Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.9 No.2 July 2024, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Unbalanced Timeliness of Financial Reporting Data Classification Using Random Forest with SMOTE

Erna Hayati[1], Fitri Nurjanah[2], Fajriyah Kurnia Laili[3]

[1,2,3]*Department of Accounting, Universitas Islam Lamongan, Indonesia*

[1]ernahayati@unisla.ac.id (*)

[2]fitrinurjanah@unisla.ac.id, [3]fajriyah5758@gmail.com

Received: 2024-05-27; Accepted: 2024-07-02; Published: 2024-07-18

Abstract— This study aims to apply the Random Forest method with SMOTE to address unbalanced data on company classifications based on the timeliness of financial reports. The data used are the financial statements of manufacturing companies in the Food and Beverage sector on the IDX from 2014 to 2022. The independent variables used are ROA, CR, DAR, and Size. The results showed that the performance of the Random Forest method after being combined with SMOTE increased compared to before SMOTE. Random Forest's best performance is derived from 60% training and 40% testing. Based on MDA and MDG values, it was found that ROA has the highest level of importance, followed by Size and CR variables. In comparison, DAR is the variable with the lowest level of importance. It means that DAR has a low impact on the timeliness of financial reports.

Keywords— Unbalanced; Random Forest; SMOTE; Timeliness of Financial Reporting.

I. INTRODUCTION

One indicator of a good financial statement is that the financial report is submitted on time. The relevance of a financial report presented in time will enhance its ability to influence investor decision-making. Financial statements must be submitted on time to be regulated by law. The study of predicting the timeliness of financial reporting through the variables that influence it becomes very interesting and beneficial to stakeholders.

Random Forest is a method of classification and regression developed from the Classification and Regression Tree (CART) method by applying the method of bagging and random feature selection [1]. Random Forest is one of the ensemble learning methods, so this method has the advantages of high accuracy and the ability to work on large datasets [2]. The Random Forest algorithm has advantages over other deep learning methods in terms of performance. Random forest has the advantages of simpler formulation, ease of application, and less computing time [3]. Random Forest can be applied to classify companies based on their timeliness in the publication of financial statements.

In the case of corporate classification based on the timeliness of financial reports, researchers were faced with imbalanced data, where companies had categories on time more than non-on time. Data imbalance is a crucial issue in the case of classification [4]. On unbalanced data, categories whose minorities are often classified into majority categories [5]. The accuracy of the majority class is higher than that of the minority class on all classification algorithms [6]. The Synthetic Minority Oversampling Technique (SMOTE) is a highly proposed method for addressing unbalanced data issues in classification cases. This method balances minority class data with majority data by synthesizing from minority data [7]. The results of a study conducted by [8], [9], [10], and [11] found that the accuracy of Random Forest methods after using SMOTE was higher than before. [12] stated that Random Forest and AdaBoost with SMOTE produced better sensitivity. In this study, the Random Forest method was combined with SMOTE to address unbalanced data classification data based on the timeliness of financial reports.

II. RESEARCH METHODOLOGY

The data used in the modelling is data from 2014 to 2022. The research focuses on the Food and Beverage manufacturing company listed in BEI, which was selected as one of the sectors that was not very affected by the COVID-19 pandemic. Variable timeliness of financial reporting is data category, i.e., not on time (0) and on time (1). A company is categorized on time if it makes its annual financial statements at the latest at the end of the third month after the date of the annual report. It's regulated in the Bapepam and LK Rules Number X.K.2 [13]. The independent variables that affect the timeliness of financial reporting are defined as four variables: profitability (Return on Asset/ROA), liquidity (Current Ratio/CR), leverage (Debt to Total Asset Ratio /DAR), and size of the company (Ln total asset/Size).

The classification method in this study uses Random Forest combined with SMOTE to address unbalanced data. Important variables in the classification with the method of Random Forest are measured using Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) [14]. The MDA is one measure that indicates how much accuracy will be reduced if free variables are not followed in the model one by one. Meanwhile, the MDG shows the stability of independent variables.

The performance of the random forest method is measured with accuracy, recall, specificity, and area under the ROC curve (AUC). Accuracy is the percentage of true prediction for the entire data. Recall is the proportion of true positive

DOI : https://doi.org/10.25139/inform.v9i2.8327

prediction results for the total positive actual data. Specificity is the proportion of true negative prediction results for the total negative actual data. In comparison, AUC is a measurement of model performance that shows how accurate a model is in classifying positive and negative observations [15]. AUC is generally used to compare the performance of classification methods [16]. Figure 1 illustrates the stages undertaken in the analysis of this research.

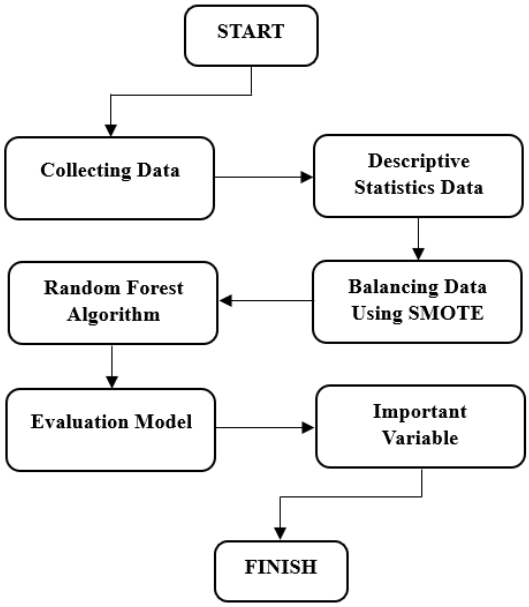

Figure 1. Research Steps

III. Result and Discussion

A. *Descriptive Statistics of Data*

Descriptive Statistics data of each category (on time and not on time) can be seen in Table I. Results of descriptive statistics show that companies that are on time have a higher rate of profitability than companies that aren't on time. The average ROA of a company on time was 0,0836, while a company not on time is – 0,0154. A negative ROA average of a firm not in time indicates that the average company is losing. The company's losses are bad news that prompted its management not to announce it publicly. It is concerned that it will create a negative sentiment against the company's stock price. The average liquidity result (CR) found that the CR value of a company's on-time (2,7219) is smaller than that of a not-on-time company (4,6286). The average CR value of a company that is not on time more than 3 indicates that the company is not optimally using the assets it owns.

In comparison, the company on time has an ideal CR value. This means that the company is still able to meet its current liabilities. The value of the DAR in companies that are on time also has a smaller average (0,43949) than not on time (0,5664). The high value of DAR indicates that companies that aren't on time are at risk of having large debts. This is bad news for shareholders. The average size of the company on time (28,6452) is bigger than the company not on time. (27,9361). Large companies have huge resources to compile and report financial reports on time.

TABLE I
DESCRIPTIVE STATISTICS OF EACH CATEGORY AND VARIABLE

Category	Variable	Min	Max	Mean	Std. Deviation
On-Time	ROA	-0,19	0,60	0,0836	0,10327
	CR	0,35	27,37	2,7219	3,19044
	DAR	0,04	0,94	0,4349	0,18872
	Size	24,49	32,83	28.6452	1,64036
Not on Time	ROA	-2,64	0,61	-0,0154	0,36993
	CR	0,15	98,63	4,6286	14,58322
	DAR	0,04	2,90	0,5664	0,51223
	Size	25,31	30,68	27,9361	1,33168

B. *Balancing Data Using SMOTE*

Based on data from the annual reports of the Food and Beverage sector companies from 2014 to 2022, 218 observations were obtained. Figure 2 shows the distribution of data categories on time and non-on time. Category on time in submitting financial reports of 160 observations and non-on time of 58 data. Category on time is almost three times more than non-on-time data.

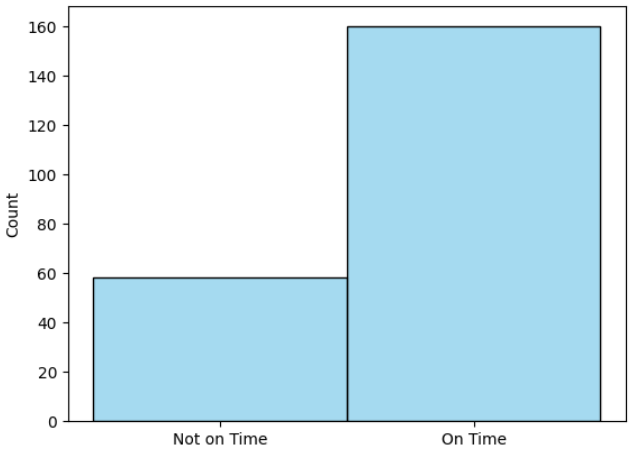

Figure 2. Data Distribution of Each Category

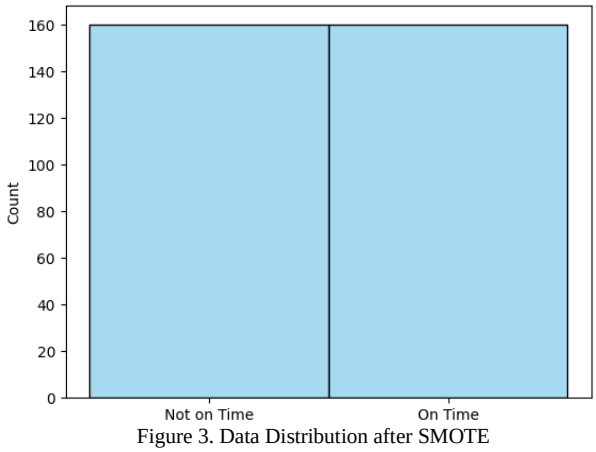

Figure 3. Data Distribution after SMOTE

DOI : https://doi.org/10.25139/inform.v9i2.8327

This study used the SMOTE method to address unbalanced data on the classification of data timeliness of financial reporting. The results of the application of SMOTE, data that are minority, category not on time, increases as much as the data that is majority (on time). The number of not-on-time categories of data that were originally 58 increased to 160. Using the SMOTE method, each major category's observations are balanced. It can be seen in Figure 3.

C. Timeliness of Financial Reports Modeling Using Random Forest

After SMOTE is applied, data is classified using the Random Forest Algorithm. Classification results before and after SMOTE are evaluated using Accuracy, Recall, Specificity, and AUC. Results of Random Forest performance measurement on the timeliness classification of financial reporting before and after using SMOTE can be seen in these Tables.

TABLE II
THE ACCURACY OF THE CLASSIFICATION USING RANDOM FOREST

Training (%)	Testing (%)	Accuracy (%)	
		Before SMOTE	After SMOTE
60	40	68,18	78,91
70	30	65,15	72,92
80	20	68,18	76,56
90	10	68,18	75

Table II shows that the accuracy rate of Random Forest before SMOTE was below 70%. After applying SMOTE, the Random Forest's accuracy rate increased to over 70%. Random Forest's accuracy rate increased from 60% training data composition and 40% testing to 90% training data and 10% testing data. The highest accuracy, 78,91%, is 60% for training and 40% for testing composition.

TABLE III
THE RECALL OF THE CLASSIFICATION USING RANDOM FOREST

Training (%)	Testing (%)	Recall (%)	
		Before SMOTE	After SMOTE
60	40	77,27	83,87
70	30	71,43	76,47
80	20	76,47	89,29
90	10	78,95	92,31

Based on the results shown in Table III, Recall on Random Forest before using SMOTE has a fairly large percentage, above 70%, but still below 80%. After using the SMOTE, Recall increases on all data compositions. Recall has increased from 60% training and 40% testing to 90% training and 10% testing. The highest Recall values occur in the composition of 90% training and 10% testing data, which is 92,31%. This indicates that the Random Forest after SMOTE can correctly predict a company's timeliness in financial reporting at 92,31%.

TABLE IV
THE SPECIFICITY OF THE CLASSIFICATION USING RANDOM FOREST

Training (%)	Testing (%)	Specificity (%)	
		Before SMOTE	After SMOTE
60	40	40,91	74,24
70	30	30,91	68,89
80	20	40	66,67
90	10	0	63,16

Table IV shows that Specificity before SMOTE has the lowest value compared to after SMOTE. The specificity before SMOTE is mostly below 50%. Even the specificity of data composition is 90% training, and 10% testing is only 0%. It shows that the company's forecast results are not timely and are 100% incorrect. Applying SMOTE to Random Forest increases specificity. The highest specificity occurs after the application of SMOTE is 74,24%. The highest specificity comes from composition, 60% training, and 40% testing.

TABLE V
THE AUC OF THE CLASSIFICATION USING RANDOM FOREST

Training (%)	Testing (%)	AUC (%)	
		Before SMOTE	After SMOTE
60	40	59,09	79,06
70	30	50,71	72,68
80	20	58,24	77,98
90	10	39,47	77,73

The result of AUC in Table V is Random Forest before applying SMOTE averages below 60% on all training and testing data compositions. Before SMOTE, AUC decreased from 60% training and 40% testing to 90% training and 10% testing. In all data compositions, the AUC value after SMOTE is greater than before SMOTE. This indicates that the best classification model is after SMOTE. The highest AUC has been obtained by 60% training and 40% testing composition, which is 79,06%.

Results from performance evaluation using Accuracy, Recall, Specificity, and AUC values show that Random Forest performance improves when combined with SMOTE. The Random Forest method's ability to predict minority categories (not on time) increases. The results of this study support the research carried out by [7], [8], [9], [10], [11], and [12]. Based on the four evaluation measures, it can be concluded that the Random Forest model has an advantage over other compositions, with a composition of 60% training and 40% testing after SMOTE.

D. Important Variable in Classification Using Random Forest

Based on the MDA and MDG values generated by the Random Forest model (Figure 4 and Figure 5), it can be seen

DOI : https://doi.org/10.25139/inform.v9i2.8327

that ROA has the highest level of importance in affecting the timeliness of financial reporting. In comparison, the DAR has the lowest level of interest. It means that DAR has little influence on the timeliness of financial reporting. Size and CR have their respective level of importance in the second and third positions.

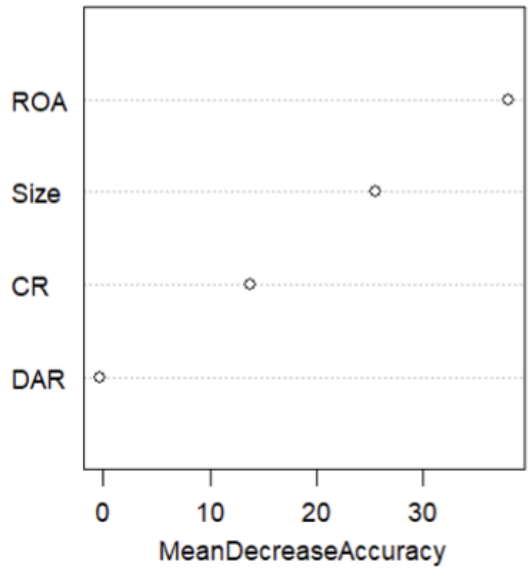

Figure 4. Determination of Importance Variable Using Mean Decrease Accuracy (MDA)

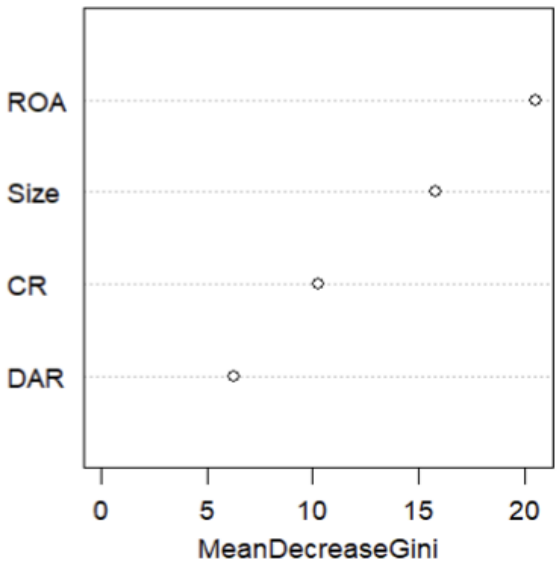

Figure 5. Determination of Importance Variable Using Mean Decrease Gini (MDG)

Profitability (ROA) is the ability of a company to generate profits. If the profits generated by the company are good, then management will be on time for the financial reporting [17]. Corporate profit publication is good news for investors. Investors can estimate the dividend obtained based on published profit information. The results of this study show that ROA has a dominant influence on the timeliness of financial reporting. These results are consistent with the research carried out by [18], [19], and [20].

Size is in the second order, affecting financial reports' timeliness. The results of this study are also in line with [19], [21], and [22]. Companies with large assets tend to have large resources to produce financial reports quickly.

CR is the third sequence of variables that influence the timeliness of financial reporting. It can be concluded that CR does not significantly influence the timeliness of financial reporting. This could happen if companies focused more on paying out current debt and dividing debt into shareholders. So, these management decisions encourage the entity to publish the annual report to shareholders immediately. It's in line with [22] and [23].

This research also found that DAR has the lowest influence on the timeliness of financial reporting. This indicates that the company's total debt does not affect the submission of financial statements on time. It's in line with [20].

IV. Conclusion

The results of the research that has been described provide a lot of information about the impact of the addition of the SMOTE method in addressing unbalanced data in the case of the timeliness of financial reporting. Random Forest with SMOTE yields good performance compared to those without SMOTE. They are based on performance evaluations using four measurements: Accuracy, Precision, Specificity, and AUC, Random Forest with a composition of 60% training and 40% testing after SMOTE performs best. The data composition yielded accuracy values of 78,91%, recall of 83,87%, Specificity of 74,24%, and AUC of 79,06%. While the level of importance of variables is based on the MDA and MDG values, the variables with the highest interest rates are ROA, second size, third CR, and last DAR. These results are expected to help investors and stakeholders determine the factors determining whether a company is on time to publish annual reports.

Acknowledgement

The authors thanked Universitas Islam Lamongan, Faculty of Economics and Business, and Litbangpemas for funding this research.

Reference

[1] A. M. A. Rahim, I. Y. R. Pratiwi and M. A. Fikri, “Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique dan Random Forest Clasifier,” *Indonesia Journal of Computer Science*, vol. 12, pp. 2995-3011, October. 2023.

[2] I. A. Dahlan, “Klasifikasi Cuaca Provinsi DKI Jakarta Menggunakan Algoritma Random Forest dengan Teknik Oversampling,” *Jurnal Teknoinfo*, vol. 16, pp. 87-92, January. 2022.

[3] B. Biswal and P.K. Biswal, "Robust Classification of Neovascularization Using Random Forest Classifier Via Convoluted Vascular Network," *Biomedical Signal Processing and Control*, vol. 66, April. 2021.

[4] L. Cahyaningrum, A. Luthfiarta and M. Rahayu, "Sentiment Analysis on the Impact of MBKM on Student Organizations Using Supervised Learning with Smote to Handle Data Imbalanced," *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, pp. 58-66, January. 2024.

[5] A. A. Khan, O. Chaudhari, and R. Chandra," A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced

DOI : https://doi.org/10.25139/inform.v9i2.8327

Problems: Combination, Implementation and Evaluation," *Expert Systems with Applications*, vol. 244, pp. 1-29. 2024.
[6] H. Suryono, H. Kuswanto and N. Iriawan," Two-Phase Stratified Random Forest for Paddy Growth Phase Classification: A Case of Imbalanced Data," *Sustainability*, vol. 14, pp. 1-13. 2022.
[7] N. N. Sholihah, and A. Hermawan, "Implementation of random Forest and SMOTE Methods for Economics Status Classification in Cirebon City," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, pp. 1387-1397, December. 2023.
[8] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, S. Annamalai, "Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier," *J Med Syst*, vol. 43, July. 2019.
[9] I. Zain, K. Fithiasari, E. O. Permatasari, T. A. Nastiti, N. N. S. Mardyono, R. Pujihasvuty, and S. L. Nasution, "Imbalanced data analysis of adolescent risk behavior of drug abuse using random forest, "in *AECon 2020: Proceedings of The 6th Asia-Pacific Education And Science Conference, AECon 2020, 19-20 December 2020, Purwokerto, Indonesia*, p. 136. European Alliance for Innovation, 2021.
[10] I. Yulianti, A. Rahmawati and T. Mardiana," The Effectiveness Analysis of Random Forest Algorithms with SMOTE Technique in Predicting Lung Cancer Risk," *Jurnal Riset Informatika*, vol. 4, pp. 207-213, March. 2022.
[11] V. M. Putri, M. Masjkur and C. Suhaeni, "Performance of SMOTE in a random forest and Naïve bayes classifier for imbalanced Hepatitis-B vaccination status," *J. Phys.: Conf. Ser.* 1863 012073, 2021.
[12] A.A. Rosita, A. Kurnia, and A. Djuraidah, "Evaluation of ensemble method for multiclass classification on unbalanced data," in *AIP conf. Proc*, vol. 2662, December. 2022.
[13] Bapepam-LK. Decision of The Chairman of Badan Pengawas Pasar Modal dan Lembaga Keuangan Number: KEP-346/BL/2011 about Submission of Financial Reports by an Issuer or Public Company. 2011. Jakarta.
[14] E. Christy and K. Suryowati, "Analisis Klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta Menggunakan Metode Random Forest," *Jurnal Statistika Industri dan Komputasi*, vol. 6, pp. 69-76, January. 2021.
[15] E. C. P. Witjaksana, R. R. Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Random Forest dan Algoritma Artificial Neural Network untuk Klasifikasi Penyakit Diabetes," *e-Proceeding of Engineering*, vol. 8, pp. 9773-9781, October. 2021.
[16] M. S. Maulana, R. Sabarudin and W. Nugraha," Prediksi Ketepatan Waktu Mahasiswa Diploma dengan Komparasi Algoritma Klasifikasi," *Jurnal Sistem dan Teknologi Informasi*, vol.7, pp.202-206, July. 2019.
[17] M. Waris and B. H. Din, "Impact of corporate governance and ownership concentrations on timelines of financial reporting in Pakistan," *Cogent Business & Management*, vol. 10, January. 2023.
[18] T. Herninta, "Faktor-Faktor yang Mempengaruhi Ketepatan Waktu Penyampaian Laporan Auditan Kepada Stakeholder," *ESENSI*: *Jurnal Manajemen Bisnis*, vol. 23, pp. 333–348, September – December. 2020.
[19] D. Wicaksono, "Pengaruh Profitabilitas, Kepemilikan Institusional dan Ukuran Perusahaan terhadap Ketepatan Waktu Penyampaian Laporan Keuangan (Studi Empiris pada perusahaan Sektor Industri Barang Konsumsi yang Terdaftar di Bursa Efek Indonesia Periode 2014-2018)," *KINERJA Jurnal Ekonomi dan Bisnis*, vol. 3, pp. 183 – 197, June. 2021.
[20] C. F. Wibowo and M. H. Saleh, "Pengaruh Profitabilitas, Leverage, dan Ukuran Perusahaan terhadap Ketepatan Waktu Pelaporan Keuangan dengan Kualitas Auditor sebagai Variabel Moderating (Studi Empiris pada perusahaan Sub Sektor Makanan dan Minuman yang terdaftar di Bursa Efek Indonesia Tahun 2017-2019)," Sekolah Tinggi Ilmu Ekonomi Indonesia, 2020.
[21] S. Ginting and S. E. Natasha, "Pengaruh Ukuran Perusahaan, Profitabilitas, dan Solvabilitas, Terhadap Ketepatan Waktu Pelaporan Keuangan pada Perusahaan Keuangan yang Terdaftar di Bursa Efek Indonesia Periode 2015-2017, " *Jurnal Wira Ekonomi Mikroskil*, vol. 11, pp. 1-12, April. 2021.
[22] S.Y.U.P. Putri and I. Wahyudi, "Pengaruh Umur Perusahaan, Ukuran Perusahaan, Likuiditas dan Profitabilitas terhadap Ketepatan Waktu Penyampaian Laporan Keuangan Perusahaan pada Masa Covid-19 (Studi pada Perusahaan Properti yang Terdaftar di BEI Tahun 2019-2020)," *AKSELERASI: Jurnal Ilmiah Nasional*, vol. 4, pp. 25 – 37, March. 2022.
[23] J. Carolina, and V. C. L. Tobing, "Pengaruh Profitabilitas, Likuiditas, Solvabilitas dan Ukuran Perusahaan terhadap Ketepatan Waktu Penyampaian Laporan Keuangan pada perusahaan Manufaktur di BEI," *Jurnal Akuntansi Barelang*, vol. 3, pp. 45-54, June. 2019.

DOI : https://doi.org/10.25139/inform.v9i2.8327