Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

Stock Industry Sector Prediction Based on Financial Reports using Random Forest Method

Kamil Elian Zhafran*, Deni Saepudin

School of Computing, Telkom University, Bandung, Indonesia
Email: zhafrankamil@student.telkomuniversity.ac.id, denisaepudin@telkomuniversity.ac.id
Correspondence Author Email: zhafrankamil@student.telkomuniversity.ac.id
Submitted: **06/08/2024**; Accepted: **10/09/2024**; Published**: 12/09/2024**

Abstract−This study aims to predict the stock industry sector on the Indonesia Stock Exchange (IDX) based on financial reports using the Random Forest method. Implementing this machine learning approach is crucial due to the complexity of financial data, which demands robust and adaptive methods for accurate predictions. The dataset comprises financial data from companies across 10 industrial sectors on the IDX, spanning 2010-2022, and includes 17 features from each financial report. Notably, there is an imbalance in the number of companies per sector, with sector B representing 14.76% and sector G only 1.98%. This imbalance introduces bias in data analysis, thus necessitating the application of the SMOTE oversampling method to address it. The research process involves data cleaning, splitting the data into 80% training and 20% testing sets, applying the SMOTE oversampling technique, and comparing predictions from imbalanced and balanced datasets. The Random Forest method is chosen for its capability to handle complex datasets for industrial sector classification. Evaluation results indicate that without oversampling, the model achieves an accuracy of 73.57%, precision of 74.29%, recall of 73.57%, and an F1-score of 73.51%. With oversampling, these metrics improve to an accuracy of 80.21%, precision of 81.34%, recall of 80.21%, and an F1-score of 80.45%.

Keywords: Industrial Sector Predictions; Indonesia stock exchange; Financial statements; Random Forest; SMOTE

1. INTRODUCTION

Companies listed on the Indonesia Stock Exchange (IDX) routinely issue financial reports every quarter and year [1]. The use of machine learning in research based on financial report predictions is a very relevant and interesting topic. The main problem faced is how the Random Forest method can be used to predict stocks in each company in the industrial sector on the IDX. Traditional financial analysis methods are often unable to handle very large and complex amounts of data, resulting in less accurate predictions [2]. Therefore, a more effective solution is needed to improve the accuracy of the company's financial report predictions [3]. One of the expected solutions is the use of the Random Forest method which has been proven effective in various studies to make predictions based on complex data [4].

This study aims to predict the shares of companies in the industrial sector listed on the IDX based on financial reports using the Random Forest method. The Random Forest method is an ensemble learning technique that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. This method is particularly effective for handling large datasets and complex interactions between variables, making it a powerful tool for financial prediction. The dataset used in this study includes financial data from companies listed on the IDX in the period 2010 to 2022. There are 10 company sectors on the IDX, and this study uses 17 features from each financial report. The data processing process includes data cleaning, data splitting by dividing training and testing data with a ratio of 80% and 20%, overcoming the imbalance in the number of each sector with the oversampling technique using SMOTE, and feature scaling using StandardScaler.

Several related studies have shown the superiority of the Random Forest method in various contexts. H. Van Der Heijden (2022) illustrated that the Random Forest method is superior to Linear Discriminant Analysis (LDA) in predicting industrial sectors, with Random Forest achieving 83% accuracy, while LDA only achieved 70%, indicating superiority in higher classification [5]. Y. S. Soekamto et al. (2023) found that Random Forest gave better results than Neural Network in predicting property categories in Surabaya, with Random Forest achieving 82% accuracy compared to 75% for Neural Network, indicating the superiority of this model in handling complex and varied data [3]. P. Chakri et al. (2023) also identified Random Forest as one of the best models in predicting financial accounting data, achieving 85% accuracy compared to the linear regression model which only achieved 70%, indicating superiority in exploratory data analysis [6]. C. Lohrmann and P. Luukka (2019) showed that Random Forest is more effective than Support Vector Machine (SVM) in classifying S&P500 intraday returns, with Random Forest achieving 80% accuracy while SVM only achieved 73%, indicating the model's ability to handle highly volatile data [7]. In addition, O. D. Madeeh and H. S. Abdullah (2021) found that Random Forest is more effective than techniques such as K-Nearest Neighbors (KNN) and Naive Bayes in predicting the stock market, with Random Forest achieving 83% accuracy, compared to KNN which achieved 72% and Naive Bayes which only achieved 68%, indicating higher accuracy and resilience to imbalanced data [2].

This study focuses primarily on the Indonesian stock market, especially companies listed on the Indonesia Stock Exchange (IDX), using more comprehensive financial report data from 2010 to 2022. One of the problems faced in this study is the imbalance in the number of companies in several sectors, resulting in the existence of a majority sector and a minority sector. The majority sector is represented by sector B, while the minority sector is represented by sector G. To overcome this imbalance, the oversampling method is used. The oversampling method works by increasing the number of samples in each sector to achieve balance between sectors [8]. This study also compares the

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

results of datasets that have not been addressed with the oversampling method and those that have been addressed with the oversampling method using SMOTE. The aim is to evaluate the effectiveness of this technique in improving the performance of the prediction model [9]. The implementation of the oversampling method is expected to overcome the bias caused by data imbalance and improve the accuracy of predictions of the stock industry sector on the IDX.

The purpose of this study is to develop an accurate prediction model for the stock industry sector on the IDX based on the company's financial statements using the Random Forest method. And to compare the evaluation results between datasets with the oversampling method and datasets without the oversampling method determined based on the values of 4 aspects, namely accuracy, precision, recall and accuracy.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This study predicts the stock industry sector in all companies in the industrial sector on the Indonesian stock exchange. The system design in this study uses a flowchart to describe the process of the system from start to finish which can be seen in Figure 1. The process begins with data input which then goes through a preprocessing stage for data cleaning and preparation. After that, the data is divided into two sets, namely training data and testing data. Training data is used to train the model, while testing data is used to evaluate the trained model. The results of the model performance evaluation are then analyzed to assess model performance, and this process ends with the Random Forest performance evaluation stage.

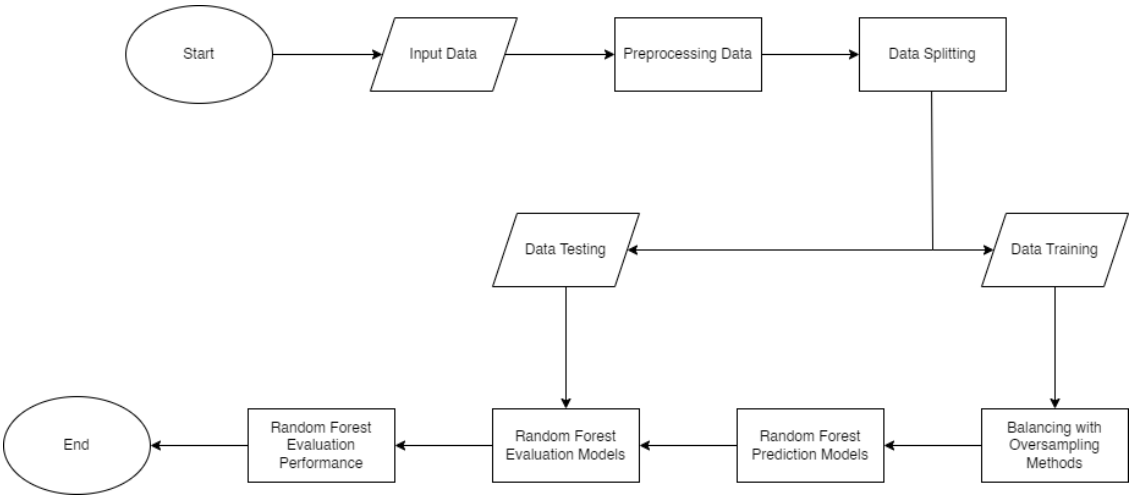

Figure 1. Research stages

2.2 Input Data

In this section, there are 10 industrial company sectors and here we will describe the sectors in the IDX (Indonesia Stock Exchange). After the description, a dataset will be inputted from the financial reports of several companies that have been selected from the Indonesia Stock Exchange (IDX). The data used is from the financial reports of each company in the industrial sector on the Indonesia Stock Exchange with a range from 2010 to 2022. The description of the industrial company sector can be seen in table 1.

To ensure that the data used is representative and accurate, companies that have complete financial reports during the research period are selected. This aims to avoid bias that may arise due to incomplete or inconsistent data. In addition, the data that has been collected is then compiled and analyzed to gain an initial understanding of the distribution of sectors and the financial characteristics of companies in each sector. This analysis is an important basis in the data pre-processing stage and the development of prediction models.

Table 1 provides a detailed breakdown of the industrial company sectors included in this study. Each sector is assigned a target class code, IDX-IC code, description, and an anonymized code. The target class codes range from 0 to 9, representing different industrial sectors, such as Energy, Basic Materials, Industrials, Consumer Non-Cyclicals, Consumer Cyclicals, Healthcare, Financials, Properties & Real Estate, Technology, and Infrastructures. This categorization is crucial for the subsequent data analysis and modeling processes, as it allows for a structured approach to handling the diverse financial data of companies within these sectors. The anonymized codes (e.g., ENE for Energy) ensure confidentiality while enabling clear identification of each sector throughout the study

Table 1. Industrial Company Sector

Target Class Code	IDX-IC Code	Description	Anonym
0	A	Energy	ENE
1	B	Basic Materials	BAS

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

2	C	Industrials	IND
3	D	Consumer Non-Cyclicals	CNC
4	E	Consumer Cyclicals	COC
5	F	Healthcare	HEA
6	G	Financials	FIN
7	H	Properties & Real Estate	PRO
8	I	Technology	TEC
9	J	Infrastructures	INF

2.3 Preprocessing Data

Data preprocessing, also known as data pre-processing, involves several important steps to ensure data quality and consistency. The first step involves removing irrelevant characters such as commas and quotation marks from the dataset, while the minus sign is retained to retain relevant negative values [6]. Next, all blank or missing values are filled with zeros to ensure that no missing data can affect the analysis results [10]. After that, each column of relevant financial data is converted into a numeric format, with special handling for data that cannot be converted into numbers, which are then marked as NaN (Not a Number). All rows containing NaN values are then removed to ensure that the dataset used in the analysis only contains valid data and is ready for further processing [6].

Subsequently, data scaling is conducted using the StandardScaler method, which standardizes features by adjusting them to have a mean of zero and a standard deviation of one [1]. This standardization process is crucial to ensure that each feature is on the same scale, preventing the machine learning model from assigning disproportionate weight to features with larger scales. By employing StandardScaler, all features in the dataset are normalized, thereby enhancing model performance and ensuring consistency in evaluation [10].

$$Z = \frac{X-\mu}{\sigma} \tag{1}$$

The formula depicted represents the Z-score, a statistical measure used to determine the relative position of an individual data point within a dataset. The Z-score (Z) is computed by subtracting the mean (μ) from the original data value (X) and dividing the result by the standard deviation (σ). This standardized metric allows for the comparison of data points by indicating how many standard deviations a particular value deviates from the mean. The Z-score is essential in statistical analysis for identifying outliers and understanding the distribution of data in relation to the mean.

This section also describes 17 features used from the financial statements of each company in this industrial sector, which can be seen in table 2. The 17 features used in the dataset are taken from the financial statements of companies in the industrial sector. These features cover various financial components that are important for analyzing company performance, such as assets, liabilities, and income statement components. These features include Cash & Marketable Securities, Receivables, Inventories, Total Current Assets, Short Investments, Total Non-Current Assets, Total Assets, Current Liabilities, Non-Current Liabilities, Total Liabilities, Total Equity, Gross Profit, Total Revenue, Income From Operations, Income Before Tax, Net Income For The Period, and Total Comprehensive Income. By using these features, the model can categorize companies into the appropriate industrial sector based on relevant financial indicators.

Table 2. Company financial report features

Common Size Component	Financial Statement
Assests	Cash & Marketable Securities
	Receivables
	Inventories
	Total Current Assets
	Short Investments
	Total Non-Current Assets
	Total Assets
Liabilities	Current Liabilities
	Non-Current Liabilities
	Total liabilities
Equity	Total Equity
	Gross Profit

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

Common Size Component	Financial Statement
Income Statement	Total Revenue
	Income From Operations
	Income Before Tax
	Net Income For The Period
	Total Comprehensive Income

2.4 Data Splitting

Data separation, also known as data splitting, is a technique that divides data into distinct subsets to facilitate model training and evaluation. Specifically, the dataset is divided into two parts: the training set, which comprises 80% of the data, is utilized for training the model, while the testing set, encompassing the remaining 20%, is reserved for evaluating the model's performance [3]. This division ensures that the model is trained on a substantial portion of the data while being rigorously tested on a separate, unseen portion to assess its predictive accuracy and generalization capabilities [2].

2.5 Data Training

Data training refers to the subset of the dataset used to train the machine learning model. In this process, the training data is used to fit the model, allowing it to learn and identify patterns within the data. The goal is to teach the model to make accurate predictions based on the features provided.

2.6 Data Testing

Data testing is the phase where the trained machine learning model is evaluated using a separate subset of the dataset that was not used during the training phase. The purpose of data testing is to assess the model's performance and generalization ability on unseen data.

2.7 Random Forest

A crucial process in machine learning is model selection, which is conducted using various metrics and techniques depending on the type of task the model performs, such as classification, regression, or clustering [11]. For this study, the Random Forest method is selected [12]. The Random Forest Classifier is a machine learning algorithm that constructs multiple decision trees to make robust and stable predictions [13]. Essentially, this algorithm uses a training dataset. The model-building process with the Random Forest Classifier is illustrated in Figure 2.

In Figure 2, the dataset is divided into several decision trees, labeled as "Decision Tree-1," "Decision Tree-2," ..., up to "Decision Tree-N." Each decision tree is built using a randomly selected subset of the data, both in terms of samples and features. This randomness introduces diversity among the trees, which is a key feature of the Random Forest algorithm.

Each decision tree splits the data at various nodes, represented by circles in the diagram. The red and blue colors indicate different branches or decisions made at each node, leading to different outcomes or predictions at the leaf nodes (circles at the bottom). The output of each decision tree is referred to as "Result-1," "Result-2," and so forth, up to "Result-N."

The results from each decision tree are then combined using the "Majority Voting/Averaging" method. For classification tasks, the final prediction is determined based on the majority vote from all the trees. This process yields the "Final Result," which is the ultimate prediction made by the Random Forest model based on the aggregated results of all the decision trees.

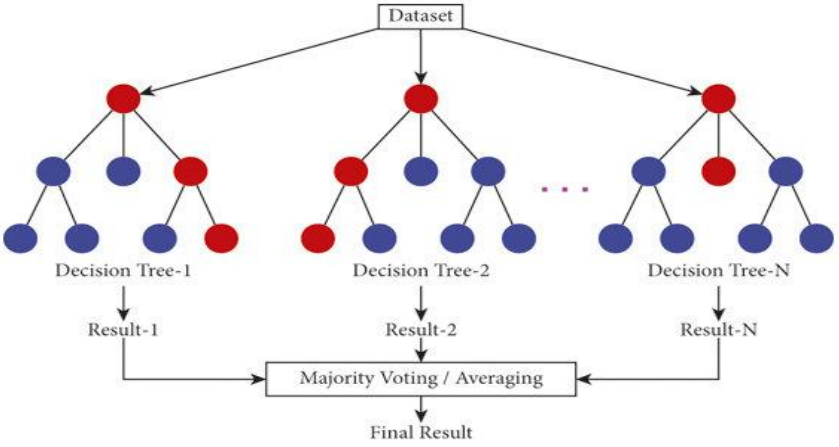

Figure 2. Random forest classifier

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

2.8 Synthetic Minority Over-sampling Technique

The Synthetic Minority Over-sampling Technique (SMOTE) is an oversampling method used to address class imbalance in datasets [14]. This imbalance occurs when the number of samples in one or more classes is significantly lower than in other classes, which can lead to biased predictive models [15]. SMOTE generates new synthetic samples from the minority class instead of merely duplicating existing ones [16].

The SMOTE process begins by randomly selecting samples from the minority class. For each selected sample, several nearest neighbors from the same class are identified based on Euclidean distance in the feature space. SMOTE then creates new synthetic samples by interpolating between the selected sample and its neighbors [11]. This is done by choosing a point along the line segment that connects the original sample and its neighbors in the feature space.

This method effectively reduces the likelihood of overfitting, a common issue with simple oversampling methods that duplicate minority class samples. SMOTE generates more diverse samples within the minority class, thus helping the model learn from a more representative and balanced dataset. This technique is particularly useful in various machine learning applications, especially when dealing with datasets that have highly imbalanced class distributions [17].

2.9 Balancing with Oversampling Methods

After the training data is separated from the testing data through the data splitting process, the next step is to handle the data imbalance in the dataset. Data imbalance occurs when the number of samples from several classes in a dataset is very different, causing bias in the machine learning model trained with the data. In the context of this study, the oversampling method is applied to overcome the imbalance problem.

The oversampling method used is the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE works by creating new synthetic samples from the minority class by interpolating between existing samples. This process increases the number of samples in the minority class so that it is more balanced with the majority class. This technique helps the model to learn the patterns of the minority class better, thereby improving the accuracy and performance of predictions.

After oversampling is applied, the balanced training data is used to train the Random Forest model. Models trained with balanced data are expected to provide more accurate and fair predictions for all classes in the dataset. This process ensures that the model is not only well trained on the majority class but is also able to recognize and predict the minority class well. This implementation is important to obtain better model performance evaluation results and avoid bias in predicting the stock industry sector on the IDX.

2.10 Random Forest Prediction Models

Once the training data is balanced, the next step is to build a pipeline with the Random Forest training model. To identify the optimal hyperparameter combination that enhances the model's performance, GridSearchCV is utilized. GridSearchCV conducts an extensive search over the pre-defined hyperparameter space using cross-validation techniques to evaluate each combination. This approach ensures that the Random Forest model is optimized to achieve the best possible performance [18].

2.11 Random Forest Evaluation Models

In the evaluation phase, the performance of the trained model is assessed using the test data. The first step involves leveraging the trained model to make predictions on the test data, generating predictions that are then compared with the actual values. Model performance evaluation is conducted by calculating the confusion matrix, which depicts the number of correct and incorrect predictions made by the model for each class. Additionally, model accuracy is calculated to determine the proportion of correct predictions out of the total predictions made. To provide a more comprehensive overview of the model's performance, a classification report is compiled, encompassing metrics such as precision, recall, and F1-score for each class, as well as macro and micro averages for the overall model [19].

2.12 Random Forest Evaluation Performance

To ensure that the machine learning model can make accurate predictions, a process known as performance evaluation is conducted. In this section, performance evaluation is based on the results from the previously trained and tested model using the Random Forest method. This performance evaluation utilizes multi-class confusion matrix assessment, which maps the model's predictions into multiple different classes, rather than just two classes as in binary cases. The multi-class confusion matrix provides a detailed overview of how the model predicts each class, including the number of correct predictions for each class (True Positives), as well as prediction errors such as False Positives and False Negatives that occur between classes, as illustrated in Table 1. This evaluation is crucial for understanding how well the model recognizes and differentiates between the different classes in the dataset and for identifying areas where the model may need improvement.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

Table 3. Confusion matrix 3 classes

		Predicted		
		A	B	C
Actual	A	TP	FP	FP
	B	FN	TP	FP
	C	FN	FN	TP

True Positive (TP) refers to the count of correctly classified positive instances, indicating the model's ability to identify actual positive cases. False Negative (FN) represents the number of positive instances that the model incorrectly classified as negative, highlighting missed detections. False Positive (FP) denotes the count of negative instances that were incorrectly classified as positive, reflecting false alarms. True Negative (TN) refers to the number of correctly classified negative instances, demonstrating the model's accuracy in identifying true negative cases. These metrics are fundamental in assessing the performance of a classification model, providing insights into its ability to correctly differentiate between positive and negative instances.

In this section, we will use general steps to evaluate the efficiency and accuracy of the prediction model, which provides 4 measurements, namely precision, recall, accuracy and F1-score. The following is an explanation of the equations of the 4 measurements.

$$Accuracy = \frac{\sum_{i=1}^{K} TP_i}{N} \tag{2}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{3}$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{4}$$

$$F_i = \frac{2 \, x \, P_i \, x \, R_i}{P_i + R_i} \tag{5}$$

The description specifies the key variables employed in the evaluation of classification models. The K denotes the number of data classes, the N refers to the total number of sample data, The P_i represents the precision for class i, the R_i signifies the recall for class i, and the F_i indicates the F1-score for class i. These variables are integral to the computation of crucial performance metrics, such as accuracy, precision, recall, and the F1-score, which are essential for assessing the overall performance and reliability of classification models in various datasets.

The accuracy metric indicates how often the model makes correct predictions. Recall measures how well the model identifies all actual positive cases, while precision measures the proportion of positive predictions that are actually positive. The F1-Score is the harmonic mean of recall and precision, providing a balance between the two metrics. Consequently, the confusion matrix helps assess the overall effectiveness of the model and reveals the types of errors the model makes [20].

3. RESULT AND DISCUSSION

3.1 Dataset

The dataset was collected from companies in 11 different sectors listed on the Indonesia Stock Exchange (IDX) over the period from 2010 to 2022. The data was sourced from the financial reports of these companies, comprising 17 features utilized in this dataset. The imbalance in the dataset is evident from the distribution of companies across different sectors, as shown in Table 4. The sector with the most data is Consumer Non-Cyclicals (Sector D) with 716 companies, while the sector with the least data is Financials (Sector G) with 85 companies, indicating a significant imbalance in the dataset.

This imbalance, characterized by the disproportionate number of companies in each sector, can adversely affect the predictive performance of the model, as it tends to be more trained on sectors with more data and less on those with fewer data. To address this imbalance and enhance model accuracy, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. This method generates a more balanced dataset across all sectors, allowing the model to provide better and less biased predictions. By implementing oversampling, the model gains a better understanding of the characteristics of each sector, ultimately leading to improved predictive performance.

The imbalance in the number of companies across different sectors necessitates the use of oversampling methods like SMOTE to ensure that the model can effectively learn from and predict for all sectors, not just those with more data.

Table 4. Distribution of company sectors

Sector	Company	Percentage

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

A	437	10.18%
B	634	14.76%
C	480	11.18%
D	716	16.67%
E	715	16.65%
F	155	3.61%
G	85	1.98%
H	412	9.59%
I	96	2.24%
J	375	8.73%
K	189	4.40%

3.2 Confusion Matrix

In Figure 3, the confusion matrix analysis after applying the oversampling method shows that the model performs well in identifying certain classes. For instance, class 0 (Energy) with 98 actual class members has minimal prediction errors, indicating that the model is effective in identifying this class. However, class 1 (Basic Materials), with 143 actual class members, shows significant prediction errors towards class 2 (Industrials) with 104 members and class 3 (Consumer Non-Cyclicals) with 139 members. This may be due to similar feature characteristics between these classes. Conversely, class 2 (Industrials) shows minimal prediction errors across all classes, indicating clear and distinct data characteristics. Classes 3 (Consumer Non-Cyclicals) through 10 (Infrastructures) also exhibit minimal prediction errors, showing that the model can identify these classes quite well. The specific and distinct characteristics of these classes likely facilitate accurate predictions. Overall, while the model performs well in identifying most classes, some classes like Basic Materials frequently misclassify into other classes with similar data characteristics.

In Figure 4, the confusion matrix without applying the oversampling method indicates a drop in model performance. Class 0 (Energy) with 88 actual class members shows greater prediction errors compared to the results after oversampling. Class 1 (Basic Materials) with 121 actual members also shows increased prediction errors towards class 2 (Industrials) with 111 members and class 3 (Consumer Non-Cyclicals) with 138 members. This suggests that without oversampling, the model struggles to distinguish between classes with similar feature characteristics. Similarly, class 2 (Industrials) shows increased prediction errors compared to the oversampled model. Classes 3 (Consumer Non-Cyclicals) through 10 (Infrastructures) follow a similar pattern, with higher prediction errors without oversampling, indicating that data imbalance affects the model's predictive accuracy.

From these analyses, it can be concluded that applying the oversampling method with SMOTE significantly improves the model's performance in identifying classes with imbalanced data. The oversampled model shows higher accuracy and fewer prediction errors compared to the model without oversampling. Therefore, data balancing techniques like oversampling are crucial for improving the predictive accuracy of machine learning models in cases of data imbalance.

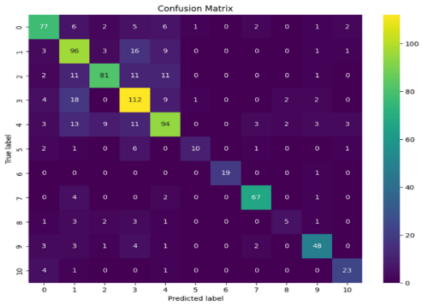

Figure 3. Confusion Matrix without oversampling

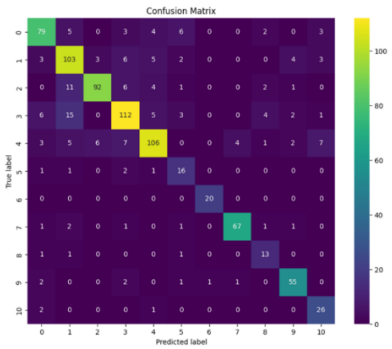

Figure 4. Confusion Matrix with oversampling

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

3.3 Classification Report

The process of implementing the Random Forest algorithm to derive precision, recall, accuracy, and F-1 score values involves several key steps. First, to handle imbalanced data, the SMOTE (Synthetic Minority Over-sampling Technique) is applied to the training data, generating synthetic samples for the minority class to achieve a balanced dataset. Following this, the Random Forest model is trained using the balanced training data. Once the model is trained, it is evaluated using the testing set. During this evaluation, the model's predictions are compared with the actual values to construct a confusion matrix. This matrix is then used to calculate the precision, recall, accuracy, and F-1 score for each class. Precision measures the proportion of true positive predictions out of all positive predictions made by the model, recall measures the model's ability to identify all actual positive cases, and the F-1 score provides a harmonic mean of precision and recall. Accuracy reflects the overall proportion of correct predictions made by the model. The detailed metrics for all classes are presented in the tables, offering a comprehensive view of the model's performance before and after the application of oversampling

Based on Table 5, which shows the classification report without oversampling, it is evident that the model's performance is lower compared to Table 6, which presents the results after applying oversampling. Before using oversampling, the model's overall accuracy was 73.57%, with lower precision, recall, and F-1 scores for several classes. For instance, in class 5 (Healthcare), the precision was only 83.33%, recall was 47.62%, and the F-1 score was 60.61%, indicating the model's inability to distinguish the features of this class from others.

After applying oversampling, as shown in Table 6, the overall accuracy increased to 80.21%. Performance improvements were observed in almost all classes, with class 6 (Financials) showing a precision of 95.24%, recall of 100%, and an F-1 score of 97.56%. This indicates that oversampling aids the model in recognizing and learning the features of classes more effectively, particularly for those with previously imbalanced data. For example, in class 5 (Healthcare), the precision improved to 51.61% and the F-1 score to 61.54%, demonstrating that despite ongoing challenges, the model provided more balanced predictions after applying the oversampling technique.

Overall, the use of the oversampling method not only improved the model's overall accuracy from 73.57% to 80.21% but also enhanced the precision from 74.29% to 81.34%, recall from 73.57% to 80.21%, and F-1 score from 73.51% to 80.45%. This underscores the importance of proper data preprocessing, especially in the context of imbalanced data, to improve the performance of machine learning models and provide fairer and more balanced predictions across all classes.

Table 5. Classification report without oversampling

Class	Precision	Recall	F-1 Score
0	77.78%	75.49%	76.62%
1	61.54%	74.42%	67.37%
2	82.65%	69.23%	75.35%
3	66.67%	75.68%	70.89%
4	70.15%	66.67%	68.36%
5	83.33%	47.62%	60.61%
6	100.00%	95.00%	97.44%
7	89.33%	90.54%	89.93%
8	55.56%	31.25%	40.00%
9	81.36%	77.42%	79.34%
10	76.67%	79.31%	77.97%
Accuracy	73.57%	73.57%	73.57%
Macro Avg	76.82%	71.15%	73.08%
Weighted Avg	74.29%	73.57%	73.51%

Table 6. Classification report with oversampling

Class	Precision	Recall	F-1 Score
0	80.61%	77.45%	79.00%
1	72.03%	79.84%	75.74%
2	91.09%	78.63%	84.40%
3	80.58%	75.68%	78.05%
4	84.13%	75.18%	79.40%
5	51.61%	76.19%	61.54%
6	95.24%	100.00%	97.56%
7	93.06%	90.54%	91.78%
8	56.52%	81.25%	66.67%
9	84.62%	88.71%	86.61%
10	65.00%	89.66%	75.36%
Accuracy	80.21%	80.21%	80.21%
Macro Avg	77.68%	83.01%	79.65%

Building of Informatics, Technology and Science (BITS)
Volume 6, No 2, September 2024, Page:1002 −1011
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i2.5743

Weighted Avg	81.34%	80.21%	80.45%

4. CONCLUSION

The conclusions of this study indicate that the use of the Random Forest method with the application of oversampling through the SMOTE technique successfully improved the model's performance in predicting industrial stock sectors on the Indonesia Stock Exchange (IDX). The evaluation of the model without oversampling showed an overall accuracy of 73.57%, with precision of 74.29%, recall of 73.57%, and F1-score of 73.51%, highlighting the challenges faced by the model in handling imbalanced data, leading to less accurate predictions for minority classes. In contrast, the model with oversampling achieved a significant improvement in performance, with an overall accuracy of 80.21%, precision of 81.34%, recall of 80.21%, and F1-score of 80.45%, demonstrating the effectiveness of the data balancing technique in enhancing the model's ability to accurately identify and predict classes with previously imbalanced data. Overall, this study makes a significant contribution to the application of machine learning for financial report prediction in the Indonesian stock market. The primary focus of this research is on the use of the Random Forest method, which has proven to be effective in classifying industrial stock sectors with high accuracy and fairness after applying the oversampling technique. The comparison between models using the oversampling technique and those that do not demonstrates that oversampling significantly enhances the model's ability to address data imbalance, resulting in more accurate and representative predictions across all industrial sectors.

REFERENCES

[1] W. Budiharto, "Data science approach to stock prices forecasting in Indonesia during Covid-19 using Long Short-Term Memory (LSTM)," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00430-0.
[2] O. D. Madeeh and H. S. Abdullah, "An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, 2021, doi: 10.1088/1742-6596/1804/1/012008.
[3] Y. S. Soekamto, M. Chandra, T. Wiradinata, R. Tanamal, and T. R. D. Saputri, *Property Category Prediction Model using Random Forest Classifier to Improve Property Industry in Surabaya*. Atlantis Press International BV, 2023. doi: 10.2991/978-94-6463-144-9_24.
[4] D. Makariou, P. Barrieu, and Y. Chen, "A random forest based approach for predicting spreads in the primary catastrophe bond market," *Insur. Math. Econ.*, vol. 101, no. Breiman 2001, pp. 140–162, 2021, doi: 10.1016/j.insmatheco.2021.07.003.
[5] H. van der Heijden, "Predicting industry sectors from financial statements: An illustration of machine learning in accounting research," *Br. Account. Rev.*, vol. 54, no. 5, p. 101096, 2022, doi: 10.1016/j.bar.2022.101096.
[6] P. Chakri, S. Pratap, Lakshay, and S. K. Gouda, "An exploratory data analysis approach for analyzing financial accounting data using machine learning," *Decis. Anal. J.*, vol. 7, no. March, p. 100212, 2023, doi: 10.1016/j.dajour.2023.100212.
[7] C. Lohrmann and P. Luukka, "Classification of intraday S&P500 returns with a Random Forest," *Int. J. Forecast.*, vol. 35, no. 1, pp. 390–407, 2019, doi: 10.1016/j.ijforecast.2018.08.004.
[8] H. Daori, "Predicting Stock Prices Using the Random Forest Classier," 2022, [Online]. Available: https://doi.org/10.21203/rs.3.rs-2266733/v1
[9] P. Ghosh, A. Neufeld, and J. K. Sahoo, "Forecasting directional movements of stock prices for intraday trading using LSTM and random forests," *Financ. Res. Lett.*, vol. 46, no. December 2018, 2022, doi: 10.1016/j.frl.2021.102280.
[10] M. Vijh, D. Chandola, V. A. Tikkiwal, and A. Kumar, "Stock Closing Price Prediction using Machine Learning Techniques," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 599–606, 2020, doi: 10.1016/j.procs.2020.03.326.
[11] B. Mohammadi ivatlood, C. Spampinato, R. Chopra, K. C. Lee, and S. S. Roy, "Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on South Korean companies," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 33, no. 1, p. 62, 2020, doi: 10.1504/ijahuc.2020.10026453.
[12] A. M. N. Alzubaidi and E. S. Al-Shamery, "Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in telecom industry," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 1406–1421, 2020, doi: 10.11591/ijece.v10i2.pp1406-1421.
[13] X. Zhong and D. Enke, "Predicting the daily return direction of the stock market using hybrid machine learning algorithms," *Financ. Innov.*, vol. 5, no. 1, 2019, doi: 10.1186/s40854-019-0138-0.
[14] A. Bin Omar, S. Huang, A. A. Salameh, H. Khurram, and M. Fareed, "Stock Market Forecasting Using the Random Forest and Deep Neural Network Models Before and During the COVID-19 Period," *Front. Environ. Sci.*, vol. 10, no. July, pp. 1–10, 2022, doi: 10.3389/fenvs.2022.917047.
[15] E. González-Núñez, L. A. Trejo, and M. Kampouridis, "A Comparative Study for Stock Market Forecast Based on a New Machine Learning Model," *Big Data Cogn. Comput.*, vol. 8, no. 4, 2024, doi: 10.3390/bdcc8040034.
[16] K. Kaczmarczyk and M. Hernes, "Financial decisions support using the supervised learning method based on random forests," *Procedia Comput. Sci.*, vol. 176, pp. 2802–2811, 2020, doi: 10.1016/j.procs.2020.09.276.
[17] J. Shen and M. O. Shafiq, "Short-term stock market price trend prediction using a comprehensive deep learning system," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00333-6.
[18] N. Rouf *et al.*, "Stock market prediction using machine learning techniques: A decade survey on methodologies, recent developments, and future directions," *Electron.*, vol. 10, no. 21, 2021, doi: 10.3390/electronics10212717.
[19] P. Sadorsky, "A Random Forests Approach to Predicting Clean Energy Stock Prices," *J. Risk Financ. Manag.*, vol. 14, no. 2, 2021, doi: 10.3390/jrfm14020048.
[20] T. P. Ogundunmade, A. A. Adepoju, and A. Allam, "Stock Price Forecasting: Machine Learning Models with K-fold and Repeated Cross Validation Approaches," *Mod. Econ. Manag.*, no. June, 2022, doi: 10.53964/mem.2022001.