Observation of Imbalance Tracer Study Data for Graduates Employability Prediction in Indonesia

Ferian Fauzi Abdulloh, Majid Rahardi, Afrig Aminuddin, Sharazita Dyah Anggita, Arfan Yoga Aji Nugraha
Computer Science Faculty, University of AMIKOM Yogyakarta
Yogyakarta, Indonesia

***Abstract*—Tracer Study is a mandatory aspect of accreditation assessment in Indonesia. The Indonesian Ministry of Education requires all Indonesia Universities to anually report graduate tracer study reports to the government. Tracer study is also needed by the University in evaluating the success of learning that has been applied to the curriculum. One of the things that need to be evaluated is the level of absorption of graduates into the working industry, so a machine learning model is needed to assist the University Officials in evaluating and understanding the character of its graduates, so that it can help determine curriculum policies. In this research, the researcher focuses on making a reliable machine learning model with a tracer study dataset format that has been determined by the Government of Indonesia. The dataset was obtained from the tracer study of Amikom University. In this study, SVM will be tested with several variants of the algorithm to handle imbalanced data. The study compared SMOTE, SMOTE-ENN, and SMOTE-Tomek combined with SVM to detect the employability of graduates. The test was carried out with K-Fold Cross Validation, with the highest accuracy and precision results produced by SMOTE-ENN SVM model by value of 0.96 and 0.89.**

***Keywords*—Tracer study; support vector machine; synthetic minority oversampling technique; SMOTE; employability**

I. Introduction

A decent University can be seen from the level of absorption of its graduates in working world, thus many universities are trying to improve the quality of their graduates [1], [2]. That is the reason why the Indonesian Ministry of Education requires all Universities to always report the results of tracer study anually for measuring University graduates employability. Tracer study is also a requirement for higher education accreditation set by the National Accreditation Board for Higher Education (BAN-PT) [3], [4].

Currently we live surrounded by data, data circulating around us can be collected and processed to produce new knowledge [5], including tracer study data. These data can be collected and processed to improve the quality of human resources and curriculum that can increase the absorption of university graduates in industries.[1], [6].

One of the machine learning models that have been widely used to meet these needs is classification [7], [8]. Using classification algorithm we can predict whether an alumni has the possibility of being absorbed in a job quickly or not [9].

There are many classification algorithms that are popularly used, one of which is the Support Vector Machine, from previous research the SVM algorithm is very well used to predict the employability of graduates [10], but basically the final result of an algorithm does not only depend on the quality of the algorithm used but also on the quality of the dataset applied to the algorithm, one of the criteria to get a reliable machine learning model is that the dataset must be balanced, to balance the dataset there are 2 methods, namely oversampling and undersampling, one of the oversampling algorithms that can be used is SMOTE, SMOTE itself has several variants, namely SMOTE, SMOTE ENN, and SMOTE Tomek[11], [12].

This study aims to find out the best method for predicting the employability of higher education alumni using the Amikom University tracer study dataset with attributes and formats determined by the Indonesian Ministry of Education which can be accessed on the web http://tracerstudy.kemdikbud.go.id/ frontend/.

II. Literature Review

A. Classification

Classification is a type of machine learning algorithm where the computer will automatically predict the class of a data from the input data given [7]. Several classification algorithms commonly used for tracer studies include Naive Bayes, Neural Network, SVM, Logistic Regression, etc [9], [13], [14]. In previous works, Tracer Study Data in Indonesia was analyzed using those classification algorithms, without using SMOTE or another imbalanced data handler model.

B. Balance Data

Balanced dataset is data in which the comparison of each data in a class is balanced, the data in which each class has a significantly different amount, the dataset is called imbalance. Unbalanced classes are a common problem in machine learning classification where there is a disproportionate ratio in each class. Class imbalances can be found in various fields , moreover in tracer study case. Classes that have more data are often called majority classes and classes that have less data are called minority classes[15]–[17].

C. Support Vector Machine

The Support Vector Machine algorithm is one of the algorithms included in the Supervised Learning category, which means that the data used for machine learning is data that has a previous label[18], [19]. So that in the decision-making process, the machine will categorize the testing data into labels that are in accordance with its characteristics.

Support Vector Machine is one of the machine learning algorithms that can be used for classification, where this algorithm will generate the best hyperplane where this hyperplane will separate the classes in the dataset [20], [21].

$$w.x - b = 0 \tag{1}$$

where:

w = Weight Vector

x = Input Vector

b = Bias

D. SMOTE

SMOTE is one of the algorithms that can be used to balance a dataset, using an oversampling approach, in which this algorithm will generate synthesis data from the minority class so that the minority class has the same amount of data as the majority class [15], [22].This synthetic data is obtained based on the value of k-neighbours from minority data.

$$\Delta(A, B) = W_A W_B \sum_{i=1}^{N} \delta(v1.v2)^r \tag{2}$$

$\Delta(A, B)$: observed distance between A & B
$W_A W_B$: observed weight
N: amounts of predictor variables
r: value of 1 (Manhattan distance) or 2 (Euclidean dist)
$\delta(v1.v2)^r$: the distance between observations A and B for each explanatory variable, with the formula;

$$\delta(v1.v2) = \sum_{i=1}^{n} \left| \frac{c_{1i}}{c_1} - \frac{c_{2i}}{c_2} \right| \tag{3}$$

$\delta(v1.v2)$: the distance between observations A and B which is included in the i variable

c_{1i}: the number of the 1st category which is included in the i-th explanatory variable category

c_{2i}: the number of the 2nd category which is included in the i-th explanatory variable category

c_1: number of category 1

c_2: number of category 2

n: the number of categories in the i-th explanatory variable

k: Constant

In this study, researchers will compare three variants of the SMOTE algorithm, namely, SMOTE, SMOTE ENN and SMOTE Tomek. SMOTE Tomek uses a combination of the SMOTE algorithm which is a balancing algorithm with an oversampling approach combined with ENN and Tomek which is an undersampling algorithm, where ENN and Tomek function to delete synthetic data that has similarities to the majority data so that data balance is obtained where each data class has a clear difference [11], [23].

III. Research Method

The dataset used in this study is data obtained from questionnaires filled out by alumni of Amikom University in 2018. The questionnaires that have been distributed are then filled out by (many) respondents and stored in csv form. The process can be seen in the Fig 1.

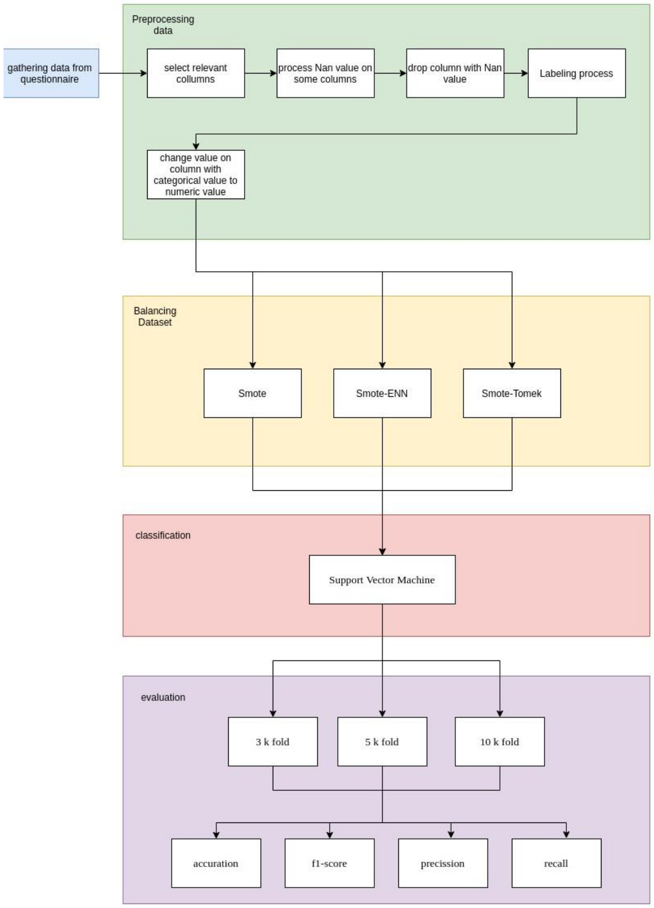

Fig. 1. Research Overview

A. Selection of Attributes and Collection of Survey Results

The first stage of this research is to collect the results of the questionnaire; which later the results from this questionnaire will be presented in csv form so that thereafter it can be processed using a predetermined model. There are 145 collumns consists of their hardskill level after graduate, sex, how long they study in college, when they start to search jobs, and many more, including the label (alumni employability). All of the atributes can be accessed at http://tracerstudy.kemdikbud.go.id/ frontend/.

B. Labeling Data

Data labeling is done by taking each respondent's answer to the question "How long did it take you to get your job after graduation?" In this research, based on that question, labels are divided into three classes. If a student gets a job before graduating from University, then the data will be labeled as “1”. If the student gets a job three months or less after they graduate from University then it will be labeled “2”. If the student takes more than 3 months to get a job get a job after graduation it will be labeled “3”.

C. Data Preprocessing

In this process, preprocessing of data is carried out by converting data labeled string into integer form and also filling empty values in all existing columns with zero values, and deleting values with remaining null data. This have to be done to avoid anomalies in the mathematical modeling.

D. Data Balancing

In practice, classification requires balanced data, balanced data is data where each label has the same amount, if each label has a significantly different amount then the dataset is called imbalanced. Class that has more data is the majority class and the class that has less data is called the minority class [24].

In this study, to overcome the imbalanced data, SMOTE algorithm is used, SMOTE is an algorithm that is useful for balancing the amount of data with an oversampling approach, the SMOTE algorithm will create synthesis data obtained based on the value of k-neighbours from minority data [25].

E. Classification

After the data balancing process, the classification process is carried out with the Support Vector Machine algorithm.

F. Testing

The model testing process uses the K-Fold Cross Validation algorithm with Folds determined to be 3, 5, and 10 Folds. This is done so that the test is more valid and vary [26].

IV. Results

In this study, we will classify the normalized tracer study dataset. After collecting and normalizing the dataset , the dataset will be divided into three classes based on when the alumni got a job, the first class will contain data on alumni who got a job before graduating, less than or three months after graduation, and more than three months after graduating. Fig. 2 showed us the amount of data that has imbalance class. Fig. 3 showed that the amount of dataset significantly altered in every observation using different types of SMOTE.

There are three models of balancing algorithm that will be compared, those are SMOTE, SMOTE ENN and SMOTE Tomek algorithms when applied to the support vector machine classification algorithm. The best model will be calculated based on the average value of f1, accuracy, precision, and recall.

The SMOTE algorithm is a data balancing algorithm with an oversampling approach where the number of minority classes will be increased to balance the majority class. Fig. 4-6 show the dataset after being applied to SMOTE, SMOTE ENN and SMOTE Tomek algorithms

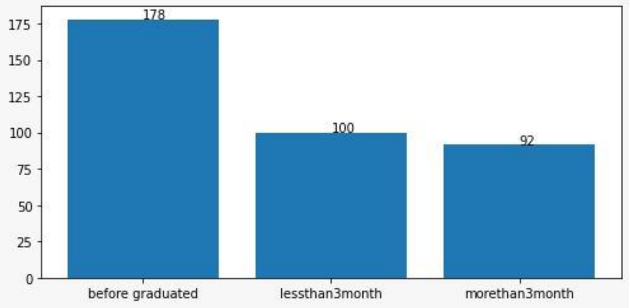

Fig. 2. Dataset before Balancing

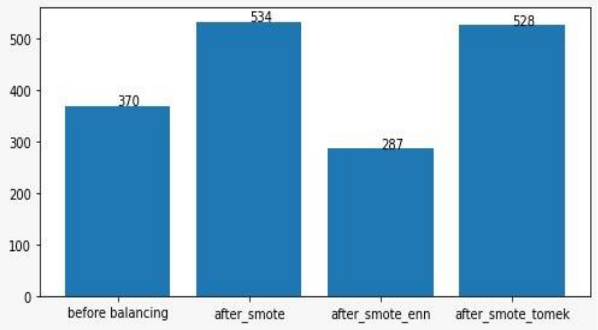

Fig. 3. All Dataset Amounts before and after Balancing

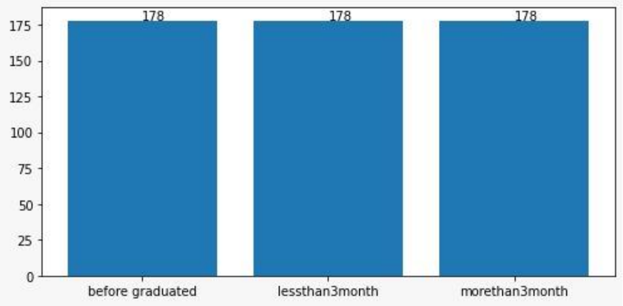

Fig. 4. Dataset after SMOTE

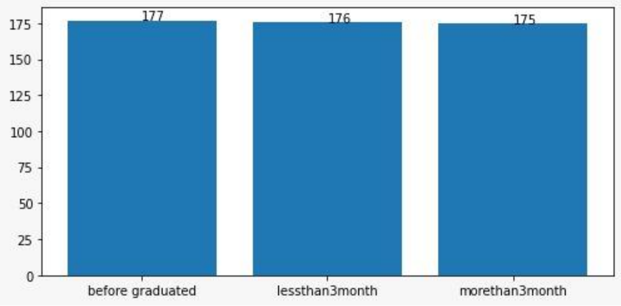

Fig. 5. Dataset after SMOTE-Tomek

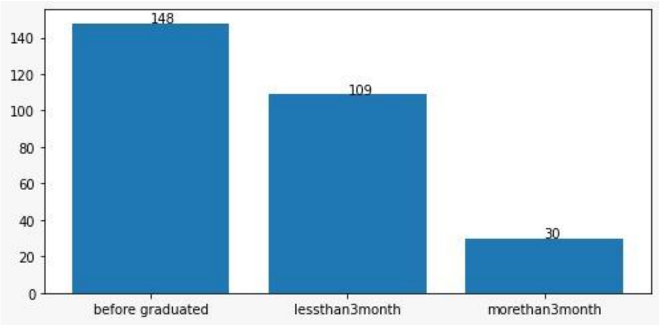

Fig. 6. Dataset after SMOTE-ENN

After the dataset are being processed by SMOTE and SMOTE-TOMEK algorithms, it produces classes that have balanced amount of data. But it did not happen in the SMOTE ENN algorithm, SMOTE ENN created a more normal dataset, this is because when data has an absolute balance, sometimes it may result in overfitting.[12]

Furthermore, after getting the data that we have balanced, the data will be applied to the Support vector machine classification algorithm and for model level measurements, cross fold validation measurements will be used with 3, 5 and 10 fold values for accuracy, f1 score, recall and precision for each model.

TABLE I. RESULT OF ACCURACY & F1 SCORE 3-FOLD

balancing	accuracy			avg	F1 score			avg
-	0.86	0.77	0.76	0.79	0.84	0.73	0.73	0.76
smote	0.78	0.86	0.96	0.86	0.77	0.86	0.96	0.86
smote-enn	0.98	0.94	0.94	0.95	0.97	0.86	0.86	0.89
smote-Tomek	0.86	0.91	0.95	0.90	0.86	0.91	0.95	0.90

TABLE II. RESULT OF RECALL & PRECISION 3-FOLD

balancing	precision			avg	recall			avg
-	0.86	0.82	0.82	0.83	0.83	0.72	0.70	0.75
smote	0.78	0.88	0.96	0.87	0.78	0.86	0.96	0.86
smote-enn	0.98	0.95	0.95	0.96	0.96	0.82	0.82	0.86
smote-Tomek	0.87	0.92	0.96	0.91	0.86	0.91	0.95	0.90

Shown in Table I and II, the experiment is done by using three fold cross validation to test the f1 score, accuracy, precision, and recall from SVM with SMOTE, SMOTE-TOMEK, and SMOTE_ENN and the results obtained that this research scenario has an average f1 accuracy result. score, precision and recall using SVM alone are 0.79.0.76, 0.83, 0.75 and after data balancing, the f1 score, precision and recall are respectively as follows

Smote : 0.86, 0.86, 0.87, 0.86

smote -enn : 0.95, 0.89 ,0.96, 0.86

Smote-tomek : 0.90, 0.90, 0.91, 0.90

These results indicate that the Three-Fold SMOTE, SMOTE-Tomek, and SMOTE-ENN validations are proven to be able to increase the accuracy value of SVM itself, with the highest average value generated by SVM Smote-ENN.

TABLE III. RESULT OF ACCURACY 5-FOLD

	balancing	accuration					avg
1	-	0.93	0.80	0.78	0.86	0.85	0.84
2	smote	0.86	0.75	0.89	0.94	0.99	0.88
3	smote-enn	0.93	0.98	0.98	0.96	0.91	0.95
4	smote-Tomek	0.97	0.82	0.91	0.94	0.99	0.92

TABLE IV. RESULT OF F1 SCORE 5-FOLD

	balancing	f1-score					avg
1	-	0.93	0.76	0.74	0.85	0.84	0.88
2	smote	0.86	0.75	0.89	0.94	0.99	0.82
3	smote-enn	0.89	0.96	0.99	0.92	0.79	0.91
4	smote-Tomek	0.97	0.82	0.90	0.94	0.99	0.92

TABLE V. RESULT OF RECALL 5-FOLD

	balancing	recall					avg
1	-	0.91	0.75	0.73	0.84	0.81	0.80
2	smote	0.86	0.75	0.89	0.94	0.99	0.88
3	smote-enn	0.86	0.94	0.98	0.89	0.76	0.88
4	smote-Tomek	0.97	0.82	0.91	0.94	0.99	0.92

TABLE VI. RESULT OF PRECISION 5 FOLD

	balancing	precision					avg
1	-	0.96	0.79	0.84	0.87	0.92	0.87
2	smote	0.86	0.78	0.91	0.95	0.99	0.89
3	smote-enn	0.95	0.99	0.99	0.97	0.93	0.96
4	smote-Tomek	0.97	0.84	0.92	0.95	0.99	0.93

In the test scenario using five cross fold validation that are shwon at Table III, IV, V and VI, the average results of the f1 score accuracy, precision and recall are 0.84, 0.88, 0.87, 0.80 after data balancing the f1 score accuracy, precision and recall values are equal to

Smote: 0.88, 0.82, 0.89, 0.88

Smote -enn: 0.95, 0.91 ,0.96, 0.88

Smote-tomek: 0.92, 0.92, 0.93, 0.92

then it can be seen from the data that the values of accuracy, precision, recall and f1 are close to perfect which indicates an overfitting, this is triggered by the distribution of test data that is less than the previous experiment.

TABLE VII. RESULT OF NO-SMOTE & SMOTE 10-FOLD

Sub set	**Support vector machine**							
	Without balancing data				**smote**			
	acuracy	precission	recall	f1	acuracy	precission	recall	F1
1	0.86	0.91	0.82	0.84	0.89	0.89	0.89	0.89
2	0.95	0.97	0.93	0.95	0.85	0.86	0.85	0.85
3	0.86	0.90	0.83	0.85	0.76	0.79	0.76	0.76
4	0.76	0.75	0.71	0.72	0.81	0.84	0.81	0.82
5	0.70	0.73	0.62	0.62	0.87	0.91	0.87	0.87
6	0.81	0.88	0.76	0.77	0.94	0.95	0.94	0.94
7	0.95	0.95	0.93	0.94	0.98	0.98	0.98	0.98
8	0.78	0.78	0.73	0.75	0.94	0.95	0.94	0.94
9	0.86	0.92	0.83	0.85	1.00	1.00	1.00	1.00
10	0.86	0.92	0.83	0.86	0.98	0.98	0.98	0.98
avg	**0.84**	**0.87**	**0.80**	**0.81**	**0.9**	**0.91**	**0.9**	**0.9**

TABLE VIII. RESULT OF SMOTE-ENN & SMOTE-TOMEK 10-FOLD

Sub set	**Support vector machine**							
	Smote ENN				**Smote tomek**			
	acuracy	**precission**	**recall**	**f1**	**acuracy**	**precission**	**recall**	**F1**
1	0.97	0.97	0.89	0.92	1.00	1.00	1.00	1.00
2	0.93	0.95	0.86	0.89	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	0.81	0.85	0.81	0.80
4	0.97	0.97	0.89	0.92	0.87	0.88	0.87	0.87
5	0.97	0.98	0.97	0.97	0.92	0.93	0.92	0.92
6	1.00	1.00	1.00	1.00	0.94	0.95	0.94	0.94
7	0.97	0.97	0.89	0.92	1.00	1.00	1.00	1.00
8	0.96	0.97	0.89	0.92	0.91	0.91	0.91	0.91
9	0.93	0.95	0.78	0.81	1.00	1.00	1.00	1.00
10	0.89	0.92	0.74	0.77	0.98	0.98	0.98	0.98
avg	**0.95**	**0.96**	**0.89**	**0.91**	**0.94**	**0.94**	**0.94**	**0.94**

Just like the previous two experiments in the 10 cross fold validation experiment that can be read in Table VII and Table VIII, before the application of balancing the data model, the accuracy value was equal to 0.84, f1 was equal to 0.81, precision was equal to 0.87 and recall is 0.8, then after SMOTE being implemented, there was an increase in the accuracy of the f1 score, precision and recall. The four values increase after data balancing is done. The value of f1 score accuracy, precision and recall is equal to getting the average result

Smotes : 0.90, 0.90, 0.90, 0.91

smooth-enn : 96, 91 ,96, 89

Smote-tomek : 0.94, 0.94, 0.94, 0.94

However, in this experiment, it can be seen that there is an overfitting of the SVM model that uses a data balancing algorithm in several folds which is marked by perfect accuracy in all 3 algorithms. This happens because the test data is only 10% of the entire dataset, it can also be seen in the ENN and Tomek algorithms, cases of overfitting occur more than in the smote algorithm, this is due to the significant difference between the classes in the dataset after the application of the enn and tomek algorithms which is getting worse. enlarge the difference in the data in each class.

V. CONCLUSION

In this study, data balancing algorithms smote, and smote tomek can be used to produce balanced data in terms of the balance ratio formula. Both of these algorithms also produce accuracy, f1 score, precision and recall which are quite significant considering the results presented. However, compared to the SMote-ENN algorithm which produces a poor balance ratio value, the smote tomek and smote algorithms have a lower accuracy value of f1 score, precision and recall. Several fold-cross validation were performed to analyze the data, and found that SMOTE-ENN has the best accuracy in general. In 10-Fold Validation Without SMOTE produced 0.84 in accuracy, using SMOTE it produced 0.9 in accuracy, using SMOTE-Tomek it has 0.94 in accuracy point, and the last one SMOTE-ENN has 0.95 in accuracy.

The SMOTE-ENN-SVM algorithm produces a model with better quality, this can be seen from the accuracy score in each experiment which is higher than other algorithms. In the future, because Tracer Study Data that has many collumns and vary type of data, it would be better to perform feature selection algorithms to select the best feature to be analyzed.

ACKNOWLEDGMENT

This research has been fully supported by grants from the AMIKOM University (#RP-1585824970000).

REFERENCES

[1] A. C. Albina and L. P. Sumagaysay, “Employability tracer study of Information Technology Education graduates from a state university in the Philippines,” Social Sciences & Humanities Open, vol. 2, no. 1, p. 100055, 2020, doi: 10.1016/j.ssaho.2020.100055.

[2] A. F. Hasibuan1, S. M. Silaban2, F. Lubis3, and R. R. Prayogo, “Tracer Study Exploration of Medan State University Graduates,” 2020. [Online]. Available: http://bit.ly/traceralumniunimed2021

[3] P. W. Yunanto, A. Idrus, V. M. Santi, and A. S. Hanif, “Tracer study information system for higher education,” IOP Conference Series:

Materials Science and Engineering, vol. 1098, no. 5, p. 052107, Mar. 2021, doi: 10.1088/1757-899x/1098/5/052107.

[4] Shelly Andari, Aditya Chandra Setiawan, Windasari, and Ainur Rifqi, “Educational Management Graduates: A Tracer Study from Universitas Negeri Surabaya, Indonesia,” IJORER : International Journal of Recent Educational Research, vol. 2, no. 6, pp. 671–681, Nov. 2021, doi: 10.46245/ijorer.v2i6.169.

[5] A. Aminuddin, “Android Assets Protection Using RSA and AES Cryptography to Prevent App Piracy,” 2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020, pp. 461–465, Nov. 2020, doi: 10.1109/ICOIACT50329.2020.9331988.

[6] Y. Nugraheni, S. Susilawati, S. Sudrajat, and A. Apriliandi, “Tracer Study Analysis of Vocational Education in Politeknik Negeri Bandung With Exit Cohort as an Approach,” 2018.

[7] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[8] F. Ernawan, A. Aminuddin, D. Nincarean, M. F. A. Razak, and A. Firdaus, “Three Layer Authentications with a Spiral Block Mapping to Prove Authenticity in Medical Images,” International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, 2022, doi: 10.14569/IJACSA.2022.0130425.

[9] D. C. Casuat and D. E. Festijo, “Predicting Student’s Employability using Machine Learning Approach,” in IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) , 2019, vol. 6th.

[10] A. Binti, A. Rahman, L. Tan, and C. K. Lim, “Supervised and Unsupervised Learning in Data Mining for Employment Prediction of Fresh Graduate Students,” Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, pp. 2–12, 2017.

[11] B. Jonathan, P. H. Putra, and Y. Ruldeviyani, “Observation Imbalanced Data Text to Predict Users Selling Products on Female Daily with SMOTE, Tomek, and SMOTE-Tomek,” 2020.

[12] J. Wang, “Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques,” Mathematical Biosciences and Engineering, vol. 19, no. 10, pp. 10407–10423, 2022, doi: 10.3934/mbe.2022487.

[13] A. U. Umar, “Student Academic Performance Prediction using Artificial Neural Networks,” International Journal of Computer Applications, vol. 178, pp. 24–29, 2019.

[14] B. Heriyadi, U. Verawardina, and T. E. Panggabean, “Tracer Study Analysis for the Reconstruction of the Mining Vocational Curriculum in the Era of Industrial Revolution 4.0 Student at Doctoral Program (S3) Vocational Education, Faculty of,” 2021.

[15] A. Fernández, S. García, F. Herrera, and N. v Chawla, “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary,” 2018.

[16] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, “Assessing feature selection method performance with class imbalance data,” Machine Learning with Applications, vol. 6, p. 100170, Dec. 2021, doi: 10.1016/J.MLWA.2021.100170.

[17] M. Khushi et al., “A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data,” IEEE Access, vol. 9, pp. 109960–109975, 2021, doi: 10.1109/ACCESS.2021.3102399.

[18] M. Tripathi, “Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM,” Journal of Artificial Intelligence and Capsule Networks, vol. 3, no. 3, pp. 151–168, Jul. 2021, doi: 10.36548/jaicn.2021.3.001.

[19] M. Rahardi, A. Aminuddin, F. F. Abdulloh, and R. A. Nugroho, “Sentiment Analysis of Covid-19 Vaccination using Support Vector Machine in Indonesia,” IJACSA) International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022, [Online]. Available: https://t.co/h5x41UO3tF

[20] V. K. Chauhan, K. Dahiya, and A. Sharma, “Problem formulations and solvers in linear SVM: a review,” Artificial Intelligence Review, vol. 52, no. 2. Springer Netherlands, pp. 803–855, Aug. 15, 2019. doi: 10.1007/s10462-018-9614-6.

[21] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, “A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting,” Annals of Data Science 2021, pp. 1–26, Jun. 2021, doi: 10.1007/S40745-021-00344-X.

[22] A. Yaqin, M. Rahardi, and F. F. Abdulloh, “Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student’s Graduation in XYZ University,” International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, p. 2022, 2022, doi: 10.14569/IJACSA.2022.0130652.

[23] Z. Xu, D. Shen, T. Nie, and Y. Kou, “A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data,” Journal of Biomedical Informatics, vol. 107, Jul. 2020, doi: 10.1016/j.jbi.2020.103465.

[24] Y. Yan, R. Liu, Z. Ding, X. Du, J. Chen, and Y. Zhang, “A parameter-free cleaning method for SMOTE in imbalanced classification,” IEEE Access, vol. 7, pp. 23537–23548, 2019, doi: 10.1109/ACCESS.2019.2899467.

[25] S. Feng, J. Keung, X. Yu, Y. Xiao, and M. Zhang, “Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction,” Information and Software Technology, vol. 139, p. 106662, Nov. 2021, doi: 10.1016/J.INFSOF.2021.106662.

[26] A. Aminuddin and F. Ernawan, “AuSR2: Image watermarking technique for authentication and self-recovery with image texture preservation,” Computers and Electrical Engineering, vol. 102, p. 108207, Sep. 2022, doi: 10.1016/J.COMPELECENG.2022.108207.