p–ISSN: 2723 - 6609 e-ISSN: 2745-5254
Vol. 5, No. 10, October 2024 http://jist.publikasiindonesia.id/

Prediction of Stock Industry Sectors Listed on the Indonesia Stock Exchange (IDX) based on Financial Statements with the Random Forest Method

I Kamil Elian Zhafran[1*], Deni Saepudin[2]
Universitas Telkom Bandung, Indonesia
Email: zhafrankamil@students.telkomuniversity.ac.id[1*], denisaepudin@telkomuniversity.ac.id[2]
*Correspondence

ABSTRACT

Keywords: random forest; indonesian stock exchange; industrial sector predictions; financial reports.

This research aims to predict the stock industry sector listed on the Indonesia Stock Exchange (BEI) based on financial reports using the Random Forest method. The dataset used in this research includes financial data from companies listed on the IDX in the period 2010 to 2022. The data processing process includes data cleaning, handling class imbalance with oversampling techniques using SMOTE, and feature scaling using StandardScaler. The Random Forest model is used to classify companies into appropriate industry sectors. The evaluation results show that the model has good performance with an overall accuracy of 80.21%. Several classes showed very good performance, such as the Financials class with precision of 95.24%, recall of 100%, and F-1 score of 97.56%. However, some classes show lower performance, such as the Healthcare class with a precision of 51.61% and an F-1 score of 61.54%. The confusion matrix indicates that the model can identify most classes accurately, although there are several classes with prediction errors.

Introduction

Since companies listed on the Indonesia Stock Exchange (IDX) issue financial statements every quarter and year, the use of machine learning in prediction-based research is a very interesting topic. (Soekamto et al., 2023). Machine learning methods can help in making accurate predictions for a company's financial statements, thus making this case even more relevant. The main focus of this research is to develop more comprehensive predictions and improve prediction accuracy by using new techniques in machine learning. (Roy et al., 2020).

Current methods of financial analysis are often incapable of handling very complex amounts of data, resulting in less accurate predictions. Previous research, such as those conducted by (Van Der Heijden, 2022), shows that the Random Forest method predicts the performance of industrial sectors better than other methods, such as Linear Discriminant Analysis (LDA). However, this study has not fully studied the potential of

Random Forest in the Indonesian stock market, especially in a changing market situation such as the COVID-19 pandemic.

This final project will solve the above problem by using the Random Forest method. This method will be used to predict the company's industrial sector by classifying the dataset and obtaining accurate results along with a classification report containing Precision, Recall, and F-1 Score values.

The focus of this research is how the author creates a prediction model using the Random Forest method to predict industrial sectors based on financial statements, as well as how to classify businesses into the right industrial sectors. This research is limited to financial statement data from companies listed on the Indonesia Stock Exchange in a certain period and the Random Forest method for classification.

The limitation of the problem in this study is that the features used come from the income statement and balance sheet of each annual financial statement of all industrial sector companies listed on the Indonesia Stock Exchange, with the period of each company whose financial statements are taken for the past 12 years.

Several studies have been conducted on financial predictions in the industrial sector based on financial statements. In 2022, Hans Van Der Heijden made stock predictions for industrial sector companies in the North American region (NAICS) based on the financial statements of each company. The Linear Discriminant Analysis (LDA) method and the Random Forest method aim to make a comparison between 2 types of methods, namely the non-linear method and the linear method. (Alzubaidi & Al-Shamery, 2020). The results obtained are linear methods, namely the Random Forest method provides a higher accuracy value compared to the LDA method so it can be concluded that the Random Forest method is superior in predicting the industrial sector compared to LDA. (Van Der Heijden, 2022). In 2022, Omar A. et al conducted a study aimed at comparing the accuracy of machine learning models in predicting stock market index prices before and during the COVID-19 period. The methods used in this journal are Autoregressive Integrated Moving Average (ARIMA) to conduct statistical analysis as well as Neural Network and Random Forest in modeling the non-linear structure of the data. A comparison of the accuracy of different models shows that the Neural Network in Autoregressive and the Autoregressive Random Forest model is the best for forecasting stock index prices for different periods. (Chakri et al., 2023).

In 2021, Perry Sadorsky explored machine learning methods using the bagging decision trees, random forest, and logit models methods in predicting stock prices for companies engaged in the Clean Energy and Technology sector. In addition to predicting stock prices, this study also aims to identify the importance of variable factors in the random forest method and provide valuable insights into the use of machine learning methods to predict stock prices (Lohrmann & Luukka, 2019). In 202, Omar D. Madeeh, et al. conducted research using the KNN and Random Forest methods which aimed to develop and present an efficient prediction model based on machine learning techniques used in predicting stock market prices. This study uses the K-Nearest Neighbor (KNN) and Random Forest (RF) algorithms and analyzes the performance of both in predicting stock market movements (Madeeh & Abdullah, 2021). In 2019, Lohrrman Christoph, et al. conducted research with 2 methods, namely Random Forest and Fuzzy Similarity and Entropy Measure (FSAE). In the above study, a comparison between the FSAE and Random Forest methods will be carried out in predicting calcification. This journal shows that Random Forest has a higher performance in accuracy compared to FSAE which can be concluded that the Random Forest method is more effective in predicting classification than FSAE (Sadorsky, 2021).

This study seeks to show that the Random Forest method is effective and relevant in dealing with the problem of predicting the industrial sector in the research by understanding all aspects of the financial statements of each industry. (Omar et al., 2022). The main purpose of this study is to determine the influential and relevant features in making predictions of the industry sector based on the company's financial statements. In addition, this study shows that a machine learning-based approach can be used to improve the ability to predict companies' financial statements in the industrial sector. (Kaczmarczyk & Hernes, 2020). In addition, the purpose of this study is to develop and validate a prediction model that uses the Random Forest algorithm to put companies into the right industry sectors based on their public financial report data. Therefore, it is expected that this research will make a significant contribution to the predictive field and help improve the accuracy of the predictions of the industrial sector.

Method

The system design in this study is to use a flowchart to describe the process of the system from start to finish:

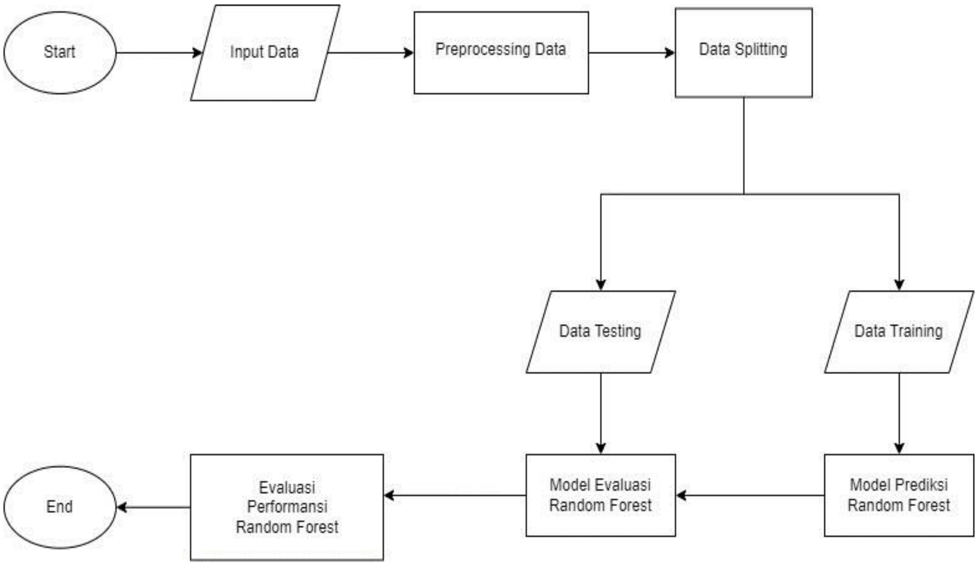

Start
Input Data
Preprocessing Data
Data Splitting
Data Testing
Data Training
Model Prediksi Random Forest
Model Evaluasi Random Forest
Evaluasi Performansi Random Forest
End

Figure 1 System Design

Preprocessing Data

Data pre-processing, also known as data pre-processing, is typically done by eliminating inappropriate data to eliminate any problems that may occur during data processing due to the large amount of inconsistent data. In addition, the data will be changed in this process so that the system can understand it better. In this section, the missing values will be checked on the dataset. In this section, 17 features that will be used from the financial statements of each company in this industry sector will also be described, which can be seen as follows:

Table 1 Features of Financial Statements

Common Size Component	Financial Statement
Assets	Cash & Marketable Securities
	Receivables
	Inventories
	Total Current Assets
	Short Investments
	Total Non-Current Assets
	Total Assets
	Current Liabilities

Liabilities	Non-Current Liabilities
	Total liabilities
Equity	Total Equity
Income Statement	Gross Profit
	Total Revenue
	Income From Operations
	Income Before Tax
	Net Income For The Period
	Total Comprehensive Income

Data Splitting

Data splitting, also known as data splitting, is a technique that divides data into two or more parts that form a subset of data. Typically, the first part is separated to be used to test or evaluate the data, and the second part is used to train the model. In this part, the dataset will be divided into 2 parts, namely data training and data testing.

Data Training

Training data commonly known as training data is part of a data set that is provided as model learning material. The goal is for the model to be able to generalize (find patterns) of data to be used in predicting new data. In this section, the dataset division used will consist of 80% of the total dataset.

Data Testing

In this section, the dataset division used in the testing data consists of 20% of the total dataset.

Random Forest Prediction Model

Data scaling in machine learning is the process of adjusting the range of feature or variable values in a dataset. The main goal of data scaling is to ensure that all features are at the same scale so that no feature dominates the other when the machine learning model is training the data. In this section, data scaling, or data standardization, is done using the StandardScaler from the Scikit-Learn Library. Scikit-Learn, also known as scikit-learn, is a bibliotech of the Python programming language used for machine learning. For statistical modeling and machine learning, the institute provides a variety of tools, including classification, regression, grouping, and dimension reduction. To improve the features, StandardScaler is one of the tools in the SciKit-Learn library. It works by altering the data so that it has a mean of zero and a variance of one, in other words, the data is standardized. The StandardScaler uses the mean and standard deviation of the data to do scaling. This process is important because many machine learning algorithms work better or faster when the features in the dataset are at a similar scale. The equation of the StandardScaler can be calculated by the following equation:

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

Information:

Z = scaled value.

X = the original value of the data.

μ = the average value of the data.

σ = standard deviation from the data.

Model Evaluasi Random Forest

An important process in machine learning is model evaluation, this is done with various metrics and techniques depending on the type of task performed by the model, such as classification, regression, or clustering. In the evaluation of this model, the Random Forest method will be used. A Random Forest Classifier is a machine-learning algorithm that builds many decision trees to make stronger and more stable predictions. This algorithm uses a training dataset. (Makariou et al., 2021).

Random Forest Performance Evaluation

To ensure that machine learning models can make accurate predictions on data that has never been seen before (known as data testing or new data), a process called performance evaluation is performed. In this part, a performance evaluation will be carried out with the results of the evaluation model that was carried out previously and the method that will be used, namely in this final project, the Random Forest method will be used. (Ogundunmade et al., 2022). Evaluating performance will be based on the assessment of the confusion matrix, which is a matrix whose test scores have been distributed through the creation of 2 classes as can be seen in the table below:

Table 3
Confusion Matrix

Prediction	Positive	Negative
Positive	HCMC	FN
Negative	**FP**	**TN**

Information:

True Positive (TP): determines the number of positive cases that have been adequately classified.

False Negative (FN): determines the number of inadequately classified positive cases

False Positive (FP): determines the number of negative events that are inadequately classified.

True negative (TN): determines the number of negative events that have been adequately classified.

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 equation of the 4 measurements:

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

Information:

1. Accuracy is the proportion of all correct predictions by measuring how often the model makes correct predictions, for both positive and negative classes.
2. Precision is the proportion of truly positive predictions that indicate how accurate the model is when classifying an instance as positive. High precision means there are fewer false positives.
3. Recall is the proportion of correctly identified actual positive instances which indicates how well the model identifies all positive instances. High recall means there are fewer false negatives.
4. F1-Score is the harmonic average of precision and recall which measures the balance between precision and recall. It is useful when we want to consider both aspects at the same time, especially when the class distribution is unbalanced.

How often the model makes correct predictions is indicated by accuracy. Recall measures how well the model finds all the true positive cases, and its precision measures the proportion of positive predictions of the model that are positive. The F1-Score is the harmonic average score for recall and precision, which provides a balance between the two metrics. Therefore, the confusion matrix helps measure the overall effectiveness of the model and shows the types of errors that the model makes.

Results and Discussion

Dataset

Table 4
Company Sector Distribution

Sector	Company	Percentage
A	437	10.18%
B	634	14.76%
C	480	11.18%
D	716	16.67%

And	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%

The dataset collection was carried out from each company in 11 different sectors on the Indonesia Stock Exchange in the range of 2010 to 2022. The dataset is taken based on the financial statements owned by each company. From the financial report, data was obtained from the features used in this dataset. There are 17 features used in this dataset. An indication of the imbalance of the dataset can be seen from the distribution of each company in each sector as shown in table 4.

There are sectors with the highest amount of company data, namely 716 companies (Sector D - Consumer Non-Cyclicals), and sectors with the least amount of company data, namely 85 companies (Sector G - Financials), which shows that there is instability in this dataset. The imbalance in question is the imbalance in the amount of company data from each sector on the Indonesia Stock Exchange. There is a sector with the highest amount of company data, namely the Consumer Non-Cyclicals sector, with 716 companies, and there is a sector with the lowest number of company data, namely the Financials sector, with 85 companies. (González-Núñez et al., 2024). With a significant comparison between the Consumer Non-Cyclicals and Financials sectors, this condition can be referred to as an imbalance in the dataset.

This imbalance can affect the performance of prediction models because models tend to be more trained in sectors with a larger amount of data and less trained in sectors with a smaller amount of data.

To overcome this imbalance and improve the accuracy of the model, an oversampling method using the SMOTE (Synthetic Minority Over-sampling Technique) technique was applied to the training data. This method results in a more balanced amount of data across each sector so that the model can provide better predictions and is not biased towards specific sectors. With oversampling, the model can better understand the characteristics of each sector.

Classification Report

Table 5
Classification Report

Class	Precision	Recall	F-1 Score

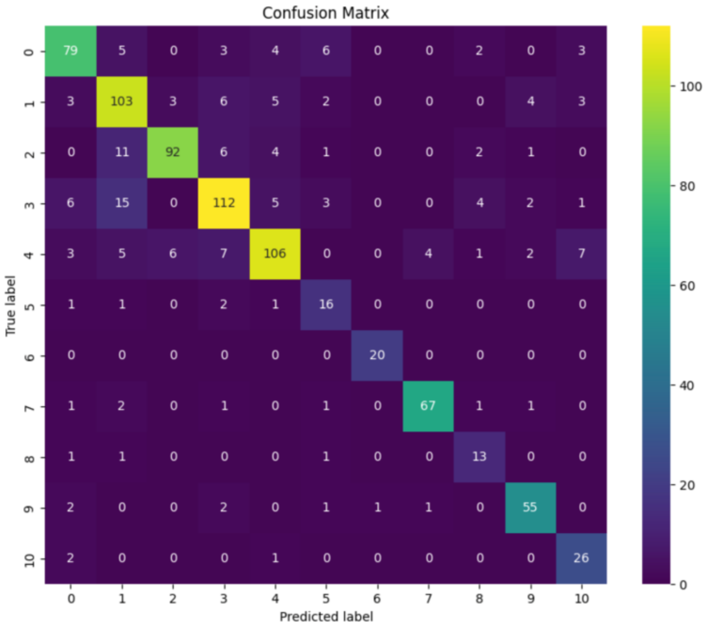
0	80.61%	77.45%	79.00%
1	72.03%	79.84%	75.74%
2	91.09%	78.63%	84.40%
3	80.58%	75.68%	78.05%
4	84.13%	75.18%	79.40%
5	51.61%	76.19%	61.54%
6	95.24%	100.00%	97.56%
7	93.06%	90.54%	91.78%
8	56.52%	81.25%	66.67%
9	84.62%	88.71%	86.61%
10	65.00%	89.66%	75.36%
Accuracy	80.21%	80.21%	80.21%
Macro Avg	77.68%	83.01%	79.65%
Weighted Avg	81.34%	80.21%	80.45%

After the scaling process, the training data that experienced instability in the number of companies per sector was overcome by the oversampling method using SMOTE (Synthetic Minority Over-sampling Technique). This technique results in a more balanced amount of data in each sector so that the model can provide better predictions and is not biased towards certain sectors. The results of the model after training and prediction are shown in the classification performance table above.

From the table, it can be seen that the model has good overall performance with an accuracy value of 80.21%. The precision, recall, and F-1 score metrics for each class show different variations, where some classes have very good performance, such as class 6 (Financials) with 95.24% accuracy, 100% recall, and an F-1 score of 97.56%. This shows that the data for that class is very well defined in the model. On the other hand, class 5 (Healthcare) showed lower performance with a precision of 51.61% and an F-1 score of 61.54%, which indicates a challenge in distinguishing the features of this class from other classes. (Daori et al., 2022).

Overall, the application of data scaling and oversampling methods has helped improve model performance and provide better and balanced prediction results across all classes. This shows the importance of data preprocessing in the machine learning process to get optimal results.

Confussion Matrix



	0	1	2	3	4	5	6	7	8	9	10
0	79	5	0	3	4	6	0	0	2	0	3
1	3	103	3	6	5	2	0	0	0	4	3
2	0	11	92	6	4	1	0	0	2	1	0
3	6	15	0	112	5	3	0	0	4	2	1
4	3	5	6	7	106	0	0	4	1	2	7
5	1	1	0	2	1	16	0	0	0	0	0
6	0	0	0	0	0	0	20	0	0	0	0
7	1	2	0	1	0	1	0	67	1	1	0
8	1	1	0	0	0	1	0	0	13	0	0
9	2	0	0	2	0	1	1	1	0	55	0
10	2	0	0	0	1	0	0	0	0	0	26

Gambar 3. Confussion Matrix Random Forest

Based on the confusion matrix analysis, it can be concluded that the model as a whole shows quite good performance with some classes having very accurate predictions. Class 0 (Energy) has 98 actual class members and very minimal prediction errors in each class, indicating that the model is very effective in identifying these classes. This may be due to the very clear and different data characteristics of this class. (Vijh et al., 2020).

However, class 1 (Basic Materials) which has 143 actual class members shows many prediction errors in class 2 (Industrials) which has 104 members, and class 3 (Consumer Non-Cyclicals) which has 139 members. This may be due to similar feature characteristics or data that is not sufficiently separate between these two classes. In contrast, class 2 (Industrials) has minimal prediction errors in each class, indicating that the data of this class also has distinct and distinct characteristics.

Grades 3 (Consumer Non-Cyclicals) to 10 (Infrastructures) showed minimal prediction errors in each class, indicating that the model was able to identify these classes quite well. The data for these classes may have very specific and different characteristics, making it easier for the model to make accurate predictions. Overall, although the model has good accuracy in identifying most classes, there are some classes such as Basic Materials that often experience prediction errors against other classes that have similar data characteristics.

Conclusion

This study uses the Random Forest method to predict the stock industry sector on the Indonesia Stock Exchange (IDX) based on financial statements. Data from companies listed on the IDX in the period from 2010 to 2022 was processed using data cleaning, oversampling techniques using SMOTE, and feature scaling using StandardScaler.

The results showed that the Random Forest model had good overall accuracy, with some classes showing high performance. However, there is data instability between sectors that affects model performance, where sectors with a higher amount of data tend to be more trained. The oversampling method with SMOTE has successfully helped overcome this instability, resulting in more balanced and fair predictions across all sectors. This study emphasizes the importance of data preprocessing and balancing techniques in improving the performance of machine learning models for industrial sector prediction on the IDX.

Bibliography

Alzubaidi, A. M. N., & Al-Shamery, E. S. (2020). Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in the telecom industry. *International Journal of Electrical & Computer Engineering (2088-8708), 10*(2).

Chakri, P., Pratap, S., & Gouda, S. K. (2023). An exploratory data analysis approach for analyzing financial accounting data using machine learning. *Decision Analytics Journal*, *7*, 100212.

Daori, H., ALHARTHI, M., ALANAZI, A., ALZAHRANI, G., ABOROKBAH, M., & Aljehane, N. (2022). *Predicting Stock Prices Using the Random Forest Classifier*.

González-Núñez, E., Trejo, L. A., & Kampouridis, M. (2024). A Comparative Study for Stock Market Forecast Based on a New Machine Learning Model. *Big Data and Cognitive Computing*, *8*(4), 34.

Kaczmarczyk, K., & Hernes, M. (2020). Financial decisions support using the supervised learning method based on random forests. *Procedia Computer Science*, *176*, 2802–2811.

Lohrmann, C., & Luukka, P. (2019). Classification of intraday S&P500 returns with a Random Forest. *International Journal of Forecasting*, *35*(1), 390–407.

Madeeh, O. D., & Abdullah, H. S. (2021). An efficient prediction model based on machine learning techniques for prediction of the stock market. *Journal of Physics: Conference Series*, *1804*(1), 12008.

Makariou, D., Barrieu, P., & Chen, Y. (2021). A random forest-based approach for predicting spreads in the primary catastrophe bond market. *Insurance: Mathematics and Economics*, *101*, 140–162.

Ogundunmade, T. P., Adepoju, A. A., & Allam, A. (2022). Stock price forecasting: Machine learning models with K-fold and repeated cross-validation approaches. *Mod Econ Manag*, *1*.

Omar, A. Bin, Huang, S., Salameh, A. A., Khurram, H., & Fareed, M. (2022). Stock market forecasting using the random forest and deep neural network models before and during the COVID-19 period. *Frontiers in Environmental Science*, *10*, 917047.

Roy, S. S., Chopra, R., Lee, K. C., Spampinato, C., & Mohammadi-ivatlood, B. (2020). Random forest, gradient boosted machines and deep neural network for stock price forecasting: a comparative analysis on South Korean companies. *International Journal of Ad Hoc and Ubiquitous Computing*, *33*(1), 62–71.

Sadorsky, P. (2021). A random forests approach to predicting clean energy stock prices. *Journal of Risk and Financial Management*, *14*(2), 48.

Soekamto, Y., Chandra, M., Wiradinata, T., Tanamal, R., & Saputri, T. R. D. (2023). *Property Category Prediction Model Using Random Forest Classifier to Improve Property Industry in Surabaya*.

Van Der Heijden, H. (2022). Predicting industry sectors from financial statements: An illustration of machine learning in accounting research. *The British Accounting Review*, *54*(5), 101096.

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia Computer Science*, *167*, 599–606.