Jurnal Komtika (Komputasi dan Informatika),
Vol. 5 No. 2 | Nov 2021

Prediction of University Student Performance Based on Tracer Study Dataset Using Artificial Neural Network

Zahrina Aulia Adriani[1*], Irma Palupi[2]
[1,2]School of Computing, Telkom University, Bandung
*email: zahrinaadriani@student.telkomuniversity.ac.id

DOI: https://doi.org/10.31603/komtika.v5i2.5901

Received: 01-10- 2021, Revised:22-10-2021, Accepted: 25-10-2021

ABSTRACT

In order to increase student performance, several universities use machine learning to analyze and evaluate their data so that it enables to improve the quality of education in the university. To get a new insight from the tracer study dataset as the relevance between university performance and student capability with business and industries work, the author will develop a model to predict student performance based on the tracer study dataset using Artificial Neural Network (ANN). For obtaining attributes that correspond to labels, Phi Coefficient Correlation will be used to select the attributes with high correlation as Feature Selection. The author is also performing the oversampling method using Synthetic Minority Oversampling Technique (SMOTE) because this dataset is imbalanced and evaluates the model using K-Fold Cross-Validation. According to K-Fold Cross Validation, the result shows that K = 3 has a low standard deviation of evaluation score and it's the best candidate of K to split the dataset. The average standard deviation is 0.038 for all score evaluations (Accuracy, Precision, Recall, and F-1 Score). After applied SMOTE to treating the imbalanced dataset with the data splitting 65 training data and 35 testing data, the accuracy value increases by 10% from 0.77 to 0.87.

Keywords: *Artificial Neural Network, Imbalanced Dataset, K-Fold Cross-Validation, Student Performance, Tracer Study.*

INTRODUCTION

In order to increase student performance in university, some effort can be made to improve the quality of education. One of the efforts that University of Nigerian made for student entrance selection is that they used evaluation results from previous education attached to the registration process to predict students' academic performance using the Artificial Neural Network (ANN) to avoid candidates manipulating the system against the certification value of General Certificate Examination (GCE) [1]. Katsina State Institute of Technology and Management (KSITM) also conducted research on first-semester students by predicting the grades that a student will get in the following semester based on performance and activities carried out during the first semester using Artificial Neural Network (ANN) [2].

Several researchers also have been conducted regarding student performance prediction using machine learning with various datasets. Research [11] used the dataset of 391 students from matriculation classes and 505 students from diploma classes from Information Management System (SIMS) 's students in UiTM. The dataset's attributes are grades of student subjects such as Digital Systems, Signal and Systems 2, Mathematics 2, Materials, CGPA. Other than using student's grades, the attributes of the dataset can use other factors such as living area condition, social media interaction, extracellular activity, study time, family education, and status of drug addiction to predict yearly student performance in

Bangabandhu Sheikh Mujibur Rahman Science and Technology University (BSMRSTU) [12].

Based on those researches, we know that those university data can be used as a measurement to improve the performance of the university by evaluating their student data. One of advantage evaluating student performance is the information about student capability during learning can tell the university if they need to make the change for their program so it can support students in the work field. In this research, the author will use a tracer study dataset to develop a student performance prediction using Artificial Neural Network. The purpose of developing this model is to ensure whether tracer study dataset from the university can be used to evaluate student performance like other datasets from researches that have been mentioned before. Indicators that will be used to measure student performance during lectures are waiting time to get the first job after graduating from the university. Based on collected data from the university, there are 1002 Alumni from 2015 - 2019 that graduated from university [14]. Tracer study dataset contains of the evaluation about student activity experience, university contribution and also the condition before the student get their first job. From the content of tracer dataset, there is interrelatedness for evaluating student performance. This research use K-Fold Cross Validation to find the best data splitting ratio and SMOTE for treating imbalanced dataset in tracer study dataset. Using Artificial Neural Network as the learning machine is based on Research [3]. Research [3] uses several machine learning techniques such as Artificial Neural Network (ANN), Decision Tree (DT), Logistic Regression, and Naïve Bayes to predict student performance. After those models being evaluated with classification and ROC index, ANN has the highest accuracy that is 77.04.

METHOD

This research is generally develop using Python language and use several libraries such as Pandas library for data preparation, Phik library to search correlation, Sklearn and Keras for develop the model and do K-Fold cross validation, Imblearn for treating imbalanced dataset, and Seaborn to visualize the result. Figure 1 explain the workflow of the system.

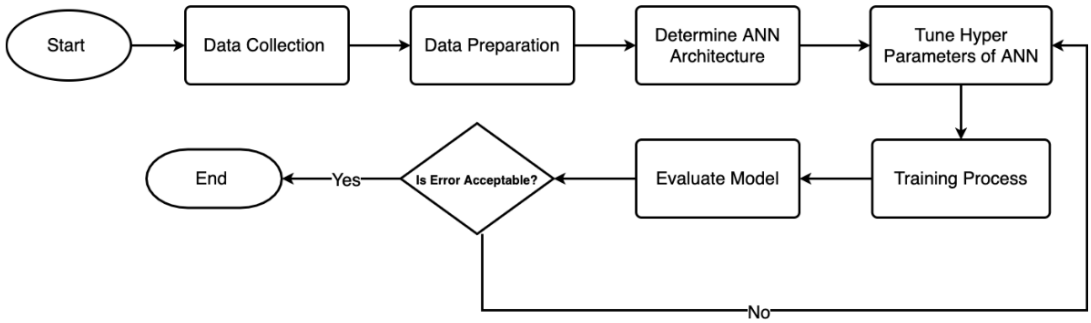

Figure 1. System Workflow

Data Collection

Data used for developing student performance prediction system is the result of survey conducted in university from 2015-2019. A further explanation of data attributes is shown in Table 1.

Table 1. Tracer Study Dataset Description

Attributes	Description	Attributes	Description
bulan_tahun_lulus	Graduation year	bekerja_tekanan*	Working under pressure
nim	Student ID	manajemen_waktu*	Time management
prodi	Department	kerja_mandiri*	Self-employed
nama mahasiswa	Student's name	bekerja_org_lain*	Working with others
jenis_kelamin	Sex	memecahkan_masalah*	Troubleshoot problem
status_kerja	Employment status	negosiasi*	Negotiation
spesifikasi_kerja	Work field Specification	memecahkan_masalah*	Troubleshoot Problem
hbngn_studi_kerja	Study and work field relation	negosiasi*	Negotiation
tngkt_didik_utk_kerja	Education's background	kemampuan_analisa*	Analysis ability
jenis_perusahaan	The type of company	loyalitas_integritas*	Loyalty and Integrity
kompetisi_dlm_bidang*	Inside knowledge	tanggung_jawab*	Responsible
kompetisi_luar_bidang*	Outside knowledge	manajemen_proyek*	Managing project
pengetahuan_umum*	General Knowledge	presentasi_ide*	Presenting idea
keterampilan_internet*	Internet skill	membuat_laporan*	Creating report
keterampilan_komputer*	Computer skil	lama_cari_kerja*	Waiting time first job
berfikir_kritis*	Thinking ability	keterampilan_riset*	Research skills
keterampilan_belajar*	Self-learning	org_beda_budaya*	Working with different cultured people
jml_perusahaan_lamar	Companies that student applied	inisiatif*	Conducting initiatives
jml_perusahaan_respon	Companies respond to student applied	kepemimpinan*	Leadership

***** Rated of Student ability and university contribution

Data Preparation

Data Preparation is a step that must be done to prepare raw data into more informative data so it can be trained using machine learning models. Stages that will be carried out in this process are data cleaning, data transformation and future selection.

In data cleaning, dropping outliers from the dataset and imputing data for missing value using the average for numerical data and mode for categorical data are being done in this step. For data transformation, categorical attributes will be encoded to numerical using ordinal encoding for attributes that has ordinal relationship and one-hot encoding for attributes that have non ordinal association [11]. Besides encoding, the target of the dataset will be grouped into 3 classes, namely students who can work before graduation, students who get a job in 1-6 months after graduation, and students who get a job more than 6 months after graduation. The last process is feature selection. Feature selection method that is Phi Correlation Coefficient. Phi Coefficient (Mean Square Coefficient) is a coefficient used to calculate the correlation of dichotomous (categorical) variables. Phi Coefficient ranges [-1,1]. The formula for ϕ (phi) is often given in terms of a shortcut notation for frequencies in the four cells, called the fourfold table that is represented in Table 2 [4].

Table 2. Azen and Walker Fourfold table Phi Coefficient Notation

n_{11}	n_{12}
n_{21}	n_{22}

Based on the fourfold table, the association between variables can be found with following Equation (1):

$$\phi = \frac{n_{11}n_{12} - n_{21}n_{12}}{\sqrt{(n_{11} + n_{12})(n_{21} + n_{22})(n_{11} + n_{21})(n_{12} + n_{22})}} \quad (1)$$

where n_{11},n_{12},n_{21},n_{22} are non-negative counts of number of observations.

Artificial Neural Network

Artificial Neural Network (ANN) is an information processing paradigm inspired by the way biological nervous systems, such as the brain, process information. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems [5]. Figure 2 shows how Artificial Neural Network works.

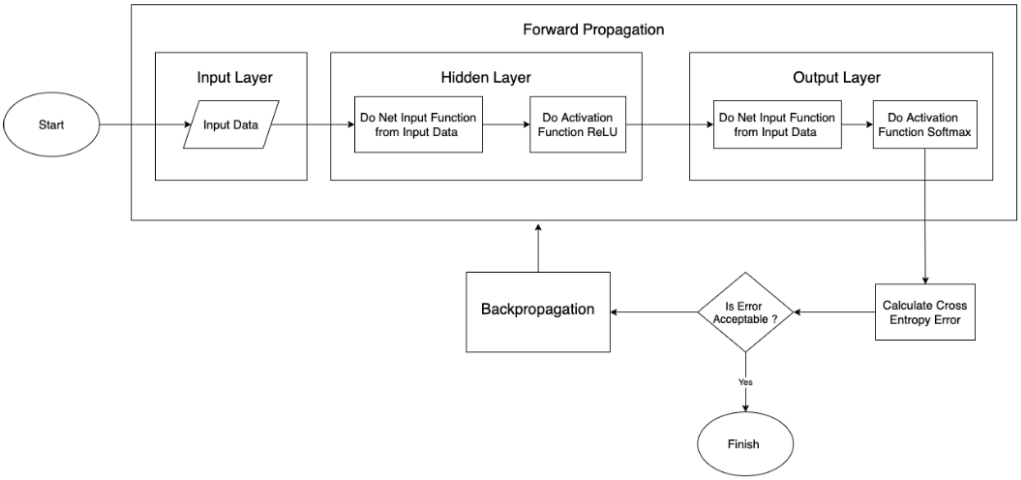

Figure 2. ANN Training Workflow

From Figure 2, The first stage is Forward Propagation. Forward Propagation (Forward Pass) is a calculation process passed by the neural network from input to output layer. The processes are performed on this algorithm are data input, net input function (Equation (2)), activation function using ReLU (Equation (3)) in hidden layer and Softmax (Equation (4)) in output layer.

$$n_k = \sum_{j=1}^{m} w_{kj} x_j \quad (2)$$

where x_1, x_2, …, x_m are the input signal, w_1, w_2, …, w_{km} are the respective synaptic weight of neuron k, and n_k is a linear combination output due to the input signals.

$$f(x) = \max(0, x) \quad (3)$$

$$f(x_i) = \frac{e^{(x_i)}}{\sum_{j=1}^{K} e^{x_j}} \quad (4)$$

where *x* is the input value that will be entered into activation function.

The second stage is calculated loss function. Loss function is a function that measures how well the performance produced by the model in predicting the target. In this research, the authors used Cross Entropy because the target of this dataset is categorical type [6]. Loss Function can be represented with following Equation (5).

$$E = -\sum_d t_d \log a_k^L \quad (5)$$

where t_d is the gold probability of the k^{th} neuron in the output layer (one-hot encoded target label), a_k^L is a result of activation function (Softmax Function) in output layer (predicted label).

The third stage is Backpropagation. This process will be performed using gradient descent that can calculate gradients one layer at a time when the iteration retreats from the last layer to avoid redundant calculation [7]. The rules of gradient descent that can be used for updated weight ΔW is proportionality with $-\frac{\partial E}{\partial W}$ and it can be calculated using chain rule with the following Equation (6).

$$\Delta W_{kj}^l = -\epsilon \frac{\partial E}{\partial a_k^l} \frac{\partial a_k^l}{\partial z_k^l} \frac{\partial z_k^l}{\partial W_{kj}^l} \quad (6)$$

where ϵ is the learning rate, a_k^l is the result activation function in layer l, z_k^l is the weighted sum of activation function in layer l and W_{kj}^l is weight in layer l.

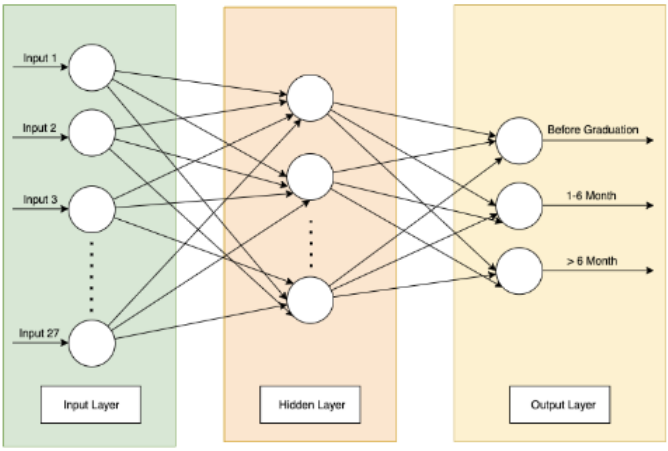

Figure 3. ANN Architecture Design

Figure 3 represents Multi-layer ANN's architecture for this research. It consists of input layer, 1 hidden layer and output layer. The number of neurons contained in the input layer will be the same with number of attributes according to feature selection that is 27. While the number of output layers are 3 neurons and using Softmax activation function according to the desired output of the study. For hidden layer, the number of neurons will be determined using tuning hyper parameters method and it gets 91 neuron. For Activation Function in Hidden layer will use ReLU because Rectified linear units are an excellent default choice of hidden unit [7].

Tuning ANN's Hyperparameter

Hyper parameters that will be tuning to improve model optimization are learning rate, epoch, batch size, patience and number of neurons in hidden layer. Tuning hyper parameters will be done using 2 methods: Random Search and Bayesian Optimization with Gaussian Process.

Training Process

In the training process, there are some scenarios that will be conducted by splitting the dataset into several ratio using K-Fold Cross-Validation with K = [2,6]. This method works by randomly dividing the dataset into k groups, one sub sample of k is fixed to be the validation data and the rest of the groups are the training data [8]. Based on K experiment, the author will search the best K to splitting dataset using average and standard deviation of accuracy in each K in K-Fold Cross-Validation.

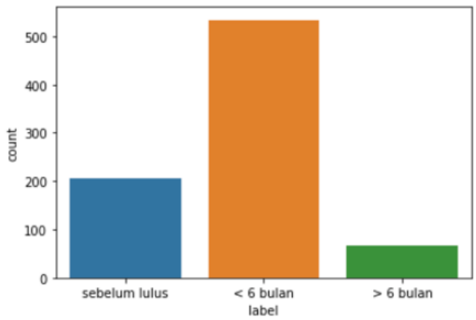

Figure 4. The distribution class in Tracer Study Dataset

From Figure 4, every class in Tracer Study Dataset has a quite large gap between those classes where sebelum_lulus is equal to 189, < 6 bulan is equal to 520, and > 6 bulan is equal to 67. It can be concluded that the classes in Tracer Study Dataset are not approximately equally represented or it can be called as Imbalanced Dataset. For treating this problem, the method that will be used is SMOTE method.

Evaluation Model

The result of model will be evaluated using Confusion Matrix. Confusion Matrix a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known [9].

Table 3. Confusion Matrix

Classification		**Actual Class**	
		Positive (P)	**Negative (N)**
Predicted Class	Positive (P)	TP	FP
	Negative (N)	FN	TN

Based on Table 3, Confusion Matrix returns 4 values from a combination of predicted and actual values such as TP (True Positive), FP (False Positive), FN (False Negative), TN (True Negative). Based on Confusion Matrix, the performance model can be measured by several score such as Accuracy, Precision, Recall and F-1 Score.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F - 1\ Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (10)$$

RESULT AND DISCUSSION

Dataset Information

Based on Table 4 there are missing values in some attributes. All the attributes that contain missing value is being replaced based on the condition that has been mention in Research Method except lama_cari_kerja. lama_cari_kerja is being treated differently with other attributes because lama_cari_kerja will be transformed into label of this dataset. So, to avoid biased evaluation through label, missing values in lama_cari_kerja will be dropped directly. Besides treating missing values, handling outlier in lama_cari_kerja is also important to have better quality of distribution's data. After using Boxplot visualization, the values that detected to be outliers are lama_cari_kerja < -5 and lama_cari_kerja ≥ 10.

Table 4. Tracer Study Dataset Profiling

		Quantity
Number of Feature		39
Number of Rows		930
Type of Variable	Numerical	36
	Categorical	2
Missing Value	nim	64
	hubungan_studi_kerja	199
	spesifikasi_kerja	124
	jumlah_perusahaan_lamar	158
	jumlah_perusahaan_respon	158
	jenis_perusahaan	123
	lama_cari_kerja	155
	tingkat_pendidikan_untuk_kerja	123
Outlier in lama_cari_kerja		27

Statistic Score of Tracer Study Attributes

Based on Table 5 attributes will be used for data training are 27 that have correlation value between label above 0.5 because attributes that have correlation above 0.5 give best accuracy during experiment that carried out in features selection stage.

Table 5. Statistic Score of Tracer Study Attributes

Attributes	**Corr**	**Mean**	**SD**	**Attributes**	**Corr**	**Mean**	**SD**
kompetisi_dalam_bidang	0.845	3.447	1.122	negosiasi	0.841	3.301	1.086
kompetisi_luar_bidang	0.831	3.311	1.080	kemampuan_analisa	0.687	3.652	1.114
pengetahuan_umum	0.663	3.38	1.078	toleransi	0.677	3.501	1.129
keterampilan_intenet	0.629	3.644	1.179	kemampuan_adaptasi	0.679	3.566	1.117
keterampilan_komputer	0.634	3.667	1.160	loyalitas_integritas	0.658	3.543	1.134
berfikir_kritis	0.679	3.570	1.135	org_beda_budaya	0.670	3.569	1.148
keterampilan_riset	0.839	3.449	1.108	kepemimpinan	0.654	3.499	1.126
keterampilan_belajar	0.691	3.589	1.125	tanggung_jawab	0.708	3.503	1.108
bekerja_tekanan	0.645	3.569	1.136	inisiatif	0.659	3.521	1.153
managemen_waktu	0.854	3.498	1.115	manajemen_proyek	0.845	3.446	1.105
kerja_mandiri	0.647	3.555	1.135	presentasi_ide	0.850	3.418	1.094
bekerja_org_lain	0.655	3.614	1.151	membuat_laporan	0.835	3.437	1.104
memcahkan_masalah	0.655	3.559	1.128	minat_belajar	0.851	3.482	1.141

After implementing data profiling and data preparation action through this dataset, now this dataset has 27 attributes and 775 that are ready to be execute for next process. Table

5 also prove that there is the evaluation from study activity during learning affect student performance because those attributes have a high correlation with label targets.

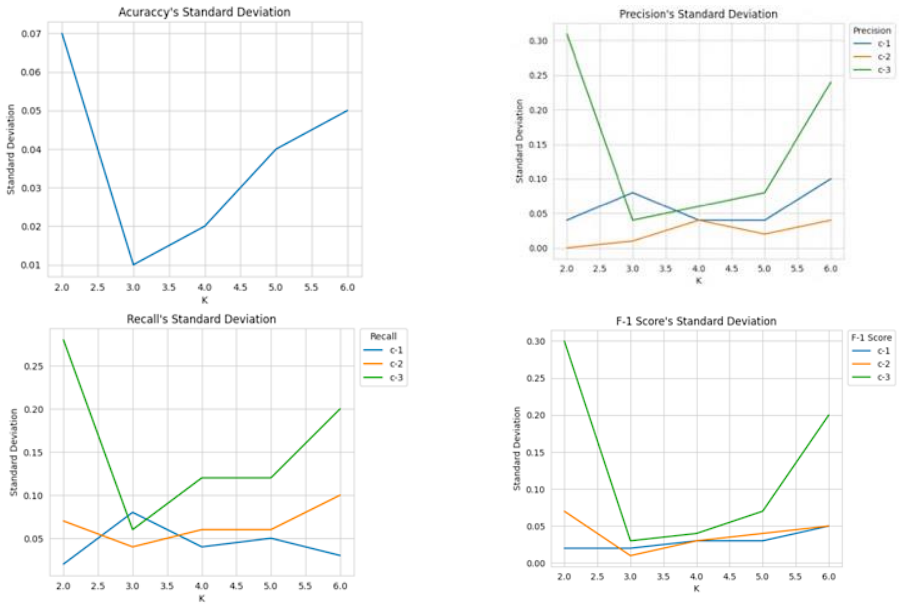

Figure 5. Line graphics of Accuracy, Precision, Recall, and F-1 Score Standard Deviation

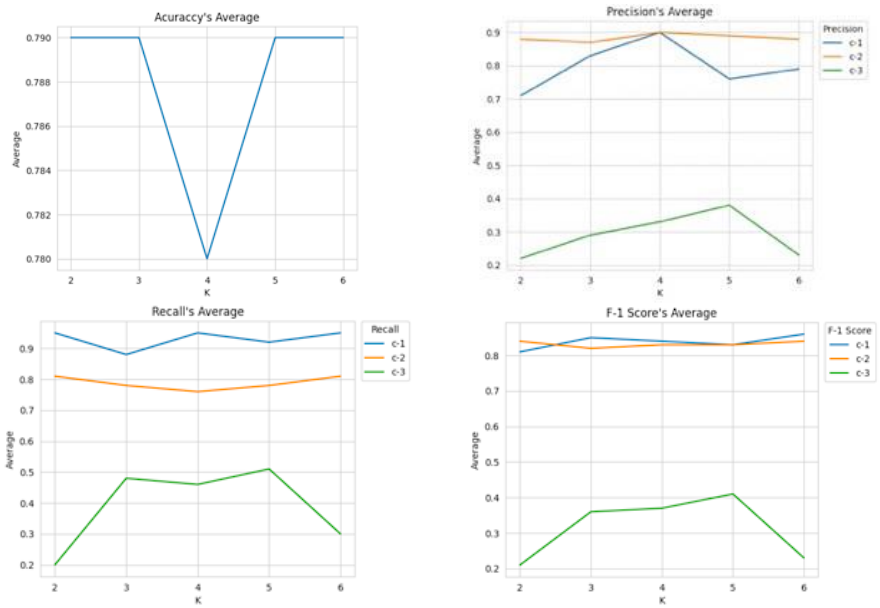

Figure 6. Line graphics of Accuracy, Precision, Recall, and F-1 Score Average

Tuning ANN's Parameter Result

After optimizing ANN Hyper parameters with 2 methods namely Random Search and Bayesian Optimization with Gaussian Process, the result has been found that accuracy from Bayesian Optimization with Gaussian Process is higher which is 0.85 {Batch Size = 16, Epoch = 81, Learning Rate = 0.04, Patience = 11, and Number of Neuron in Hidden Layer = 91} rather than Random search that only gets 0.83.

Cross-Validation Experiment

Based on Figure 5 it can be seen that k = {3,4,5} have small standard deviation of evaluation score from each class. The standard deviation of score in those k is below 0.1 except recall score from class 2 that is 0.12 in k = 4 and k = 5. The accumulation of standard evaluation score from k = {3,4,5} will be average to find which k has a minimum standard deviation evaluation score the will be seen from the average of standard deviation. The minimum of standard deviation's average will be a candidate to be the suitable k for Cross-Validation. The supporting argument underlies that statement is that standard deviation values can describe the distribution of data. If the data has a high standard deviation value, then the data value is widely spread [8].

In order to get a model that has good performance, a consistent evaluation score is needed in every fold iteration of cross-validation. Suppose the standard deviation of the cumulative evaluation score is slight. In that case, the score generated by the fold in each fold of K-Fold Cross-Validation doesn't have much diversity, and it can be stated that the data sharing does not affect the performance model. After searching for average of standard deviation in k = [3,5], the average of standard deviation of k = 3 is 0.038, k = 4 is 0.048 and k = 5 is 0.058. Therefore, k = 3 will be used to be the best candidate of k to split the dataset.

From Figure 6, it can be noticed that and line graphics of class 0 and class 1 have a distance slightly and intersect. However, compared to class 2, class 0 and class 1 have a very different score from the score produced by class 2 around [0.2,0.4]. This problem is a result of the imbalanced dataset as visualized in Figure 4. The low score evaluation made by class 2 can occur due to the uneven distribution of classes and causes misclassification of minority classes [10].

Imbalance Dataset Experiment

Based on Figure 4, it can be concluded that the tracer study dataset used as data to predict student performance is an imbalanced dataset which causes the model to misclassify the class and make the model's performance decrease, as happened in class 2 in Figure 6. Therefore, modeling will be carried out using the SMOTE method with a data splitting ratio of 65 for training data and 35 for testing data (it's because K = 3 equal to 65:35 data splitting ratio).

Table 6. Evaluation Matrices of Model K=3 and SMOTE

	K=3			**SMOTE**		
Class	0	1	2	0	1	2
Accuracy		0.78			0.87	
Precision	0.83	0.88	0.29	0.88	0.90	0.50
Recall	0.88	0.78	0.40	0.90	0.91	0.45
F-1 Score	0.85	0.83	0.36	0.89	0.91	0.47

This model will be compared with the model's results from a K-fold cross which is equivalent to data splitting from the previous model, with K = 3. Table 6 is the results of the evaluation scores of the model using K-Fold Cross Validation with K = 3 and the model using the SMOTE method with the data splitting ratio of 65:35. After using SMOTE, Precision, Recall, and F-1 Score have increased to 0.5 in Precision, 0.45 in Recall, and 0.47 in F-1 score at class 2, which indicates a minority class according to Figure 4. In addition to solving the

misclassification problem, the SMOTE method also increases the accuracy of the model in classifying data until the accuracy value increases by 10% from the previous accuracy to 0.87.

CONCLUSION

Based on the results and discussions, it can be concluded that Tracer Study Dataset has the potential to be used as a tool in evaluating and measuring student performance in predicting time that will be needed to get first job after graduating from university. Performing data processing, parameter tuning, and treating imbalanced datasets on the built ANN model is very influential in improving model performance in predicting student performance using the Tracer Study Dataset. By doing data profiling, author can indicate missing values and outliers, select 27 attributes using the Phi Coefficient Correlation that have correlation > 0.5 with the label and find that the Tracer Study Dataset is an imbalanced dataset due to the unbalanced distribution of data to classes. Optimization tuning parameters using Bayesian Optimization with Gaussian Process also helps determine the hyperparameter of the ANN model until it reaches 0.83 accuracy during the optimization process. From the results of the K-Fold Cross-validation, author can also determine the best K for the model using the Tracer Study dataset that is 3 (65% training data and 35% testing data). Because this dataset is an imbalance dataset, the oversampling method using SMOTE succeeded in increasing the model's performance in predicting the minority class so that the evaluation score increased especially in class 2 and the model accuracy increased by 10%. For future works that can be done for this research is to use tracer study dataset with more variant features that related with the student performance.

REFERENCES

[1] V. Oladokun, A. Adebanjo, and O. Charles-Owaba. Predicting students academic performance using artificialneural network: A case study of an engineering course. 2008.

[2] M. A. Umar. Student academic performance prediction using artificial neural networks: A case study.Inter-national Journal of Computer Applications, 975:8887.

[3] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi. Predicting students' performance using machine learningtechniques.JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences, 27(1):194–205,2019.

[4] J. Ekström, "The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate".2011.

[5] R. Dase and D. Pawar. Application of artificial neural network for stock market predictions: A review ofliterature.International Journal of Machine Intelligence, 2(2):14–17, 2010.

[6] I.-T. Lee, "Notes on Backpropagation with Cross Entropy," doug919.github.io. https://doug919.github.io/notes-on-backpropagation-with-cross-entropy/ (accessed Aug. 08, 2021).

[7] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*, 1st ed. 2016.

[8] J. G. Moreno-Torres, J. A. Saez, and F. Herrera, "Study on the Impact of Partition-Induced Dataset Shift on K-Fold Cross-Validation," IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 8, pp. 1304–1312, Aug. 2012.

[9] W. A. Wardana, I. A. Siradjuddin, and A. Muntasa, “Identification of pedestrians attributes based on multi-class multi-label classification using Convolutional Neural Network (CNN),” 2020.

[10] Y. Liu, A. An, and X. Huang, “Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles,” Lecture Notes in Computer Science, pp. 107–118, 2006.

[11] J. Brownlee, “Ordinal and one-hot encodings for categorical data,” Machinelearningmastery.com, 11-Jun-2020. [Online]. Available: https://machinelearningmastery.com/one-hot encoding-for-categorical-data/. [Accessed: 08-Aug-2021].

[12] P. M. Arsad, N. Buniyamin and J. A. Manan, "Prediction of engineering students' academic performance using Artificial Neural Network and Linear Regression: A comparison," 2013 IEEE 5th Conference on Engineering Education (ICEED), 2013, pp. 43-48, doi: 10.1109/ICEED.2013.6908300.

[13] Sikder, Md Fahim, Md Jamal Uddin, and Sajal Halder. "Predicting students yearly performance using neural network: A case study of BSMRSTU." 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV). IEEE, 2016.

[14] “CDC Telkom University,” Telkomuniversity.ac.id. [Online]. Available: https://cdc.telkomuniversity.ac.id/. [Accessed: 08-Aug-2021]