A Comparative Study of Feature Selection Techniques in Machine Learning for Predicting Stock Market Trends

Adi Suryaputra Paramita [1,*]; Shalomeira Valencia Winata [2]

[1,2] Information Systems Department, Universitas Ciputra Surabaya, Indonesia
[1] adi.suryaputra@ciputra.ac.id*;
* corresponding author

(Received June 4, 2023; Revised July 21, 2023; Accepted July 24, 2023; Available online August 15, 2023)

Abstract

This study aims to compare the effectiveness of three feature selection techniques, namely Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE), in predicting stock market conditions. This research uses three different datasets from Kaggle that contain stock market value prediction data. The results show that RFE performs better than PCA and IG in predicting market value with fairly precise accuracy. By using the RFE technique, this study was able to identify the most influential features in prediction, reduce the dimensionality of the data, and improve the performance of the prediction model. This provides significant benefits in the world of stocks, including improved investment decisions, reduced investment risk, improved trading strategy performance, and identification of promising investment opportunities. For future research, further comparative studies between other feature selection techniques can be conducted. This research has novelty in several aspects. First, it applies different feature selection techniques, namely Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE), in the context of stock market prediction. The use of these techniques to select the most relevant features in predicting stock market conditions provides a deeper understanding of the influence of these features on stock price movements. Furthermore, this research utilizes different datasets from Kaggle, which represent various stock market value predictions. The use of different datasets provides variation in the data and allows this research to examine the performance of feature selection techniques in various stock market contexts. In conclusion, this research provides insight into the effectiveness of feature selection techniques in stock market value prediction and provides guidance for market participants to improve investment decisions and trading performance in the stock market.

Keywords: Feature Selection, Principal Component Analysis, Information gain, Recursive Feature Elimination

1. Introduction

The stock market is a complex and ever-changing environment, where stock prices are influenced by various economic, political, social, and psychological factors [1], [2]. Predicting stock market trends is an interesting challenge in finance, as the ability to predict changes in stock prices can provide significant benefits to investors and stock traders [3], [4]. One approach that is widely used in predicting stock market trends is to use machine learning techniques. Machine learning allows computers to learn from existing stock market data and identify hidden patterns [5]. However, in the use of machine learning, proper feature selection is essential to improve prediction performance. Relevant and informative features can help in identifying patterns and trends hidden in the stock market data.

Selecting the right features in machine learning to predict stock market trends involves an in-depth analysis of various factors that affect stock price movements. These factors include company financial reports, economic and political news, market indicators, and other factors that may affect investor sentiment [5]-[7]. By selecting relevant features, machine learning models can learn patterns related to these factors and provide more accurate predictions.

In addition, in selecting the right features, it is also necessary to consider the relationship between the selected features. Some features may be interrelated or correlated, and in some cases, highly correlated features may provide similar information. Therefore, it is important to identify and select features that provide different and complementary information, so as to improve the accuracy of predictions [5].

In order to improve the prediction of stock market trends using machine learning, research and development continues to identify the most relevant and effective features. In addition, it is important to keep up with the latest developments in the stock market and update machine learning models regularly with new data. With the right approach and intelligent feature selection, machine learning can be a powerful tool in predicting stock market trends and provide significant benefits to market players.

The purpose of this study is to compare three commonly used feature selection techniques in machine learning, namely Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE), in the context of stock market trend prediction. PCA is used to reduce the dimensionality of features by keeping the largest variance in the data. IG is used to measure the information importance of each feature to the prediction class. RFE is used to iteratively eliminate features based on their role in improving the performance of the prediction model. While there have been previous studies on feature selection techniques in the context of stock market trend prediction, this research has some novelty. First, this study directly compares three different feature selection techniques, namely PCA, IG, and RFE, to see the difference in their performance and effectiveness in stock market trend prediction. Second, this study focuses on the comparison of feature selection techniques in stock market trend prediction, which can provide investors and stock traders with valuable insights into which approach is most effective in predicting stock price movements.

2. Software Defects and Data Mining

2.1. Stock Market and Machine Learning

The stock market and technology have a close relationship and influence each other in running and developing the stock market as a whole. Technology has been instrumental in facilitating the trading process, increasing accessibility, improving speed and efficiency, and providing market players with sophisticated analytical tools [5], [8]-[11]. Technology has changed the way stock trading is conducted. In the digital age, stock trading has moved from physical stock exchanges to electronic platforms. Electronic trading technologies such as Electronic Communication Networks (ECN) and Electronic Trading Platforms allow traders and investors to conduct transactions quickly and efficiently by executing orders electronically. This reduces reliance on slower human trading and allows for wider market access [12], [13]. Technology has also provided greater accessibility for investors. In the past, only large financial institutions and wealthy investors could actively participate in the stock market. However, with the advent of online trading platforms and mobile device applications, individuals can easily buy and sell stocks at a lower cost. This technology has reduced the barriers of entry to the stock market and given individuals the opportunity to invest and participate in the growth of the market.

Technology has provided great advancements in market analysis and research tools. With data analysis software and intelligent algorithms, analysts can collect, process, and analyze stock market data more quickly and efficiently [14], [15]. Technologies such as machine learning and artificial intelligence have enabled the development of better and more accurate prediction models. This helps investors and traders make better investment decisions based on in-depth analysis. Technology has also driven the development of global stock markets. Through widespread internet connectivity, investors and traders can easily access stock markets in different countries. This opens up wider investment opportunities and portfolio diversification. In addition, technology has also facilitated cross-border trading and integration of global stock markets, thereby expanding liquidity and improving market efficiency. Technology continues to play a role in innovations in the stock market. For example, the development of blockchain technology has provided the possibility to speed up and simplify the transaction settlement process and enhance security and transparency in shareholding. In addition, technologies such as big data, social media sentiment analysis and artificial intelligence continue to be adopted in an effort to improve stock market predictions and identify potential investment opportunities.

The relationship between the stock market and machine learning is very close because machine learning techniques can be used to predict stock market trends. The stock market is a complex and dynamic environment, where stock prices are influenced by many factors that are difficult to predict manually [16], [17]. Therefore, the use of machine learning is an attractive solution to analyze stock market data and produce more accurate predictions. Machine learning allows computers to learn from existing stock market data. By involving the right algorithms and models, machines can analyze historical data and identify hidden patterns. This allows investors and stock traders to better understand future stock price movements. Machine learning requires proper feature selection to improve prediction performance. In the context of the stock market, such features include company financial reports, economic and political news, market indicators, and other factors that may affect investor sentiment. By selecting relevant and informative features, machine learning models can identify patterns related to these factors and provide more accurate predictions.

Machine learning can utilize techniques such as regression, classification, and clustering to predict stock price movements. For example, using regression, machine learning models can identify the relationship between certain variables and future stock prices. Thus, investors and stock traders can use these predictions to make better investment decisions. Machine learning can also help in managing risk in the stock market. By analyzing historical data, machine learning models can identify patterns associated with certain risks. This allows investors to manage their portfolios more effectively, reducing risk and increasing potential returns. Developments in the field of machine learning also allow the use of more advanced techniques such as deep learning. Deep learning uses complex artificial neural networks to process and analyze stock market data in greater depth. This technique has demonstrated its ability to identify complex patterns and predict stock market trends with greater accuracy.

Overall, machine learning plays a significant role in analyzing and predicting stock market trends. By utilizing the ability of computers to learn patterns hidden in stock market data, investors and stock traders can make more informed investment decisions and earn greater profits.

2.2. Feature Selection

Feature Selection is the process of selecting the most relevant and informative subset of features from the set of features available in the data [6], [9]. The main objective of Feature Selection is to improve the performance of machine learning models by eliminating irrelevant or redundant features, thereby reducing data dimensionality and model complexity. Proper feature selection is very important in machine learning because it can affect the quality of predictions and computational efficiency [8]. According to experts, there are several reasons why Feature Selection is necessary:

1) Improving prediction accuracy: By selecting the most relevant features, the model can focus on the important information and ignore features that do not contribute significantly to the prediction results. This reduces the effect of noise and improves prediction accuracy.
2) Reducing overfitting: Overfitting occurs when the model is too complex and "memorizes" the training data without being able to generalize well to new data. By selecting relevant features, we can reduce model complexity and avoid overfitting, so that the model can provide more generalized and better predictions on new data.
3) Improve model interpretability: In some cases, having many features in a model can make it difficult to interpret. By performing feature selection, we can produce a simpler and more understandable model, making it easier to make decisions based on an understanding of the selected features.
4) Reduced computation time and cost: By reducing the dimensionality of the data through feature selection, the computational time required to train the model can be reduced. In addition, the use of relevant features can also reduce the need for large data processing, thereby reducing storage and processing costs.

2.2.1. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a technique used in multivariate data analysis to reduce the dimensionality of data by projecting the data into a lower space. The main objective of PCA is to identify a linear combination of the original features that explains the variation in the data with the least number of components or features [18], [19].

In the PCA process, the newly formed principal components are linear combinations of the original features. PCA calculates the covariance matrix between the features in the dataset. This covariance matrix describes the relationship between the features and a measure of how diverse they are. Next, PCA calculates the eigenvectors and eigenvalues of the covariance matrix. The eigenvector is a vector that describes the direction of the principal components, while the eigenvalue is a scalar that describes the relative importance of each principal component [20]-[22]. PCA sorts the eigenvectors based on their eigenvalues in descending order. Eigenvectors with higher eigenvalues reflect principal components that have a greater impact in explaining variations in the data. Thus, PCA allows us to select the principal components that are most significant in explaining the variation in the data [23], [24].

PCA projects data into a lower space using the eigenvectors associated with the selected principal components. This projection results in new data known as principal components. These principal components are a linear combination of the original features and describe the variation of the data in a lower dimension. PCA allows us to choose the number of principal components to use based on the eigenvalue. By choosing a lower number of components, we can reduce the dimensionality of the data, eliminate redundant or unimportant features, and still retain most of the variation in the data. This dimensionality reduction can lead to simpler interpretations, reduce computational requirements, and help in analyzing and visualizing multivariate data.

2.2.2. Information Gain (IG)

Information Gain (IG) is a concept in decision theory and machine learning used to measure how much information a feature provides in predicting a desired class or target [25]-[28]. IG is used to select the most informative or most influential features in separating and classifying data.

IG measures the difference in uncertainty or entropy before and after dividing the data by a feature. Entropy describes the level of uncertainty in the data. If a feature has a good ability to split data and reduce uncertainty, then its IG is high. If the IG is high, then the feature provides a lot of valuable new information in data classification [27]. IG is calculated by comparing the entropy before dividing the data with the entropy after dividing the data based on a particular feature. Entropy is measured by a mathematical formula that relates to the class distribution at each feature split. If the entropy after splitting becomes lower, it means that the feature has a high contribution in predicting the target class.

IG is often used in decision tree algorithms to select features that are used as splits in building the tree. The decision tree algorithm iteratively selects the feature with the highest IG at each step to split the data into increasingly homogeneous subsets within the target class. By selecting features that have a high IG, the algorithm can build a decision tree that is effective in predicting the target class. IG has some disadvantages, including the tendency to select features with a high number of values. This can lead to the selection of features that have a strong relationship with the target class, but have no direct correlation with the target class. In addition, IG is also prone to bias in unbalanced datasets, where the dominant target class may exert a greater influence on the IG calculation.

Overall, Information Gain is a measure used in decision making and machine learning to evaluate the importance or informativeness of a feature in predicting the target class. By using IG, we can select the most influential features in separating and classifying data with high accuracy.

2.2.3. Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a method in machine learning used to select the most influential features in predicting a target by iteratively removing the least influential features. The main goal of RFE is to reduce data dimensionality and improve model performance by using the most informative subset of features [29]-[32].

RFE starts by training a machine learning model on all the features available in the dataset. The model is used to measure the relative importance of each feature based on the coefficient or weight assigned by the model. Features with lower coefficients or weights are considered less influential in predicting the target. RFE removes the least influential features in the first iteration and retrains the model on the remaining subset of features. This process is repeated until the desired number of features is reached or until no more features can be removed. At each iteration, the model is re-evaluated and the features with the lowest contribution are removed.

RFE utilizes model evaluation metrics such as accuracy, precision, or area under the ROC curve to measure model performance at each iteration. By observing the change in model performance after removing features, RFE can identify features that contribute significantly to predicting the target. RFE allows the user to select the desired number of features or specify a threshold for the importance of features deemed relevant. By selecting the most influential subset of features, RFE can increase computational efficiency, reduce overfitting, and improve model interpretability.

Overall, Recursive Feature Elimination is a useful method in feature selection in machine learning. By iteratively eliminating less influential features, RFE helps in building more efficient and accurate models by retaining the subset of features that are most informative in predicting the target.

3. Methodology

3.1. Dataset Explanation

In this research, three types of datasets obtained from the Kaggle platform are used. The three datasets focus on stock market prediction, which includes historical data and stock prices.

First, the "AMZN, DPZ, BTC, NTFX adjusted May 2013-May2019" dataset contains historical data and stock prices of several companies, including Amazon (AMZN), Domino's Pizza (DPZ), Bitcoin (BTC), and Netflix (NTFX). This dataset covers the time span from May 2013 to May 2019. This data can be used for analysis and prediction of stock price movements of these companies.

Secondly, the "S&P 500 stock data" dataset contains data from the S&P 500 index, which includes major companies listed on the United States stock exchange. This dataset contains daily data on the 500 stocks listed in the index, including opening and closing prices, trading volume, and other related factors. This data can be used for analysis and prediction of overall stock market movements.

Third, the "Tesla Stock Price" dataset focuses on the company Tesla. This dataset contains historical data on Tesla's stock price over a period of time. This data includes opening, closing, trading volume, and other relevant factors. This dataset can be used to analyze and predict stock price movements specific to the company Tesla.

Using these three datasets, this research can perform stock market analysis and predictions involving specific companies (such as Amazon, Domino's Pizza, and Netflix), the overall stock market (via the S&P 500), and the company Tesla. This historical data and stock prices can provide useful insights and information in understanding stock market trends and patterns and help in developing more accurate prediction models.

3.2. Feature Selection Technique

Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE) techniques are three methods used in feature selection in machine learning. Here is the flow of the three techniques:

Principal Component Analysis (PCA) [33]:

1) Calculate the covariance matrix between the features in the dataset.
2) Calculate the eigenvector and eigenvalue of the covariance matrix.
3) Sort the eigenvectors by their eigenvalues in descending order.
4) Select the desired number of principal components based on the eigenvalue.
5) Project the data into a lower space using the eigenvectors associated with the selected principal components.

Information Gain (IG) [26]:

1) Calculate the entropy before and after dividing the data based on certain features.
2) Calculate IG by comparing the difference in entropy before and after data separation.
3) Choose the feature with the highest IG as the most informative feature.

Recursive Feature Elimination (RFE) [31]:

1) Train the machine learning model on all features in the dataset.
2) Calculate the relative importance of each feature based on the coefficients or weights given by the model.
3) Remove features with the lowest contribution.
4) Retrain the model on the remaining feature subset.
5) Repeat steps b and c until the desired number of features is reached or no more features can be deleted.

In this flow, PCA is used to reduce the dimensionality of the data by identifying linear combinations of the original features that explain the variation in the data with the least number of components. IG is used to measure the importance of features in predicting the target by comparing the entropy difference before and after splitting the data. RFE is used to select the most influential features by iteratively removing the features with the lowest contribution. Figure 1 is the flow of the feature selection test of this research.

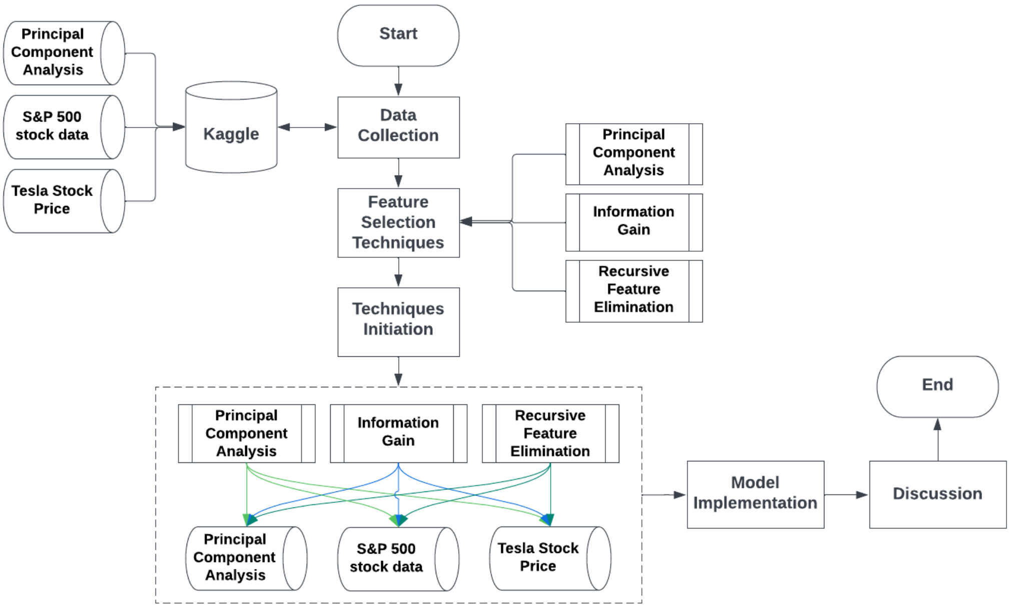
Principal Component Analysis
S&P 500 stock data
Tesla Stock Price
Kaggle
Start
Data Collection
Feature Selection Techniques
Principal Component Analysis
Information Gain
Recursive Feature Elimination
Techniques Initiation
Principal Component Analysis
Information Gain
Recursive Feature Elimination
Principal Component Analysis
S&P 500 stock data
Tesla Stock Price
Model Implementation
Discussion
End

Figure. 1. Proposed feature selection flow

3.4. Experimental Simulation

In this research, Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE) techniques will be compared using three different datasets. The use of different datasets for each technique is because each feature selection technique has different criteria and principles in selecting the most informative features. Table 1 and 2 is a summary of the techniques and data cases that will be tested in this research.

Table 1. Techniques Setup/Case

Experimental Label	Algorithm Used
X	Principal Component Analysis
Y	Information Gain
Z	Recursive Feature Elimination

Table 2. Dataset Setup/Case

Experimental Label	Algorithm Used
A	Principal Component AnalysisAMZN, DPZ, BTC, NTFX adjusted May 2013-May2019
B	S&P 500 stock data
C	Tesla Stock Price

First, the Principal Component Analysis (PCA) technique is used to reduce the dimensionality of the data by projecting the data into a lower space using a linear combination of the original features. PCA focuses on the variation of the data and looks for principal components that explain the variation with the least number of components. Therefore, the dataset used for PCA can either be a high-dimensional dataset or have features that are correlated with each other.

Second, the Information Gain (IG) technique is used to select the most influential features in predicting the target by comparing the entropy difference before and after splitting the data based on certain features. IG is more suitable for datasets that have target variables or classes to be predicted. Therefore, the dataset used for IG must include features that have a relationship or correlation with the desired target variable.

Third, the Recursive Feature Elimination (RFE) technique is used to select features by repeatedly removing the least influential features until it reaches the desired number of features or no more features can be removed. RFE has no specific criteria related to the type of dataset used. However, the dataset used must include the features to be selected and assessed for importance.

By using three different datasets for each technique, this research makes it possible to evaluate the performance and effectiveness of each technique in different contexts. Each technique has a unique approach in selecting the most informative features, and by using datasets that match the principles and criteria of each technique, it is expected to find better and optimized prediction results for each technique used.

4. Results and Discussion

4.1. Initial Experimental Result

The results showed that the Recursive Feature Elimination (RFE) technique proved to be the most superior in predicting stock market conditions. RFE is a feature selection method that iteratively removes the least influential features in predicting the target. Figure 2 below shows the accuracy performance of the three feature selection techniques on the three datasets.

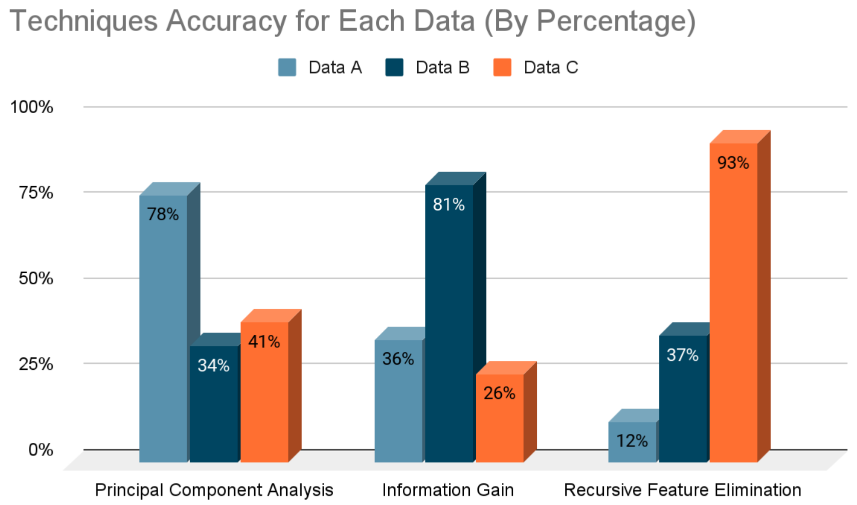

Figure 2. Techniques accuracy for each data.

This result can be explained by several factors:

1) RFE's ability to identify the most influential features: RFE systematically removes features that have the lowest contribution in predicting the target. Thus, RFE is able to filter out less relevant or uninformative features, thereby improving the quality of the remaining features and focusing attention on more important features.
2) Data dimensionality reduction: RFE effectively reduces the dimensionality of data by selecting the most influential subset of features. By reducing the number of features, RFE helps to overcome the "curse of dimensionality" problem and improve computational efficiency. In addition, dimensionality reduction can also help reduce overfitting and improve model generalization.
3) Improved model performance: By using RFE, the selected features have a more significant contribution in predicting stock market conditions. This can lead to improved model performance, such as increased accuracy, precision, recall, or other evaluation metrics. By retaining the most informative subset of features, RFE can help uncover patterns and trends hidden in stock market data.

In the context of predicting stock market conditions, RFE has the advantage of selecting the most relevant and informative features for prediction purposes. However, it is important to keep in mind that these results may vary depending on the dataset used, the machine learning algorithm chosen, and other factors. Therefore, it is important to conduct comprehensive evaluation and validation in selecting the appropriate feature selection method for each specific task and condition.

4.2. Model Implementation Result

Figure 3 below is the result of using the model with the RFE technique on the stock market dataset.

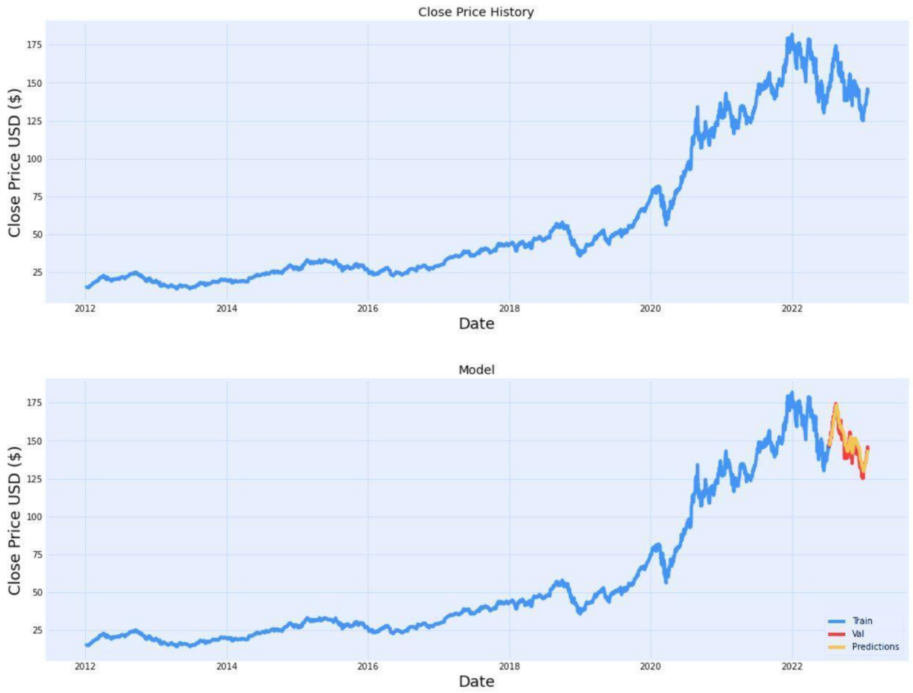

Figure 3. RFE implementation on single stock dataset.

The results of using the model with the RFE feature selection technique show that the model is able to predict market value with fairly precise accuracy. This can be explained by several factors:

Selection of more relevant features, RFE helps select a subset of features that are most influential in predicting stock market value. By focusing on the most informative features, the model has access to the most relevant information to make predictions. This can improve prediction accuracy and reduce "noise" or unimportant information in the dataset.

Reducing overfitting, RFE reduces the dimensionality of the data by selecting a subset of the most influential features. By reducing the number of features, the model is less likely to "memorize" the training data and tends to generalize better to data that has never been seen before. Thus, RFE can help reduce overfitting and improve prediction precision on new data.

Adaptability to recent cases, Models with RFE techniques have good adaptability to recent cases in the stock market. The features selected by RFE can reflect changes in trends and patterns that occur in the data. When there are changes in the stock market environment, such as changes in economic, political, or social factors, the model with RFE can identify and update the most relevant features to predict market value with high accuracy.

To prove some of these factors, Figure 4 below is the second test using the same model and technique but using a different dataset and output. It was found that using the scatterplot, the prediction results of the model for each stock had a fairly high accuracy, slightly lower than the first test but the accuracy was still appropriate to be able to prove the theory.

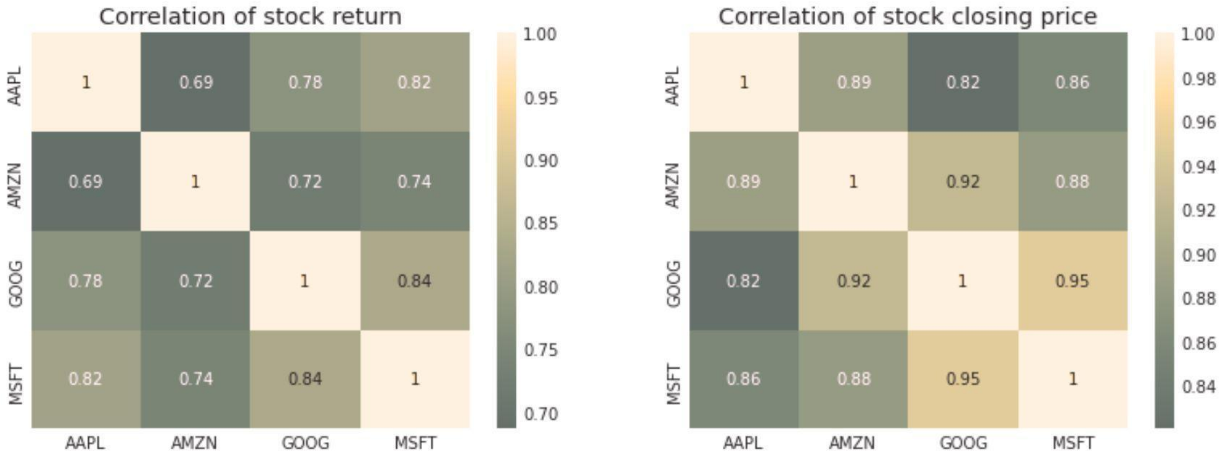

Figure 4. Scatterplot of RFE implementation on multiple stock features.

4.3. Discussion

By using feature selection techniques such as Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE), investors can identify the most influential features in predicting stock market conditions. This allows them to make more informed investment decisions based on more accurate analysis. Thus, the results of this study can help investors improve their investment decisions and optimize their portfolios. In the world of stocks, investment risk is significant. By using effective feature selection techniques, such as RFE, prediction models can be refined by focusing on the most informative features. By reducing the dimensionality of the data and selecting the most relevant features, the results of this study can help reduce investment risk by allowing investors to gain more accurate insights into stock market trends.

Trading strategies in the stock market often involve an in-depth analysis of relevant features. By using robust feature selection techniques, such as PCA or IG, traders can improve their strategies by selecting the features that are most influential in predicting stock price movements. The results of this research can help improve the performance of trading strategies by reducing confusion and gaining more detailed insights into market conditions. By using accurate feature selection techniques, the results can help identify promising investment opportunities. In a complex stock market, being able to identify trends and patterns hidden in the data is key to success. By utilizing effective feature selection techniques, investors can find potentially profitable investment opportunities and take the right steps in making investment decisions.

The results of this study can also contribute to the development of better prediction models in the stock industry. By comparing and analyzing different feature selection techniques, this research provides insight into which techniques are most effective and can provide more accurate prediction results. This can encourage further research and innovation in the development of stock market prediction models. Improvements in investment decisions, better risk management, and the development of more accurate prediction models can contribute to better economic and financial growth. By utilizing the results of this research, stock market participants, both investors and traders, can optimize their returns, reduce associated risks, and improve overall market performance. In turn, this could lead to more stable and sustainable economic growth in the financial sector.

5. Conclusion

This study investigates the use of feature selection techniques, namely Principal Component Analysis (PCA), Information Gain (IG), and Recursive Feature Elimination (RFE), in predicting stock market conditions. The results show that RFE is the most superior technique in predicting market value with fairly precise accuracy. The use of RFE helps to identify the most influential features in prediction, reduce the data dimension, and improve the performance of the prediction model.

The results of this study have important implications in the world of stocks. The use of RFE can improve investment decisions by selecting relevant features, reducing investment risk, and improving trading strategy performance. In addition, the results also help in the identification of promising investment opportunities and the development of better prediction models. By applying these findings, stock market participants can optimize their returns, reduce risks, and contribute to better economic and financial growth.

This study has several limitations that need to be noted. First, this study used three different datasets, which may affect the results and generalizability of the findings. In addition, the feature selection technique used in this study may also be affected by other factors such as the type of machine learning algorithm used or the parameters set. Furthermore, this research focuses on predicting stock market conditions, but does not consider external factors that may affect the stock market such as political events or policy changes.

To complement this study, further research can be conducted. First, further comparative studies between other feature selection techniques can be conducted to gain a deeper understanding of their relative performance in stock market condition prediction. Furthermore, research can consider external factors in the analysis, such as market sentiment or financial news, to improve the accuracy of the predictions. In addition, research could also consider using more advanced feature selection techniques or a combination of various techniques to achieve better prediction performance. Finally, research could involve testing and validating the prediction model using real-time data to evaluate the generalization ability of the model in changing market situations.

References

[1] Z. Wang, "Research on the selection of stock prediction model features for long-term stock market trends based on K-nearest neighbor algorithms," in*Proceedings - 2022 International Conference on Data Analytics, Computing and Artificial Intelligence, ICDACAI 2022*, vol. 1, no. 1, pp. 38-42, 2022.

[2] Goyal, J. S. Challa, S. Shrivastava, and N. Goyal, "Anytime Frequent Itemset Mining of Transactional Data Streams,"*Big Data Res.*, vol. 21, no. 1, pp. 100146-67, 2020.

[3] Araújo, A. Pereira, and F. Benevenuto, "A comparative study of machine translation for multilingual sentence-level sentiment analysis,"*Inf. Sci. (Ny).*, vol. 512, no. 1, pp. 1078-1102, 2020.

[4] Park and J. Kwon Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,"*Expert Syst. Appl.*, vol. 42, no. 6, pp. 2928-2934, 2015,

[5] Aslam, K. S. Mughal, A. Ali, and Y. T. Mohmand, "Forecasting Islamic securities index using artificial neural networks: performance evaluation of technical indicators,"*J. Econ. Adm. Sci.*, vol. 37, no. 2, pp. 253-271, 2021.

[6] Figueira, "Predicting grades by principal component analysis: A data mining approach to learning analytics,"*Proc. - IEEE 16th Int. Conf. Adv. Learn. Technol. ICALT 2016*, vol. 1, no. 1, pp. 465-467, 2016.

[7] Doherty, Trevor, *et al.*, "A comparison of feature selection methodologies and learning algorithms in the development of a DNA methylation-based telomere length estimator," *BMC Bioinformatics*, vol. 24, no. 1, pp. 178-208, 2023,

[8] E.Ahmadi, A. Garcia-Arce, D. T. Masel, E. Reich, J. Puckey, and R. Maff, "A metaheuristic-based stacking model for predicting the risk of patient no-show and late cancellation for neurology appointments,"*IISE Trans. Healthc. Syst. Eng.*, vol. 9, no. 3, pp. 272-291, 2019.

[9] Bertolini, S. J. Finch, and R. H. Nehm, "Quantifying variability in predictions of student performance: Examining the impact of bootstrap resampling in data pipelines,"*Comput. Educ. Artif. Intell.*, vol. 3, no.1, p. 100067, 2022.

[10] N. Chi, "Modeling and Forecasting Long-Term Records of Mean Sea Level at Grand Isle, Louisiana: SARIMA, NARNN, and Mixed SARIMA-NARNN Models,"*J. Appl. Data Sci.*, vol. 2, no. 2, pp. 1-13, 2021.

[11] T.Wahyuningsih, "Problems, Challenges, and Opportunities of Visualization on Big Data,"*J. Appl. Data Sci.*, vol. 1, no. 1, pp. 20-28, 2020.

[12] L. An, "Research on Short Video Publishing Algorithm and Recommendation Mechanism Based on Artificial Intelligence,"*J. Appl. Data Sci.*, vol. 3, no. 2, pp. 66-71, 2022.

[13] L .Ran, "Development of Computer Intelligent Control System Based on Modbus and WEB Technology,"*J. Appl. Data Sci.*, vol. 4, no. 1, pp. 15-21, 2023.

[14] Y. Kaya and T. Dumitras, "When Does Data Augmentation Help With Membership Inference Attacks?,"*Int. Conf. Mach. Learn.*,vol. 1, no.1, pp. 5345-5355, 2021.

[15] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, "Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers,"*Int. J. Mach. Learn. Cybern.*, vol. 1, no. 1,

pp. 1-16, 2022.

[16] K. Venkateswararao and B. V. R. Reddy, "LT-SMF: long term stock market price trend prediction using optimal hybrid machine learning technique,"*Artif. Intell. Rev.*, vol. 56, no. 6, pp. 5365-5402, 2023.

[17] X. Yuan, J. Yuan, T. Jiang, and Q. U. Ain, "Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market,"*IEEE Access*, vol. 8, no. 1, pp. 22672-22685, 2020.

[18] T. D. Phan, "Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia,"*Proc. - Int. Conf. Mach. Learn. Data Eng. iCMLDE 2018*, vol. 1, no. 1, pp. 8-13, 2019.

[19] Y.O. Fourar, M. Djebabra, W. Benhassine, and L. Boubaker, "Contribution of PCA/K-means methods to the mixed assessment of patient safety culture,"*Int. J. Heal. Gov.*, vol. 26, no. 2, pp. 150-164, Jan. 2021.

[20] K. P. Yang, "Cyber Democracy Versus Controlling Shareholders: The Implications of E-Voting System for Corporate Governance,"*IJIIS Int. J. Informatics Inf. Syst.*, vol. 2, no. 3, pp. 136-142, 2019.

[21] J. Zeng and S. Nantida, "Study on the Ideological and Political Practice Teaching of College Students Based on the Internet + Technology,"*IJIIS Int. J. Informatics Inf. Syst.*, vol. 6, no. 1, pp. 24-30, 2023.

[22] R. Fitriana, "Quality Analysis of Vaberaya.Banyumaskab.Go.Id Website on User Satisfaction on Vaccination Activities in Banyumas District Using Delone and Mclean Success Model,"*IJIIS Int. J. Informatics Inf. Syst.*, vol. 5, no. 1, pp. 16-24, 2022.

[23] M.S. Ejaz, M. R. Islam, M. Sifatullah, and A. Sarker, "Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition,"*1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019*, vol. 1, no. 1, pp. 1-5, 2019.

[24] N.A. Prabowo, "Social Network Analysis for User Interaction Analysis on Social Media Regarding E-Commerce Business,"*IJIIS Int. J. Informatics Inf. Syst.*, vol. 4, no. 2, pp. 95-102, 2021.

[25] O. Alqaryouti, N. Siyam, A. Abdel Monem, and K. Shaalan, "Aspect-based sentiment analysis using smart government review data,"*Appl. Opt. Comput. Informatics*, vol. 1, no. 1, pp. 1-20, 2020.

[26] B.F. Tanyu, A. Abbaspour, Y. Alimohammadlou, and G. Tecuci, "Landslide susceptibility analyses using Random Forest, C4.5, and C5.0 with balanced and unbalanced datasets,"*Catena,* vol. 203, no. 1, pp. 1-14, 2021.

[27] X. Pu, F. Chan, and A. Chong, "The Influence of Supply Chain Relationships on the Adoption of Open Standards Inter-Organizational Information Systems: A Conceptual Framework,"*Int. J. Appl. Inf. Manag.*, vol. 1, no. 3 SE-Articles, pp. 91-98, 2021.

[28] M.S. Alomari, "The Legal System for the Conversion of Commercial Companies in the Light of the Rules of the Saudi Corporate System,"*Int. J. Appl. Inf. Manag.*, vol. 2, no. 4, pp. 106-111, 2022.

[29] L. Diao, D. Niu, Z. Zang, and C. Chen, "Short-term weather forecast based on wavelet denoising and catboost,"*Chinese Control Conf. CCC*, vol. 2019-July, no. 2018, pp. 3760-3764, 2019.

[30] S. Hidayat, M. Matsuoka, S. Baja, and D. A. Rampisela, "Object-based image analysis for sago palm classification: The most important features from high-resolution satellite imagery,"*Remote Sens.*, vol. 10, no. 8, 2018.

[31] N.S. M. Nafis and S. Awang, "An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification,"*IEEE Access*, vol. 9, pp. 52177-52192, 2021.

[32] I. Nordat, B. Tola, and M. Yasin, "The Effect of Work Motivation and Perception of College Support on Organizational Commitment and Organizational Citizenship Behavior in BKPSDM, Tangerang District,"*Int. J. Appl. Inf. Manag.*, vol. 2, no. 3, pp. 37-46, 2022.

[33] B.M. Faria, L. P. Reis, N. Lau, and G. Castillo, "Machine Learning algorithms applied to the classification of robotic soccer formations and opponentteams," *2010 IEEE Conf. Cybern. Intell. Syst. CIS 2010*, vol. 1, no. 1, pp. 344-349, 2010.