The 1st International Conference On Health, Social, Sciences, and Technology (ICoHSST 2020)

Educational Data Mining to Predict Students Performance Based on Deep Learning Neural Network

1st Mussa S Abubakari
Department of Electronics & Informatics
Universitas Negeri Yogyakarta
Yogyakarta, Indonesia
Email: mussas.2019@student.uny.ac.id

2nd Suprapto
Department of Electronics & Informatics
Universitas Negeri Yogyakarta
Yogyakarta, Indonesia
Email: suprapto@uny.ac.id

Corresponding author: mussas.2019@student.uny.ac.id

***Abstract* - Recently, educational data mining has become very helpful in decision making in an educational context and hence improving students' academic outcomes. Thus, the goal of this study was to create a predictive model to predict students' academic performance based on a neural network algorithm. Authors implemented a Neural Network data mining technique using Anaconda 3 as datamining tool to extract knowledge patterns from student's dataset consisting of 131 students with 22 attributes for each student. The classification metric used is accuracy as the model quality measurement. The result indicates that when SGD optimizer was applied, the accuracy was below 80%. While, when Adam optimization technique was applied the accuracy improved to more than 96% which is more than a satisfactory percentage for our predictive model. This indicates that the suggested NN model can be reliable for prediction, especially in social science studies like education.**

Keywords - Classification, Data Mining Techniques, Educational Data Mining, Neural Network Algorithm, Predictive Model.

I. INTRODUCTION

The rapid advancement of science and technology, especially information and communication technology (ICT), is a fact which cannot be denied. The existence of ICT makes human life easier than some decades ago. ICT increasingly shows its glory in various fields of life, one of which is the educational aspect [1], [2]. The development of ICT encourages various educational institutions to use artificial intelligence to increase the effectiveness and flexibility of learning for better academic outcomes.

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' data, it has become easier to make decisions concerning students in their academic performance. 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], [4], [5].

A. Related Work

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 based 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 [6]. 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 [7].

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 [5]. 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 [8].

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 [9, 10].

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 [11]. 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 [12].

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.

B. Background

Analysis of students' 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 [13], [14]. Moreover, these data mining techniques help educational stakeholders to make quality decisions to enhance students' achievement. 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 [15].

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 [14]. Other researchers used ensemble methods such as Random Forest (RF), AdaBoosting, and Bagging as classification methods [14], [16]. In this current project, a predictive model is created based on neural network (NN) classification algorithm in predicting academic performance of students.

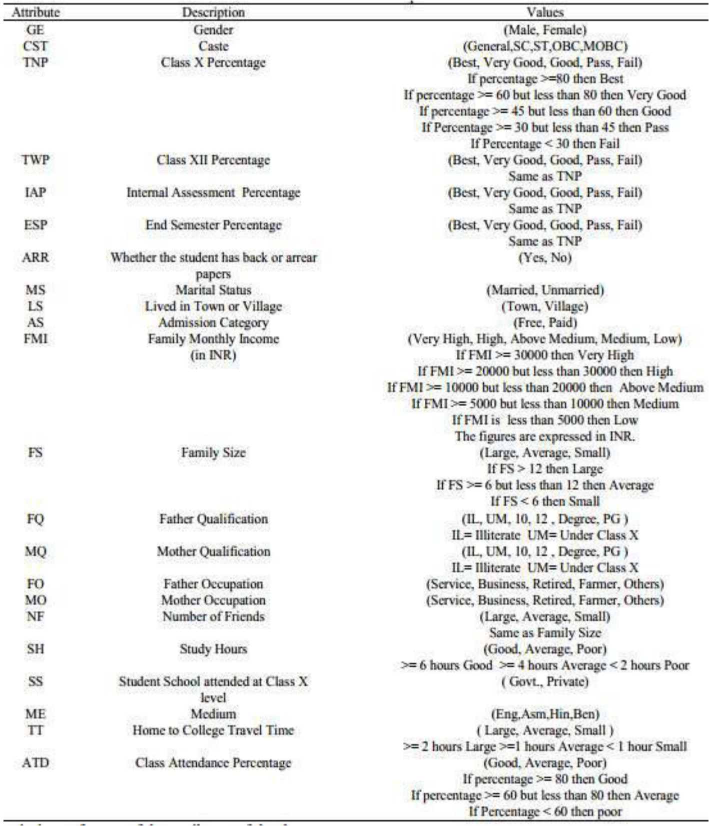
II. METHOD

A. Data Collection and Preparation

The student data implemented in this study were obtained from the study by [13]. and can be freely accessed and used from UCI Machine Learning Repository website. The dataset comprises demographic information, socio-economic features, and academic information of students. The total number of attributes of the dataset after data cleaning is twenty-two (22) consisting of 131 (instances) students. Note that, according to the source of dataset description, the dataset is supposed to have 300 instances (students) [13]., but actually the dataset only has 131 instances.

Since the dataset contains variables with different categories, there was a need to transform them into a form the computer and NN model can process. The dataset consists of three main categories of variables, namely nominal variables with two categories, variables with numerical, and nominal variables with three or more categories. 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. Table 1 shows the description of all variables in the dataset [13].

Table 1. Dataset description



Attribute	Description	Values
GE	Gender	(Male, Female)
CST	Caste	(General,SC,ST,OBC,MOBC)
TNP	Class X Percentage	(Best, Very Good, Good, Pass, Fail) If percentage >=80 then Best If percentage >= 60 but less than 80 then Very Good If percentage >= 45 but less than 60 then Good If Percentage >= 30 but less than 45 then Pass If Percentage < 30 then Fail
TWP	Class XII Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
IAP	Internal Assessment Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
ESP	End Semester Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
ARR	Whether the student has back or arrear papers	(Yes, No)
MS	Marital Status	(Married, Unmarried)
LS	Lived in Town or Village	(Town, Village)
AS	Admission Category	(Free, Paid)
FMI	Family Monthly Income (in INR)	(Very High, High, Above Medium, Medium, Low) If FMI >= 30000 then Very High If FMI >= 20000 but less than 30000 then High If FMI >= 10000 but less than 20000 then Above Medium If FMI >= 5000 but less than 10000 then Medium If FMI is less than 5000 then Low The figures are expressed in INR.
FS	Family Size	(Large, Average, Small) If FS > 12 then Large If FS >= 6 but less than 12 then Average If FS < 6 then Small
FQ	Father Qualification	(IL, UM, 10, 12 , Degree, PG) IL= Illiterate UM= Under Class X
MQ	Mother Qualification	(IL, UM, 10, 12 , Degree, PG) IL= Illiterate UM= Under Class X
FO	Father Occupation	(Service, Business, Retired, Farmer, Others)
MO	Mother Occupation	(Service, Business, Retired, Farmer, Others)
NF	Number of Friends	(Large, Average, Small) Same as Family Size
SH	Study Hours	(Good, Average, Poor) >= 6 hours Good >= 4 hours Average < 2 hours Poor
SS	Student School attended at Class X level	(Govt., Private)
ME	Medium	(Eng,Asm,Hin,Ben)
TT	Home to College Travel Time	(Large, Average, Small) >= 2 hours Large >=1 hours Average < 1 hour Small
ATD	Class Attendance Percentage	(Good, Average, Poor) If percentage >= 80 then Good If percentage >= 60 but less than 80 then Average If Percentage < 60 then poor

B. Methods and Tools

For this study, authors used Anaconda 3 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], [18], [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].

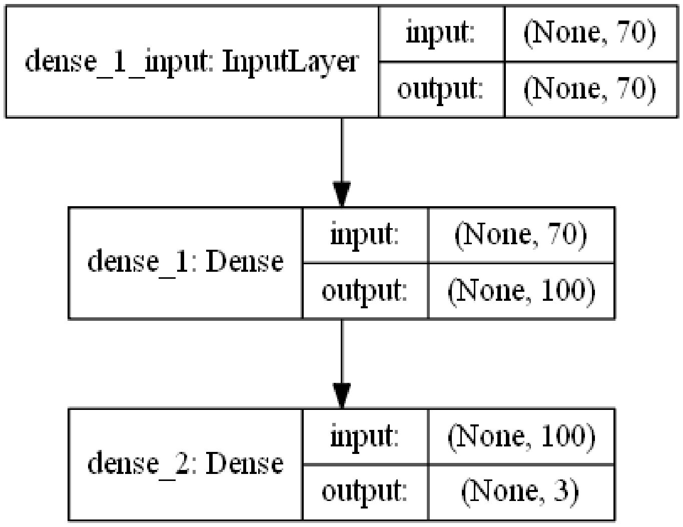
III. RESULTS AND DISCUSSION

After data transformation, the inputs increased from 22 columns (inputs) to 70 columns and the output (classification outputs) of 3 outputs making a total of 73 columns in the NN model matrix. After that, the dataset was split into train data and test data with 25% of data for testing and the remaining percentage for training.

The following step was to create a predictive model based on Neural Network algorithm to evaluate the attributes which influence directly or indirectly student's academic success. Moreover, 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 50 iteration (epochs) with 10 batch-size of inputs and then followed by model 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.

The NN predictive model consists of three layers: (1) input layer with 70 neurons, (2) hidden layer with 100 neurons and (3) output layer with 3 outputs. The input layer receives input data from 22 attributes and the output layer send output of three grade categories, namely Good, Average, and Poor. There is a hidden layer between the input layer and output layer. Figure 1 below shows the NN model structure created by a python code.

Fig. 1. The NN model structure created by a python code

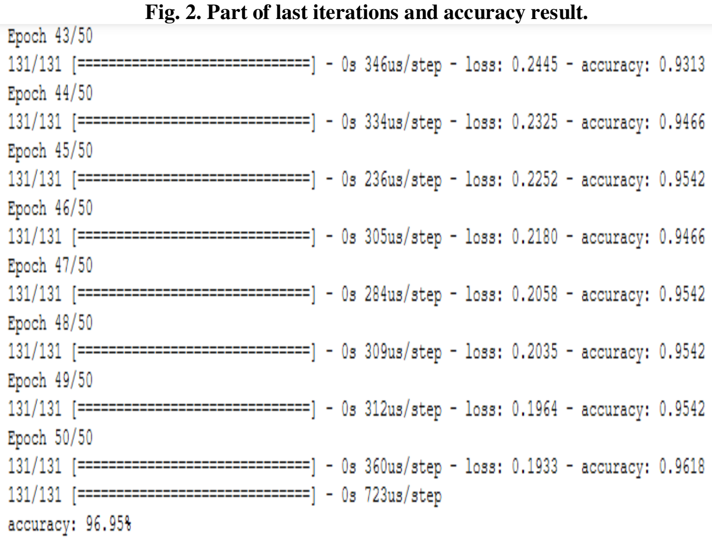


In this project, accuracy is used as the metric for measuring prediction quality of the developed NN model. Also, only NN algorithm was used for classification of the student dataset. The result of the experiment has two versions due to the implementation of two different model optimizers namely, Adam and Stochastic gradient descent (SGD).

The result indicates that when SGD optimizer was applied, the accuracy was below 80%. While, when Adam optimization technique was applied the accuracy improved to more than 96% which is more than a satisfactory percentage for our predictive model developed using NN algorithm. The knowledge patterns and results discovered in this project 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 Figure 2 below illustrate a part of last iterations and accuracy result after running the NN algorithm.

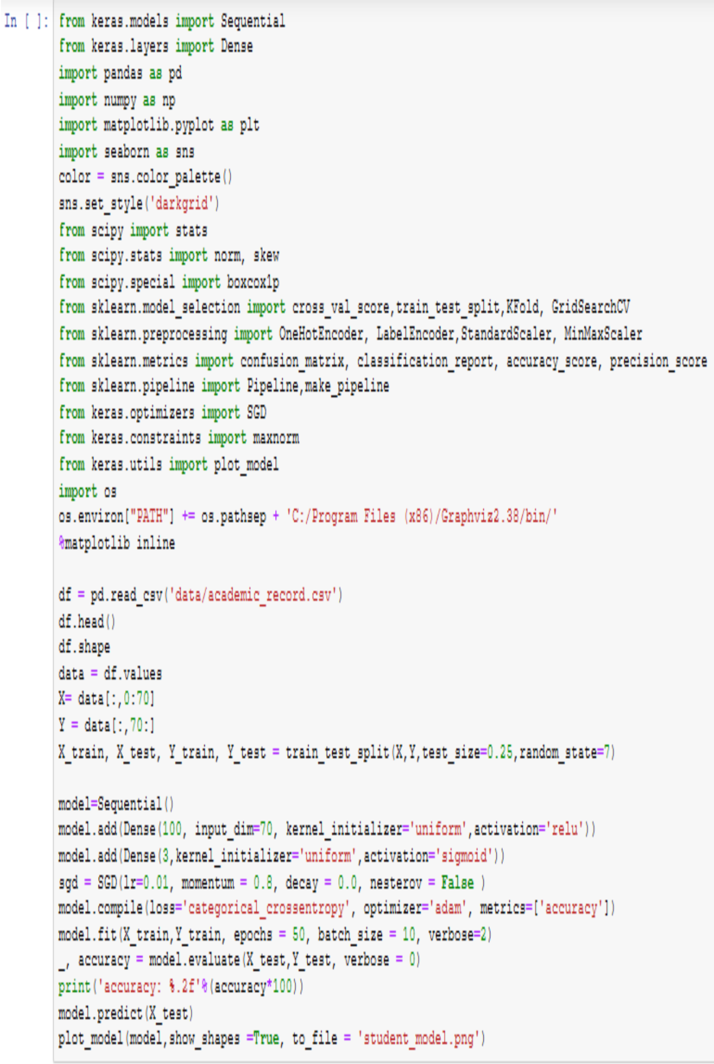
Also, Figure 3 depict the python code used to create, fit, and validate the NN model. Note that, for simplicity, the code blocks used to encode (transform) the dataset is omitted.

Fig. 2. Part of last iterations and accuracy result.



```
Epoch 43/50
131/131 [==============================] - 0s 346us/step - loss: 0.2445 - accuracy: 0.9313
Epoch 44/50
131/131 [==============================] - 0s 334us/step - loss: 0.2325 - accuracy: 0.9466
Epoch 45/50
131/131 [==============================] - 0s 236us/step - loss: 0.2252 - accuracy: 0.9542
Epoch 46/50
131/131 [==============================] - 0s 305us/step - loss: 0.2180 - accuracy: 0.9466
Epoch 47/50
131/131 [==============================] - 0s 284us/step - loss: 0.2058 - accuracy: 0.9542
Epoch 48/50
131/131 [==============================] - 0s 309us/step - loss: 0.2035 - accuracy: 0.9542
Epoch 49/50
131/131 [==============================] - 0s 312us/step - loss: 0.1964 - accuracy: 0.9542
Epoch 50/50
131/131 [==============================] - 0s 360us/step - loss: 0.1933 - accuracy: 0.9618
131/131 [==============================] - 0s 723us/step
accuracy: 96.95%
```

Fig. 3. Piece of python code used to create, fit, and validate the NN model .



```
In [ ]: from keras.models import Sequential
        from keras.layers import Dense
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        color = sns.color_palette()
        sns.set_style('darkgrid')
        from scipy import stats
        from scipy.stats import norm, skew
        from scipy.special import boxcox1p
        from sklearn.model_selection import cross_val_score,train_test_split,KFold, GridSearchCV
        from sklearn.preprocessing import OneHotEncoder, LabelEncoder,StandardScaler, MinMaxScaler
        from sklearn.metrics import confusion_matrix, classification_report, accuracy_score, precision_score
        from sklearn.pipeline import Pipeline,make_pipeline
        from keras.optimizers import SGD
        from keras.constraints import maxnorm
        from keras.utils import plot_model
        import os
        os.environ["PATH"] += os.pathsep + 'C:/Program Files (x86)/Graphviz2.38/bin/'
        %matplotlib inline

        df = pd.read_csv('data/academic_record.csv')
        df.head()
        df.shape
        data = df.values
        X= data[:,0:70]
        Y = data[:,70:]
        X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.25,random_state=7)

        model=Sequential()
        model.add(Dense(100, input_dim=70, kernel_initializer='uniform',activation='relu'))
        model.add(Dense(3,kernel_initializer='uniform',activation='sigmoid'))
        sgd = SGD(lr=0.01, momentum = 0.8, decay = 0.0, nesterov = False )
        model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
        model.fit(X_train,Y_train, epochs = 50, batch_size = 10, verbose=2)
        _, accuracy = model.evaluate(X_test,Y_test, verbose = 0)
        print('accuracy: %.2f'%(accuracy*100))
        model.predict(X_test)
        plot_model(model,show_shapes =True, to_file = 'student_model.png')
```

IV. CONCLUSION

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 was 96.95% which is more than satisfactory percentage for the 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 city. Also, in the future we expect to apply other data mining methods such as RF, DT, and others for comparison. Moreover, future experiments will add more measurement classification qualities such as Precision, sensitivity, and Recall.

V. ACKNOWLEDGMENT

I appreciate all the efforts of my close friends and my lecturer providing me the advice I needed to accomplish this work.

REFERENCES

[1] P. García-Alcaraz, V. Martínez-Loya, J. L. García-Alcaraz, and C. Sánchez-Ramírez, "The Role of ICT in Educational Innovation," no. Iv, pp. 143–165, 2019, doi: 10.1007/978-3-319-93716-8_7.

[2] N. Saxena, "The Role and Impact of Ict in Improving the Quality of Education," *Int. J. Eng. Sci. Res. Technol.*, vol. 6, no. 3, pp. 501–503, 2017.

[3] S. K. Mohamad and Z. Tasir, "Educational Data Mining: A Review," *Procedia - Soc. Behav. Sci.*, vol. 97, pp. 320–324, 2013, doi: 10.1016/j.sbspro.2013.10.240.

[4] 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, doi: https://doi.org/10.1016/j.sbspro.2014.07.117.

[5] 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.

[6] 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, [Online]. Available: http://www.ijcsmc.com/docs/papers/February2013/V2I2201310.pdf.

[7] 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.

[8] 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, doi: 10.1109/ISET.2015.33.

[9] 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.

[10] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybern. Inf. Technol.*, vol. 13, no. 1, pp. 61–72, 2013, doi: 10.2478/cait-2013-0006.

[11] 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, doi: 10.1080/0144929X.2019.1706637.

[12] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's performance," 2015, doi: 10.1109/AEECT.2015.7360581.

[13] 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, doi: 10.11591/ijeecs.v9.i2.pp447-459.

[14] 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, doi: 10.1007/978-3-030-16657-1_70.

[15] 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, doi: 10.1016/j.procs.2015.12.157.

[16] 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, doi: 10.14257/ijdta.2016.9.8.13.

[17] P. S. Janardhanan, "Project repositories for machine learning with TensorFlow," *Procedia Comput. Sci.*, vol. 171, pp. 188–196, 2020, doi: https://doi.org/10.1016/j.procs.2020.04.020.

[18] L. Hao, S. Liang, J. Ye, and Z. Xu, "TensorD: A tensor decomposition library in TensorFlow," *Neurocomputing*, vol. 318, pp. 196–200, 2018, doi: https://doi.org/10.1016/j.neucom.2018.08.055.

[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, doi: https://doi.org/10.1016/j.asr.2018.11.011.

[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, doi: https://doi.org/10.1016/j.catena.2020.104458.

[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, doi: https://doi.org/10.1016/j.eswa.2019.112875.