5thICITB

Prediction of Student's Major based on Grades and Psychological Test using Artificial Neural Network

Intan Yuniar Purbasari[1], Axvian Bagas Syah Putra[2]
[12]*Department of Informatics, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia*
intanyuniar.if@upnjatim.ac.id

ABSTRACT
New enrolled high school students are typically grouped into majors that will determine the subjects they will take during their three-years study period. Three common majors are Life Sciences, Social Sciences, and Linguistics. Several parameters used to categorize the students are grades of certain subjects (those in the National Final Examination from the previous level) and result from a psychological test taken before enrolling in high schools. This research aims to categorize new students into three common majors based on their previous academic grades and also result from the psychological tests. Each student's preference for a certain major is also put into consideration. The accuracy of the system was 100% for the training set and 79% for the test set.

Keywords: new student grouping, artificial neural network, academic grades, classification

1. INTRODUCTION

General high schools in Indonesia typically categorize their students into several groups. Student's categorization is aimed to escalate student's achievement, help teachers to teach and deliver instruction [4]. In the old times, high school students were grouped into certain majors in their third year or senior year, based on their academic grades in freshmen and sophomore years. Nowadays, the categorization occurs as early as the first year, which means that new students enrolled in high schools already have their majors determined from the beginning. The academic grades used to categorize their major were taken from their previous level of study. A comprehensive evaluation is needed to determine a student's major. Therefore, besides grades, students are also required to take a psychological test as well as being explicitly stated what major they want to take. All those three aspects will then be compiled to get a final decision of the most appropriate major for a particular student.

Artificial Neural Network (ANN) is one of the most popular artificial intelligent learning models and has been extensively used in a wide variety of real-world problems and areas. An ANN consists of a number of artificial neurons that mimics a human neural work of processing and transferring information along with a learning improvement [6][3]. In educational field, ANN has been acknowledged to build a model and predict student's course selection behavior (Kardan, 2013), student's successful study rate (Anggraeny, 2009) (Usman & Adenubi, 2013) (Isljamovic & Suknovic, 2014) (Obsie & Adem, 2018), assessing teacher's performance (Rashid & Ahmad, 2016), even detecting disorientation of learning behavior (Bajaj & Sharma, 2018). An attempt to use other approaches beside ANN was proposed by

(Superby, 2007) using techniques such as random forests, decision tree, and discriminant analysis. Another research by (Villaseñor, 2017) tried to identify the similarity between academic researchers in different institutions.

The neuron's structure in an ANN is divided into three layers: input, hidden, and output. The input layers are connected with feature values of data samples aggregated through an activation function into the second layer, which is the hidden layer. Each node in the hidden layer serves similar to the input node in the input layer. Another activation function is triggered and aggregates all values multiplied by their weights in the hidden layer and forwarded to the output layer (Rashid & Ahmad, 2016). The whole process is called the feedforward phase. Equation (1) describes the computation applied in input to the hidden layer and hidden layer to the output layer.

$$y_k = f\left(\sum_{a=1}^{N_a} w_{ak} x_a\right) \qquad (1)$$

Function f is the activation function, N_a is the total number of ith connection to the kth neuron, and x_a is the input value of the ath neuron (which is the feature input from input layer and the hidden input from the hidden layer).

The second phase is computing the learning error by using a cost function/loss function/error function which measure the difference between the target and the output of the network as calculated using Equation (2).

$$E = \frac{1}{2}\left(\sum_{a}^{N_a} (y_a - o_a)^2\right) \qquad (2)$$

The E variable measure the difference or error between y_a (the target value for neuron a) and o_a (the output value from neuron a).

The third phase is the backpropagation process to nodes in the previous layers to update weights responsible for the error value. The backpropagation can be implemented with numerous techniques and one of the most common one is Gradient Descent. Thus, updating weights is performed using Equation (3):

$$w_{ak,new} = w_{ak,old} + \eta * E_k * f'(y_k) * x_a \qquad (3)$$

In Eq. (3), $w_{ak,new}$ is the new weight between input a and output k, $w_{ak,old}$ is the old weight, η is the learning rate, E_k is the error in output k, y_k is the sum of input multiplied by weights as computed in Eq. (1), and x_a is the value of input a.

2. MATERIALS AND METHODS

This section provides the explanation of data source collection and methods to develop an ANN model and to evaluate the performance of the model on the designated test data set.

2.1 Data Set

Sample data were collected from 274 freshmen of a high school. Each data consists of 11 (eleven) attributes plus 1 (one) attribute for data class (major). The class data is categorized into 3 (three) majors: Life Sciences, Social Sciences, and Linguistics. The eleven attributes represent grades from 5 (five) subjects: Life Sciences, Social Sciences, Maths, English, and Indonesian and taken from 2 (two) different sources: Final Examination from Junior High School (excluding Social Sciences) and from rapport grades (including Social Sciences). Besides academic grades, the dataset also contains class results from psychological test and preference of each student on his or her choice of major.

All grades are numerical data in the range of 0 to 100 while psychological test result, preference, and class data are categorical data transformed to numeric (Life Sciences=1, Social Sciences=2, and Linguistics=3). Table 1 presents some data samples taken from the dataset consists of 5 (five) data. Column 1 to 9 contain grade for particular subjects, column 10 contains the student's preference on a major, column 11 is the result from the student's psychological test, and column 12 is the class.

Table 1: A sample of students' grade dataset.

LS1	LS2	M1	M2	E1	E2	I1	I2	SS	R	P	Class
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
96.5	97.5	96	77.5	95.5	86	88.5	86	91.5	1	1	1
88.5	82.5	85.5	90	86.5	74	87	88	89.5	1	1	1
91	75	91.5	80	91	64	92	84	89	1	2	1
83	77.5	84	65	82	74	89	84	83	2	1	2
81.5	72.5	82.5	60	89	68	83	92	83.5	3	3	3

2.1 Methods

The process to develop an ANN model consists of steps of data pre-processing, data training, data testing, and performance evaluating. The ANN model was implemented using Python programming language with scikit-learn module.

2.2.1 Data Pre-processing

Input dataset were loaded and divided into training and test data (with 70% for training and 30% for testing). Then a z-score normalization was applied to each training and testing data to ensure the data is normally distributed and help the ANN to converge faster.

2.2.2 Model Generating

The next step was to generate an ANN model using MLPClassifier class. ANN model has many parameters and hyperparameters that can be tuned making its implementation may have various results and finding the most optimal solution might be a challenge. In this experiment, the default implementation of MLPClassifier was applied as the following setting with some exceptions in the parameters for maximum iteration, max_iter=500, initial learning rate, learning_rate_init=0.01, and making use of previous call of fit function when repeating the training processes, warm_start=True:

```
activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9, beta_2=0.999,
early_stopping=False, epsilon=1e-08, hidden_layer_sizes=(100,),
learning_rate='constant', learning_rate_init=0.01, max_iter=500, momentum=0.9,
n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5,
random_state=None, shuffle=True, solver='adam', tol=0.0001,
validation_fraction=0.1, verbose=False, warm_start=True
```

2.2.3 Model Training

The generated model was then trained on the 70% normalized dataset. The error values for training dataset were recorded to be plotted to see how well the model has learned.

2.2.4 Model Testing and Evaluation

The trained model was tested on the 30% normalized dataset. A confusion matrix of the model's prediction was constructed to compute the precision, recall, and f-measure result and the scores on how well the model learned from training data and how well it performed on testing data is presented.

3 RESULTS AND DISCUSSION

The dataset has empty values for several cells, particularly in the 'R' column of Table 1. The column was supposed to represent the student's choice or interest on one of the three available majors (Life Sciences=1, Social Sciences=2, and Linguistics=3). However, this column having some empty values indicating that some students did not provide or have any preference of majors. For these kinds of cells, they are given the default value of 0 during pre-processing step.

Training dataset was used to generate a model by setting some hyperparameters mentioned in section 2.2.2. Training accuracy was able to reach an average of 100% and for test accuracy was about 79.5% in average. The loss function for training set is presented in Figure 1 and

the last value of error is 0.006886478337553628, while Figure 2 presents visualization of the confusion matrix.

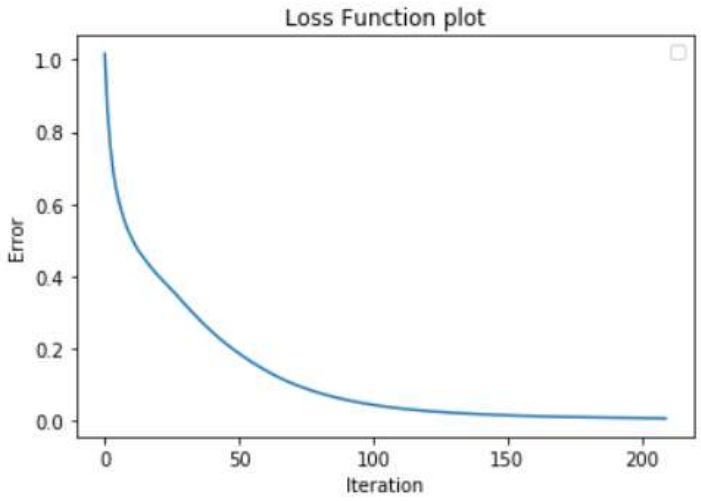

Figure 1: Loss function plotting for training set.

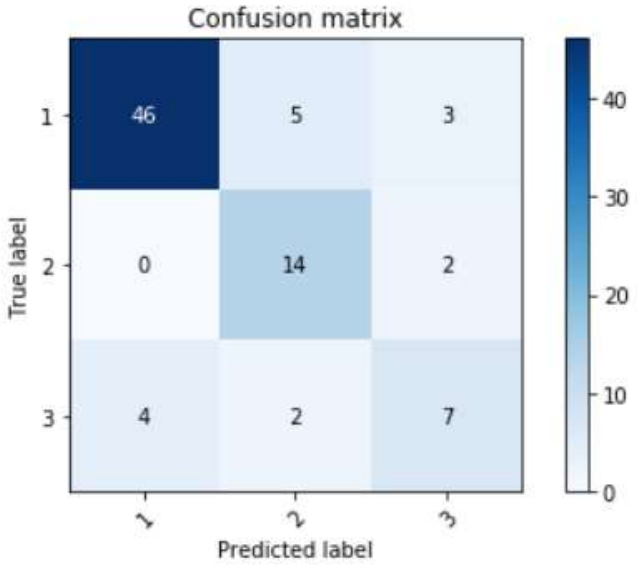

Figure 2: Confusion matrix for test set.

From Fig. 1, it can be seen that the error rate converged quite quickly since the beginning of iteration and gradually reached a stable convergence at the end of the iteration. It showed that the model has learned successfully.

In Fig. 2, class 3 has a non-dominant result for True Positive and has a quite high percentage of False Negative. This might happen because there are not many supporting data for class 3

(most students are categorized as class 1 since they have good grades in Life Sciences and Maths) and thus the ANN was unable to learn optimally to categorize class 3.

Table 2: Values for precision, recall, and f1-score from test set.

Class	precision	recall	f1-score	support
1	0.94	0.85	0.89	54
2	0.58	0.94	0.71	16
3	0.62	0.38	0.48	13
micro avg	0.80	0.80	0.80	83
macro avg	0.71	0.72	0.69	83
weighted avg	0.82	0.80	0.79	83

Table 2 explains the precision, recall, and f1-score for each class from Fig. 2. The overall test set score was 79% which combines the performances of precision and recall for each class, while the overall training set score was 100%. This high score result for training set score shows that the ANN classifier has performed very well to learn from training data while still avoiding overfitting. The learned model still has a good generalization to classify test data and achieve a good score.

Another interesting point to note is in this experiment all dataset's non-class attributes have equal contribution to the weight of the network despite not all columns represent similar values. Nine columns represent academic grades of several subjects while two columns represent preference of majors. With their dominant numbers, the academic grades columns might have bigger role in deciding the predicted outcome rather than those preferences given in the last two columns. Therefore, this might affect the predicted major made by the ANN which disagree with the true decision from the dataset. In future research, those two values might be given different weights to give them higher priorities in deciding the predicted major outcome.

4 CONCLUSIONS

This research has successfully made a classification of new students into three classes: Life Sciences, Social Sciences, and Linguistics using academic grades from the previous level of study and results from the psychological test and interest preference of each student. The average classification score is 100% for training data and 79% for testing data.

ACKNOWLEDGEMENTS

The authors would like to thank the Department of Informatics, Faculty of Computer Science Universitas Pembangunan Nasional "Veteran" Jawa Timur for its support for the publishing of this article in the conference.

REFERENCES

[1] Anggraeny, F. T., 2009. *Prediction of Student's Academic Achievement using Artificial Neural Network (Prediksi Prestasi Akademik Mahasiswa dengan Metode Jaringan Syaraf Tiruan).* Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, s.n.

[2] Bajaj, R. & Sharma, V., 2018. *Smart Education with artificial intelligence based determination of learning styles.* India, Science Direct.

[3] Hu, B., 2017. *Teaching Quality Evaluation Research Based on Neural Network for University Physical Education.* Changsha, China, IEEE, pp. 290-293.

[4] Imron, A., 2012. *Management of School-based Students (Manajemen Peserta Didik Berbasis Sekolah).* Malang: Universitas Negeri Malang.

[5] Isljamovic, S. & Suknovic, M., 2014. *Predicting Students' Academic Performance using Artificial Neural Network: A Case Study from Faculty of Organizational Sciences.* Konya, Turkey, ISRES Publishing.

[6] Kardan, A. A. a. S. H. a. G. S. S. a. S. M. R. F., 2013. Prediction of Student Course Selection in Online Higher Education Institutes Using Neural Network. *Comput. Educ.,* 65(July), pp. 1-11.

[7] Obsie, E. Y. & Adem, S. A., 2018. Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications,* 180(40), pp. 39-47.

[8] Rashid, T. A. & Ahmad, H. A., 2016. Lecturer performance system using neural network with Particle Swarm Optimization. *Computer Application in Engineering Education,* 24(4), pp. 629-638.

[9] Superby, J. V. a. N. M. a. J., 2007. Predicting Academic Performance by Data Mining Methods. *Education Economics,* 15(4), pp. 405-419.

[10] Usman, O. L. & Adenubi, A. O., 2013. Artificial Neural Network (ANN) Model for Predicting Students' Academic Performance. *Journal of Science and Information Technology,* 1(2), pp. 23-37.

[11] Villaseñor, E. A.-J. R. &. C.-C. H., 2017. Multiparametric characterization of scientometric performance profiles assisted by neural networks: a study of Mexican higher education institutions. *Scientometrics,* 110(January), pp. 77-104.