Multiclass Classification With Imbalanced Class And Missing Data

Irfan Pratama [1]
Department of Information System
Universitas Mercu Buana Yogyakarta
Yogyakarta, Indonesia
Irfanp@mercubuana-yogya.ac.id

Putri Taqwa Prasetyaningrum [2]
Faculty of Information Technology
Universitas Mercu Buana Yogyakarta
Yogyakarta, Indonesia
putri@mercubuana-yogya.ac.id

***Abstract*—In any data mining field, the presence of a good shaped data is needed. Yet in the reality, the data condition is far from the expectation as there are possible to have missing values, redundant data, and inconsistent data. There are problems with the dataset to begin with before we overcome the problem of data mining process interpretation. In the raw data level, possible problem such as missing values and data redundancy or inconsistency can be solved by some certain process called preprocessing. On the preprocessing step, the raw dataset is adjusted to the needs of the whole process, one of the adjustments is to handle missing values. Missing values is a certain condition where the expected values of the data are not recorded. The other problems that happen in the real-world dataset especially in categorical data with label or class is the imbalance distribution of the instance for each class. The imbalanced class is a condition where the distribution of the class is skewed or biased. This study emphasizing on the problem solving of missing values and imbalanced class on the dataset. K-NN imputation is a missing value handling method of this study. As for the imbalanced class problem, this study utilizes SMOTE and ADASYN for the comparison. While the dataset will further be tested by various classification methods such as Decision tree, Random Forest, and Stacking. The original dataset produced bad score from the classification process due to the imbalanced data. Then the data undergoing an oversampling process using SMOTE and ADASYN methods in hope that the accuracy will be hugely better. Yet the reality is the accuracy score do not move to the expected number at all with only averaging in 32%-37% of accuracy score in any scheme of process.**

Keywords—misisng values, imbalanced class data, multiclass dataset, oversampling, classification

I. INTRODUCTION

In recent years the digital data grow in an exponential manner. And data is become more valuable assets to anyone who seek a knowledge of what is happening currently based on the data. The educational field also can utilize this kind of action to be more effective and efficient in knowing the students based on the biography of their registers form and their majors. The expected result of the profiling is to boost the effectivity of the marketing department.

Data mining as a field that utilizes data as its source of knowledge become more and more powerful and the data condition that handled by a data mining process also become variative. The condition of the data sometimes demanding ispecial treatment and special mechanism to overcome in order to produce a good data mining result. As written by Witten, Ian et.al[1] and Han, Jiawei et.al [2] that the basic concept of data mining is as follows: prediction, association, classification, and clustering. In any of those data mining field, the presence of a good shaped data is needed. Yet in the reality, the data condition is far from the expectation as there are possible to have missing values[3], redundant data[4], inconsistent data[4], and so on. There are problems with the dataset to begin with before we overcome the problem of data mining process interpretation. In the raw data level, possible problem such as missing values and data redundancy or inconsistency can be solved by some certain process called preprocessing.

Preprocessing is a process before the core data mining process is eventually begin. On the preprocessing step, the raw dataset is adjusted to the needs of the whole process, one of the adjustments is to handle missing values. Missing values is a certain condition where the expected values of the data are not recorded. The cause of such condition is vary based on the situation. In the industrial field, missing values caused by the machine faults or errors which has to be recording the data but it is not happening in the end. In other field, missing values can occurs in the survey filling misconduct, it is when the respondent refused to fill completely or miss an item on the questionnaire[3].

The other problems that happen in the real-world dataset especially in categorical data with label or class is the imbalance distribution of the instance for each class. The imbalanced class is a condition where the distribution of the class is skewed or biased[5]. Imbalanced classification pose a threat where the predictive modeling are actually designed under the assumption that the distribution of each class is equal[5]. With that problem alone, the weak result of classification will be produced.

There are several methods and mechanism to solve those two explained problems. To handle missing values, there are several approaches that utilize basic statistics methods to more advanced machine learning procedure. The basic method to handle missing values is to delete the instance that contain

missing values or ignore the existence of the missing values[3]. The statistical approach of missing values handling method are mean/mode imputation[3]. The advance mechanism of missing values handling method is by utilizing machine learning method such as K-NN[6]. The K-NN method serves the missing values as the predictable class of the data, the setup of how many *k* values are used are affecting the result.

As for the imbalanced class problem, there are two approach to handle imbalanced data namely data level methods and algorithm level methods. Undersampling and oversampling are two common approach of data level methods. Oversampling methods such as SMOTE[7] are widely used in this case of problems. The extension of SMOTE as an oversampling method is ADASYN[8]. Those two oversampling methods are widely used in recent classification research. The algorithm level methods covers a mechanism such as one-class learning, cost-sensitive learning, threshold methods, and so on [9].

This study emphasizing on the problem solving of missing values and imbalanced class on the dataset. Looking at the reference explained before, the missing values handling method is K-NN imputation[6]. As for the imbalanced class problem, this study utilizes SMOTE[7] and ADASYN[8] for the comparison. While the dataset will further be tested by various classification methods such as Decision tree[2], Random Forest[10] and Stacking[11].

II. METHOD

A. Data

The dataset used in this research is the student biography based on their majors from a certain University. There are there are 5 years (2016-2020) dataset compiled from 13 majors. The dataset has many attributes yet not all of them are usable. As for this research there are 5 attributes namely: "Province", "Parent's Job", "Parent's Income(monthly)", "High School", and "Majors" (label). The sample of the dataset can be seen in Table 1.

TABLE I. DATASET SAMPLES

Province	*Parent's Job*	*Parent's income*	*High School*	*Majors*
JAMBI	Swasta	2	SMA	Ilmu Komunikasi
GORONTALO	PNS	1	MA	Psikologi
NUSA TENGGARA TIMUR	Swasta	1	SMA	Pendidikan Bahasa Inggris
SUMATERA UTARA	Lainnya	3	SMA	Manajemen
M A L U K U	Swasta	2	SMA	Akuntansi
JAWA TENGAH	Swasta	4	SMK	Ilmu Komunikasi
KALIMANTAN TENGAH	Swasta	4	SMA	Manajemen
JAWA TENGAH	Swasta	2	MA	Informatika
KALIMANTAN TIMUR	Lainnya	3	SMA	Psikologi
SUMATERA UTARA	PNS	0	SMA	Manajemen
RIAU	Swasta	3	SMA	Psikologi

From Table I can be seen that the *Parent's Income* and *High School* already been simplified and encoded. The High School column is originally written the specific school they went. As for the *Parent's Income*, the original data is the range amount of the monthly earning but then encoded into several numbers as category. For the other attributes, *Province* contain the Province where the come from, *Parent's Job* is the category of the jobs, and *Majors* are where they study at the time in college.

B. Research Design

The following stages is the Research Design:

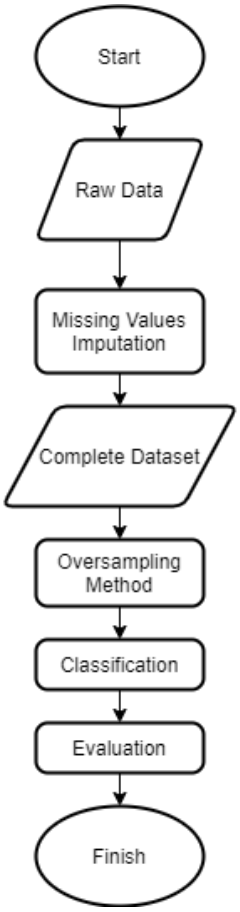

Fig. 1 Research Design

C. Preprocessing

Preprocessing is a step to do a preparation of the dataset before ongoing any data mining processes. The missing values handling and imbalanced class handling are treated in this process. The first process to be done in this stage is missing values handling. As explained before, the dataset contains

missing values in several areas of columns. The sample that contains missing values are shown in Table II.

TABLE II. SAMPLES WITH MISSING VALUES

Province	*Parent's Job*	*Parent's income*	*High School*	*Majors*
SULAWESI SELATAN	Swasta	1	MA	Informatika
DI YOGYAKARTA	Lainnya	1	?	Psikologi
DI YOGYAKARTA	Lainnya	2	?	Teknologi Hasil Pertanian
DI YOGYAKARTA	Swasta	0	SMK	Manajemen
DI YOGYAKARTA	?	?	SMA	Manajemen

As shown in the Table II, the missing values occurs randomly among the dataset. By the count of the missing values, the data set have 3.349 missing values of 10.146 instances. That number cannot be ignored as the amount of possible information loss if the instances with missing values are deleted.

The missing values imputation done using Rapidminer tools by utilizing the *Replace Missing Values* module and then add the K-NN (with $k = 5$) method to the subprocess after the dataset role set with the *Majors* column as label. The sample result of the missing values imputation can be seen in Table III in accordance to the Table II sample preview.

TABLE III. MISSING VALUES HANDLING RESULT SAMPLES

Province	*Parent's Job*	*Parent's income*	*High School*	*Majors*
SULAWESI SELATAN	Swasta	1	MA	Informatika
DI YOGYAKARTA	Lainnya	1	SMA	Psikologi
DI YOGYAKARTA	Lainnya	2	SMK	Teknologi Hasil Pertanian
DI YOGYAKARTA	Swasta	0	SMK	Manajemen
DI YOGYAKARTA	Buruh	0	SMA	Manajemen

As shown in Table III, the missing values are already been replaced with the estimated value using K-NN method. As for this step, the dataset is now completed and have no missing values in it anymore.

As explained before, the dataset has 13 *Majors* which is the label of the dataset which make this dataset is a multiclass dataset. Yet the instances distribution of each *Majors* is skewed and biased. The preview of the data distribution can be seen in Fig. 2.

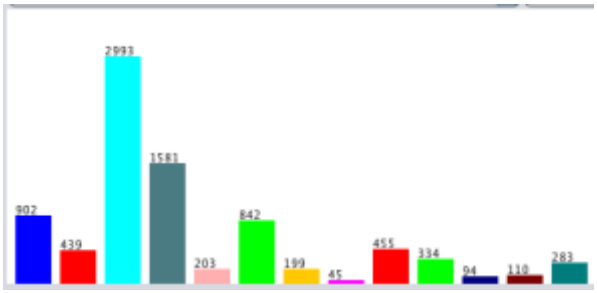

Fig. 2 Class Distribution

Fig. 2 shown the distribution of the class (Majors) of the dataset, if the chart sequenced 1 to 13 from left to right respectively, the third class have the majority number among the classes. So, this condition can be considered as an imbalanced class situation.

For the Imbalanced class handling, the imbalanced data handling library on python are utilized. The completed dataset from previous step imported into python and then treated using SMOTE and ADASYN. Both of the method is synthetizing the dataset into relatively balanced number for each class. The methods generating synthetic dataset for the minority class in consideration of the general characteristic of the class itself. The illustration of how SMOTE and ADASYN oversampling procedure works can be seen in Fig. 3.

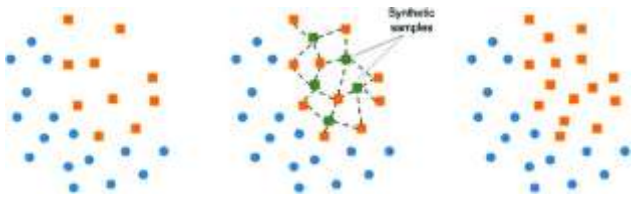

Fig. 3 Oversampling procedure illustration [12]

The result of ADASYN and SMOTE oversampling method can be seen in Fig. 4. And Fig. 5.

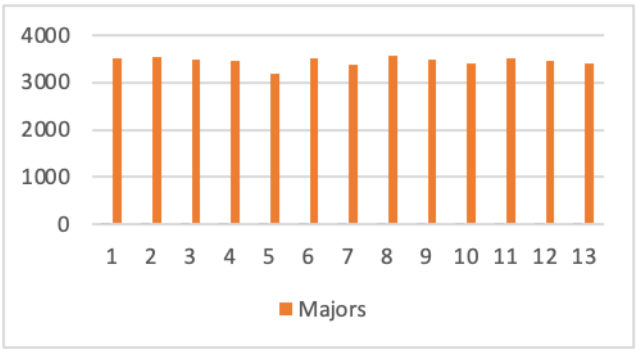

Fig. 4 ADASYN Result

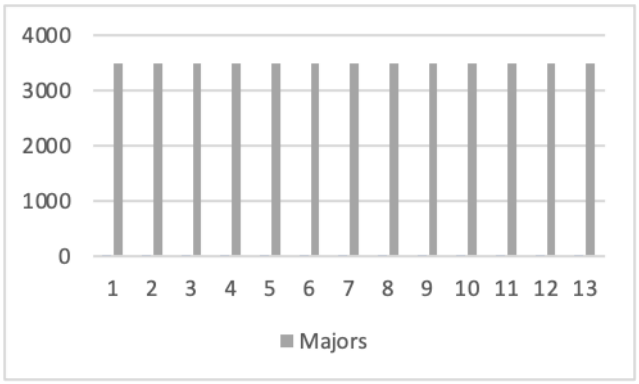

Fig. 5 SMOTE result

As shown in Fig. 4 and Fig. 5, the oversampling methods produced different result. ADASYN produced the relativity in the amount of synthetic data for each minority class, while SMOTE produced the same exact amount of the majority class for the minority class. After this stage, the classification on the balanced dataset will be started.

D. Classification

This stage will be conducted using several classification methods such as Decision Tree, Random Forest, and ensemble technique stacking. All of the classification done in WEKA. The result of each classification will be evaluated and compared with the original data.

Decision Tree is a widely used in data mining field as a solution finder for classification problem. Decision tree alter the large tabular dataset into a tree model to represent the rules of decision[2]. The tree model illustration of decision tree can be seen in Fig. 6.

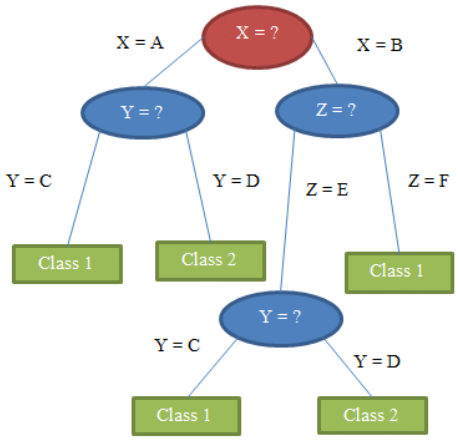

Fig. 6 Decision Tree Illustration

Decision Tree is a supervised learning method, which means the dataset should have a target class attribute or label. The label always been decided beforehand and the Decision tree working on based on that target attribute. Decision Tree consists of the following structures[13]:

Root Node: as it says, "root" means the core of the tree which placed on the top of a decision tree. This node has no incoming branches and has one or more branches. This node usually filled by the strongest attribute related to the class of the dataset.

Internal Node: a branching node that located in between nodes on a decision tree. It has one incoming branch, and has one or more outcoming branch.

Leaf Node: The last node of a decision tree, it has one incoming branch and no outcoming branch. This node is always be a class node or the target attribute.

Random Forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x},\Theta_k), k=1, ...\}$ where the $\{\Theta_k\}$ are identically independent and distributed random vectors and each tree casts a unit vote for most popular class at input **x**[10].

Stacking is an ensemble method that stack individual learners into a single powerful learner. The general principle of stacking is as follows: given *d* different learning algorithms, evaluate each of them on the predictor matrix *X*, given outcome vector *y* in a *k*-fold cross-validation. Save the out-of-fold predictions and combine them to a new data matrix *Z*. *Z* now has *d* columns and the same number of rows as *X*. Then, estimate a weighted scheme for each column of *Z* to combine to a final prediction[14].

E. Evaluation

For each classification process, the measurement of how well the classifier perform is by looking at the accuracy, precision, recall, and F-Measure. The test scheme for this research is 10-fold cross validation using WEKA.

III. Result and Discussion

The raw dataset undergoing a missing value handling stage where the missing data on the dataset are replaced by estimated values that produced by the K-NN method. K-NN can work with both numerical and nominal attribute on the dataset. And after being completed, the dataset then be balanced using the oversampling method to overcome the imbalanced dataset problem.

The dataset has been balanced using SMOTE and ADASYN oversampling method and tested using several classification methods such as, Decision tree, Random Forest and Stacking ensemble method that stacks the Decision Tree and Random Forest into one learner method. The oversampling stage produced a relatively balanced dataset according to Fig. 4 and Fig. 5 with minor differences between SMOTE and ADASYN. SMOTE method produced the exact same instances number for all class which the number of each class are based on the majority class' instances amount. While ADASYN produced different amount of synthetic data among the classes yet the method still recognizes the difference as "relatively balanced" state.

After the data balancing method, the balanced and the original completed dataset will be tested using the classification method to acquire the accuracy result and may provide the unique pattern of each class. The classification result of the original completed dataset using the classification can be seen in Table IV.

TABLE IV. CLASSIFICATION RESULT OF ORIGINAL DATASET

No	Methods	Accuracy	Precision	Recall	F-Measure
1	Decision Tree	0.370	0.283	0.370	0.292
2	Random Forest	0.359	0.288	0.359	0.306
3	Stacking	0.326	0.258	0.326	0.279

Table IV shown that the classification result is bad, the accuracy score is only reach around 32%-37%, with the Decision Tree get the biggest score out of three classification method. The original dataset then oversampled and the testing result of the SMOTE-oversampled dataset can be seen in Table V and the ADASYN-oversampled dataset test result in Table VI.

TABLE V. CLASSIFICATION RESULT OF SMOTE OVERSAMPLED DATASET

No	Methods	Accuracy	Precision	Recall	F-Measure
1	Decision Tree	0.363	0.366	0.363	0.352
2	Random Forest	0.367	0.370	0.367	0.356
3	Stacking	0.346	0.344	0.346	0.344

TABLE VI. CLASSIFICATION RESULT OF ADASYN OVERSAMPLED DATASET

No	Methods	Accuracy	Precision	Recall	F-Measure
1	Decision Tree	0.349	0.341	0.349	0.332
2	Random Forest	0.352	0.347	0.352	0.337
3	Stacking	0.327	0.315	0.327	0.313

As shown in Table V and VI that the both oversampled techniques do not work to fix the accuracy score to be better which only scored a 32%-37% accuracy for any used method. As for SMOTE oversampled dataset, the evaluation score of the classification result is not too far from the original dataset. In terms of precision the SMOTE oversampled dataset produced slightly better than the original dataset. While the ADASYN oversampled dataset shown a decrease in terms of classification accuracy score in any used methods. ADASYN reported to be better than SMOTE in several previous research related to imbalanced class problem in classification process. But the Tables shows that SMOTE have the slight lead in terms of accuracy score than ADASYN.

Normally, the oversampling technique fix the imbalanced dataset into better classification result in the end. But even some circumstances like this situation may happen in reality. The dataset is taken from a real record yet it cannot do much things to the classification's result.

The assumption of this situation is either the datamining methods is just do not fit for the dataset, or it is the dataset that has the bad quality in terms of statistical values such as normality, or correlation, or it is because the nature of the dataset that have multi class in it. In terms of classification method results, stacking technique produced lowest accuracy score while individual learner such as Random Forest and Decision Tree produced better result. Even though, the score is still close to each other and may be have no significant difference.

IV. CONCLUSION

This research is emphasizing on the pattern discovery of the student data taken from the university. Yet in ended up become a technical report of how the result's going on. The original dataset produced bad score from the classification process due to the imbalanced data. Then the data undergoing an oversampling process using SMOTE and ADASYN methods in hope that the accuracy will be hugely better. Yet the reality is the accuracy score do not move to the expected number at all. As per this result shown, the dataset may have a bad quality or it is just not fit for the classification process. As for the classification technique used, the individual learner has slightly better performance than the stacking technique. So, it can be a said that it is a confirmation research on how well the dataset work with the used method to produce result.

For the future works, the dataset may be tried on other field of data mining in order to extract the desired knowledge. Addition of instances or attribute that closely related on the student personal characteristics may increase the odd of good result be produced.

REFERENCES

[1] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[3] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. New York, NY, USA: John Wiley & Sons, Inc., 1986.

[4] M. North, *Data Mining for the Masses*. 2012.

[5] J. Brownlee, "A Gentle Introduction to Imbalanced Classification," 2019. [Online]. Available: https://machinelearningmastery.com/what-is-imbalanced-classification/. [Accessed: 20-Oct-2020].

[6] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, 2012, doi: 10.1016/j.jss.2012.05.073.

[7] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J.*

Artificial Intell. Res., vol. 16, pp. 321–357, 2002, doi: https://doi.org/10.1613/jair.953.

[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *Proc. Int. Jt. Conf. Neural Networks*, no. 3, pp. 1322–1328, 2008, doi: 10.1109/IJCNN.2008.4633969.

[9] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets : A review," *Science (80-.).*, vol. 30, no. 1, pp. 25–36, 2006.

[10] L. Breiman, "Random Forests," *Int. J. Adv. Comput. Sci. Appl.*, 2001.

[11] S. Džeroski and B. Ženko, "Is combining classifiers with stacking better than selecting the best one?," *Mach. Learn.*, vol. 54, no. 3, pp. 255–273, 2004, doi: 10.1023/B:MACH.0000015881.36452.6e.

[12] H. Kuswanto and A. Naufal, "Evaluation of performance of drought prediction in Indonesia based on TRMM and MERRA-2 using machine learning methods," *MethodsX*, vol. 6, no. March, pp. 1238–1251, 2019, doi: 10.1016/j.mex.2019.05.029.

[13] D. B. Ananda and A. Wibisono, "C4.5 DECISION TREE IMPLEMENTATION IN SISTEM INFORMASI ZAKAT (SIZAKAT) TO AUTOMATICALLY DETERMINING THE AMOUNT OF ZAKAT RECEIVED BY MUSTAHIK," *J. Inf. Syst.*, vol. 10, no. 1, 2014.

[14] C. F. Kurz, W. Maier, and C. Rink, "A greedy stacking algorithm for model ensembling and domain weighting," *BMC Res. Notes*, vol. 13, no. 1, p. 70, 2020, doi: 10.1186/s13104-020-4931-7.