negeri Yogyakarta, Yogyakarta 55281, Indonesia
E-mail: abu.mussaside@gmail.com [*], fatchul@uny.ac.id

Gilbert G. Hungilo
Department of Informatics Engineering, Graduate Program, University Atma Jaya Yogyakarta, Yogyakarta 55281, Indonesia
E-mail: gutabagaonline@gmail.com

Received:07 May 2020; Accepted: 26 July 2020; Published: 08 December 2020

Abstract: The study was aimed to create a predictive model for predicting students' academic performance based on a neural network algorithm. This is because recently, educational data mining has become very helpful in decision making in an educational context and hence improving students' academic outcomes. This study implemented a Neural Network algorithm as a data mining technique to extract knowledge patterns from student's dataset consisting of 480 instances (students) with 16 attributes for each student. The classification metric used is accuracy as the model quality measurement. The accuracy result was below 60% when the Adam model optimizer was used. Although, after applying the Stochastic Gradient Descent optimizer and dropout technique, the accuracy increased to more than 75%. The final stable accuracy obtained was 76.8% which is a satisfactory result. This indicates that the suggested NN model can be reliable for prediction, especially in social science studies.

Index Terms: Classification, Data Mining Techniques, Educational Data Mining, Neural Network Algorithm, Predictive Model.

1. Introduction

Currently, data mining has become an interesting topic for many researchers in various fields such as medicine, engineering, and even educational field. Especially in educational context, through mining of students' information, it has become easier to make decisions concerning students in their academic performance [1, 2]. The prediction of students' performance is a vital matter in educational context as predicting future performance of students after being admitted into a college, can determine who would attain poor marks and who would perform well. These results can help make efficient decisions during admission and hence improve the academic services quality [3–5].

Analysis of educational data using data-mining techniques helps extract unique information of students from educational database and use that hidden information to solve various academic problems of students by understanding learners, improve teaching-learning methods and process [6, 7]. Moreover, these data mining techniques help educational stakeholders to make quality decisions to enhance students' outcomes.

Various methods like Decision tree and Naïve Bayesian were used by many researchers for predicting learners' academic performance and make decisions to help those who need help immediately [7]. Other researchers used ensemble methods such as Random Forest (RF), AdaBoosting, and Bagging as classification methods [7, 8]. Different data mining methods can solve different educational problems such as classification and clustering. The famous known data mining method in prediction models is classification. Various deep learning algorithms like Neural Networks, are used under

classification matter [9].

In the current study, neural network (NN) classification algorithm is implemented to create a predictive model in predicting academic performance of students in a particular academic institution by using students' characteristics and their distinctive demographic data. A predictive model based on NN approach can be useful in decision making on academic success of students and therefore enhancing academic management and improving quality education.

2. Related Works

Various studies have been conducted concerning data mining in educational context for uncovering knowledge patterns from students' information for improving academic performance of students. This current study will base its theoretical background on the previous research done on the educational data mining contexts as explained below.

The study was conducted on engineering students based on different mining techniques for making academic decisions. Techniques involving classification rules and association rules for discovering knowledge patterns, were used to predict the engineering student's performance. The study experiment also clustered the students based on k-means clustering algorithm [10]. In another study, students' performance was evaluated based on association rule algorithm. The research was done by assessing the performance of students based on different features. The experiment was implemented based on real time dataset found in the school premises using Weka [11].

Baradwaj and Pal explained in their study on student's assessment by using a number of data mining methods. Their study facilitated teachers to identify students who need special attention to reduce the fail percentage and help to take valid measure for next semesters [3]. Also, another study was done to develop a classification model to predict student performance using Deep Learning which learns multiple levels of representation automatically. They used unsupervised learning algorithm to pre-train hidden layers of features layer-wisely based on a sparse auto-encoder from unlabeled data, and then supervised training was used for the parameters fine-tuning. The resulted model was trained on a relatively huge real-world students' dataset, and the experimental findings indicate the effectiveness of the proposed method to be implemented into academic pre-warning mechanism [12].

Other researchers developed models to predict students' university performance based on students' personal attributes, university performance and pre-university characteristics. The studies included the data of 10,330 students Bulgaria with every student having 20 attributes. Algorithms such as the K-nearest neighbour (KNN), decision tree, Naive Bayes, and rule learner's algorithms were applied to classify the students into 5 classes: Excellent, Very Good, Good, Bad or Average. Overall accuracy was below 69%. However, decision tree classifier showed best performance having the highest overall accuracy, followed by the rule learner [13, 14].

Recently, the study was conducted to predict user's intention to utilize peer-to-peer (P2P) mobile application for transactions. Logistic regression (LR) analysis technique together with neural network were used to predict the technology adoption. The results indicated that NN model has higher accuracy than LR model [15]. Another study proposed a student performance model with behavioral characteristics. These characteristics are associated with the student interactivity with an e-learning platform. Data mining techniques such as Naïve Bayesian and Decision Tree classifiers were used to evaluate the impact of such features on student's academic performance. The results of that study revealed that there is a strong relationship between learner behaviors and its academic achievement [16].

In this study, a predictive model is created based on neural network (NN) classification algorithm in predicting academic performance of students by using students' behavioral characteristics and their distinctive demographic data as variables. A predictive model using NN data mining approach can help in making decisions and conclusions on academic success of students hence enhancing academic management and improve education quality.

3. Methodology

3.1 Data Collection

The student data implemented in this project were obtained from educational dataset collected by [16] from learning management system (LMS) in The University of Jordan, Amman, Jordan during the study conducted in 2015. The dataset is available in the kaggle website (https://www.kaggle.com/aljarah/xAPI-Edu-Data). The dataset comprised of 480 (instances) of student records and their 16 respective attributes. These attributes were grouped into three classes, namely (i) Behavioral attributes include parents answering survey, school satisfaction, opening resources, and raised hand on class, (ii) Academic background attributes including grade Level, educational stage, and section, and (iii) Demographic features including nationality and gender. The dataset also includes 175 females and 305 males. The students have different nationalities including from Kuwait (179), USA (6), Jordan (172), Iraq (22), Lebanon (17), Tunis (12), Saudi Arabia (11), Egypt (9), from Iran, Syria, and Libya were 7 each, Morocco (4), 28 students from Palestine, and one from Venezuela.

Another attribute is school attendance having two groups based on days of class absence: 191 students exceeded 7 days and 289 students were absent under 7 days. Moreover, the dataset includes also a new kind of attribute namely parent participation having two sub attributes: Parent School Satisfaction and Parent Answering Survey. 270 parents participated in a survey answering and 210 did not, 292 parents were satisfied from the school and 188 were not. The students are

grouped into three classes based on their total grades, namely High-Level, Middle-Level, and Low-Level [8]. Appendix A summarizes the students' attributes and their description.

3.2 Methods and Data Preparation

For this study, authors used Anaconda software environment for python machine learning language together with keras machine learning library and specifically TensorFlow utility which is powerful to create and evaluate the proposed NN classification model [17–19]. Keras is a python library widely used in deep-learning that run on top of TensorFlow and Theano, providing an intuitive best API for Python in NNs [20, 21]. Since the dataset used in this study contains variables (attributes) with different categories, there was a need to transform them into a form the computer and NN model can understand. The dataset explained above consists of three main categories of variables. First are nominal variables with two categories such as gender (male or female), semester (first or second), and others. Second, are variables with numerical values such as visited resources, raised hand, and others. And third, are nominal variables with more than three categories such as grade levels (G-01 to G-12), topic (English, Math, Chemistry, and so on), and other variables as it can be seen in Appendix A.

Nominal variables with two categories were transformed using label encoder mechanism. While, those with three or more categories were transformed using one-hot encoding (dummies method). Furthermore, continuous numerical variables were transformed by normalizing them using min-max scaler mechanism for normal distribution.

4. Experiment Process and Results

After data transformation as explained above, the inputs increased from 16 inputs to 39 inputs and the output (classification outputs) of 3 outputs making a total of 42 columns in the NN model. After that, the dataset was split into train data and test data with data for testing consisting of less than 26% of all dataset and the remaining percentage for training.

The following step was to create a predictive model based on Artificial Neural Network (ANN) classification technique to evaluate the attributes which influence directly or indirectly student's academic success. ANN technique is an implementation of artificial neural network that involves training data inputs for the best accuracy achievement. A cross validation with 10-fold was used to divide the dataset for training and testing process. Then the process was followed by fitting the model by 200 iteration (epochs) with 10 batch-size of inputs and then followed by the results evaluation for generating knowledge representation. The evaluation measure used is accuracy for classification quality. Accuracy is the proportion or ratio of the total number of correct predictions to incorrectly predicted.

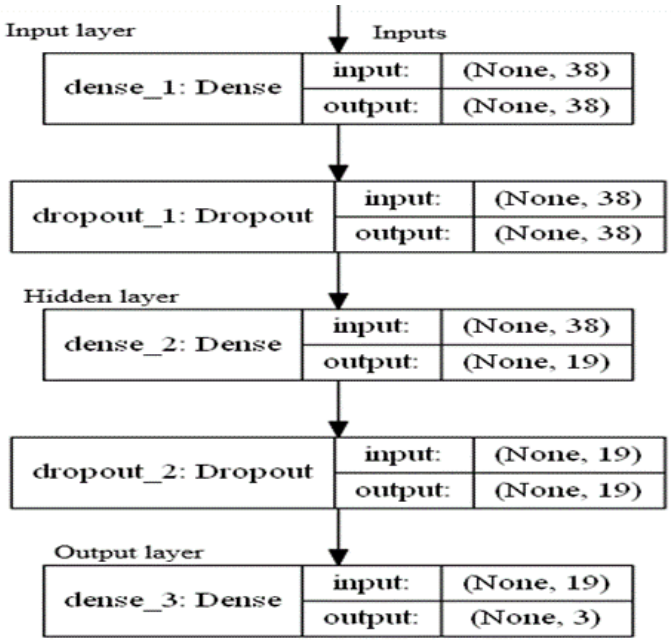

Fig. 1. The NN Model Structure.

Figure 1 above shows the NN model structure created by a python code as can be seen in the last code line in Appendix B. The NN predictive model used in this study consists of three layers: (1) input layer with 39 neurons, (2) hidden layer

with 19 neurons and (3) an output layer with 3 outputs. The input layer receives input data from 16 attributes and the output layer send output of three grade categories, namely Low (L). Middle (M), and High (H). There is a hidden layer between the input layer and output layer. Appendix B illustrate the python code used to create, fit, and validate the NN model.

In this study, we used accuracy as the metric for prediction quality of the developed NN model. Also, only NN algorithm was used for classification of the student dataset.

The result of the experiment has two versions due to the implementation of two different model (function) optimizers namely, Adam and Stochastic gradient descent (SGD) as well as due to the introduction of dropout technique to the NN model development to drop (20% of neurons were dropped in this study) loosely connected neuron. The result indicates that when we applied Adam optimization technique the accuracy was below 60%. While, when we applied the SGD optimizer the accuracy improved to more than 76%.

Moreover, the dropout technique helped to improve the accuracy value to more than 76.5%. The dropout technique is used to remove the loosely connected neurons as the NN technique performs better with fully connected neurons. The final stable result was 76.8% accuracy.

5. Conclusion and Future Work

Education is a vital element in any community for their social-economic development. Data mining techniques or business intelligence allows extracting knowledge patterns from students' raw data offering interesting chances for the educational context. Particularly, various studies have implemented machine learning techniques like Decision Tree and Random Forest to enhance the management of college resources and hence improving education quality.

In this study, the authors have presented a predictive model using NN technique to learn the patterns from students' data and predict their academic performance. By applying data mining techniques on students' database, academic stakeholders can find the important factors which have direct or indirect impacts on the student's academic success. The knowledge patterns and results discovered in this study after applying NN classification method indicate that different attributes of students have impacts on their learning process as it can be seen in the classification accuracy results. The final classification accuracy obtained in this study is 76.9% which is more than satisfactory percentage for our predictive model developed using NN algorithm.

Like other studies, this study is with some limitations too. One of which is the dataset can only be applied to the similar context as this study. Also, the results presented here involves the accuracy as the only predictive measure of model quality. Moreover, only one algorithm, NN algorithm was used for classification purpose.

For future studies, authors intend to use the localized student data from a particular university in Yogyakarta, especially from Yogyakarta State University. Also, in the future we expect to apply other data mining methods such as RF, DT, and others in the localized dataset. Moreover, future experiments will add more measurement classification qualities such as Precision, sensitivity, and Recall.

Acknowledgements

Much appreciation to my close friends who inspired me to do this work.

References

[1] S. K. Mohamad and Z. Tasir, "Educational Data Mining: A Review," Procedia - Soc. Behav. Sci., vol. 97, pp. 320–324, 2013.
[2] M. Chalaris, S. Gritzalis, M. Maragoudakis, C. Sgouropoulou, and A. Tsolakidis, "Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques," Procedia - Soc. Behav. Sci., vol. 147, pp. 390–397, 2014.
[3] B. Brijesh Kumar and P. Saurabh, "Mining Educational Data to Analyze Students‟ Performance," Int. J. Adv. Comput. Sci. Appl., vol. 2, no. No. 6, pp. 59–63, 2011.
[4] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 3, pp. 1584–1592, 2019.
[5] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," Telemat. Informatics, vol. 37, pp. 13–49, 2019.
[6] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA," Indones. J. Electr. Eng. Comput. Sci., vol. 9, no. 2, pp. 447–459, 2018.
[7] S. S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "A data mining approach to predict academic performance of students using ensemble techniques," in Advances in Intelligent Systems and Computing, 2020, vol. 940, no. March, pp. 749–760.
[8] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," Int. J. Database Theory Appl., vol. 9, no. 8, pp. 119–136, 2016.
[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in Procedia Computer Science, 2015, vol. 72, pp. 414–422.
[10] R. Singh, "An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education," Int. J. Comput. Sci. Mob. Comput., vol. 2, no. February, pp. 53–57, 2013.

[11] S. Borkar and K. Rajeswari, “Predicting students academic performance using education data mining,” Int. J. Comput. Sci. Mob. Comput., vol. 2, no. 7, pp. 273–279, 2013.
[12] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, “Predicting Students Performance in Educational Data Mining,” in Proceedings - 2015 International Symposium on Educational Technology, ISET 2015, 2016, pp. 125–128.
[13] D. Kabakchieva, “Predicting student performance by using data mining methods for classification,” Cybern. Inf. Technol., vol. 13, no. 1, pp. 61–72, 2013.
[14] D. Kabakchieva, K. Stefanova, and V. Kisimov, “Analyzing university data for determining student profiles and predicting performance,” in EDM 2011 - Proceedings of the 4th International Conference on Educational Data Mining, 2011, pp. 347–348.
[15] J. Lara-Rubio, A. F. Villarejo-Ramos, and F. Liébana-Cabanillas, “Explanatory and predictive model of the adoption of P2P payment systems,” Behav. Inf. Technol., vol. 0, no. 0, pp. 1–14, 2020.
[16] E. A. Amrieh, T. Hamtini, and I. Aljarah, “Preprocessing and analyzing educational data set using X-API for improving student’s performance,” in 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2015, 2015.
[17] P. S. Janardhanan, “Project repositories for machine learning with TensorFlow,” Procedia Comput. Sci., vol. 171, pp. 188–196, 2020.
[18] L. Hao, S. Liang, J. Ye, and Z. Xu, “TensorD: A tensor decomposition library in TensorFlow,” Neurocomputing, vol. 318, pp. 196–200, 2018.
[19] R. Orus Perez, “Using TensorFlow-based Neural Network to estimate GNSS single frequency ionospheric delay (IONONet),” Adv. Sp. Res., vol. 63, no. 5, pp. 1607–1618, 2019.
[20] V.-H. Nhu et al., “Effectiveness assessment of Keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area,” CATENA, vol. 188, p. 104458, 2020.
[21] K. Akyol, “Comparing of deep neural networks and extreme learning machines based on growing and pruning approach,” Expert Syst. Appl., vol. 140, p. 112875, 2020.

Authors’ Profiles

Mussa S. Abubakari was born in Kondoa, Tanzania in 1990. He received the B.Sc. degree in Telecommunications Engineering from the University of Dodoma, Tanzania in 2016. Currently he is the postgraduate candidate taking master degree in Electronics & Informatics Engineering Education at Universitas Negeri Yogyakarta, Indonesia. His research interests include technology enhanced learning, human computer interaction, technology acceptance, Internet of Things, mobile technologies, intelligent systems, and signal processing.

Dr. Fatchul Arifin was born on 08 Mei 1972. He received a B.Sc. in Electric Engineering at Universitas Diponegoro and PH.D. degree in Electric Engineering from Institut Teknologi Surabaya, in 1996 and 2014, respectively. Currently he is the lecturer at both undergraduate faculty of engineering and postgraduate program at Universitas Negeri Yogyakarta. His research interests include but not limited to intelligent control systems, machine learning, expert systems, and neural-fuzzy system.

Gilbert G. Hungilo is a master degree graduate from department of Informatics Engineering at the University Atma Jaya Yogyakarta, Indonesia. He received Bachelor of Science in Computer Science from the University of Dar es salaam, Tanzania. His research interests include technology adoption, big data analytics, and machine learning.

How to cite this paper: Mussa S. Abubakaria, Fatchul Arifin, Gilbert G. Hungilo. " Predicting Students' Academic Performance in Educational Data Mining Based on Deep Learning Using TensorFlow ", International Journal of Education and Management Engineering (IJEME), Vol.10, No.6, pp.27-33, 2020. DOI: 10.5815/ijeme.2020.06.04

Appendix A. Students' Attributes [16]

SN	Attribute	Description	Variable Type
1	Gender	Gender of Student: Female or Male.	Nominal(binary)
2	Nationality	Student's Origin: Kuwait, Iraq, Libya Lebanon, Egypt, USA, Morocco, Jordan, Iran, Tunis, Syria, Palestine, Saudi Arabia, Venezuela.	Nominal(dummy)
3	Birth Place	Student's Birth Place: Kuwait, Iraq, Libya Lebanon, Egypt, USA, Morocco, Jordan, Iran, Tunis, Syria, Palestine, Saudi Arabia, Venezuela.	Nominal(dummy)
4	Stage ID	Student Educational Level: High School, Middle School, Lower level.	Nominal(dummy)
5	Grade ID	Student Grade: G-01 up to G-12.	Nominal(dummy)
6	Section ID	Classroom student belongs: A, B, C.	Nominal(dummy)
7	Topic	Course Studied: Arabic, Biology, Chemistry, English, Geology, French, Spanish, IT, Math, Science, History, Quran.	Nominal(dummy)
8	Semester	School year semester: First, Second.	Nominal(binary)
9	Relation	Responsible Parent: Mom, Father.	Nominal(binary)
10	Raised hand	Frequency of raising hand in classroom: 0-100.	Numeric
11	Visited resources	Frequency of visiting course online content: 0-100.	Numeric
12	Announcements View	Frequency of checking the new online announcement: 0-100.	Numeric
13	Discussion	Frequency of participating in online discussion forums: 0-100.	Numeric
14	Parent Survey Answering	Whether Parents answered or not the survey: Yes, No.	Nominal(binary)
15	Parent School Satisfaction	Whether a parent is satisfied or not: Yes, No.	Nominal(binary)
16	Student Absence Days	The number of absence days a student was absent: Above or Under 7 days.	Nominal(binary)
17	Class	The grade class: High-Level (H): from 90-100; Middle-Level (M): from 70 to 89; Low-Level (L): from 0 to 69.	Nominal(dummy)

Appendix B. A Piece of Python Code Used to Create and Validate an NN Model

```
dataframe = dataset.values
X = dataframe[:,0:38]
Y = dataframe[:,38:41]
```

```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.25,random_state=10)
```

```
model = Sequential()
model.add(Dense(38, input_dim = 38, activation='relu', kernel_initializer = 'uniform', W_constraint=maxnorm(3)))
model.add(Dropout(0.2))
model.add(Dense(19, activation='relu', kernel_initializer = 'uniform', W_constraint=maxnorm(3)))
model.add(Dropout(0.2))
model.add(Dense(3,activation = 'softmax'))

sgd = SGD(lr=0.01, momentum = 0.8, decay = 0.0, nesterov = False )

model.compile(loss = 'categorical_crossentropy', optimizer= sgd, metrics = ['accuracy'])
```

```
model.fit(X_train,Y_train, epochs = 200, batch_size = 10, verbose=1)
_, accuracy = model.evaluate(X_test,Y_test, verbose = 1)
print('accuracy: %.2f'%(accuracy*100))
```

```
model.predict(X_test)
```

```
plot_model(model,show_shapes =True, to_file = 'student_model.png')
```

