Building of Informatics, Technology and Science (BITS)
Volume 99, No 99, Bulan 2023 Page: 999−999
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v9i9.999

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 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.