p-ISSN 2622-3082
e-ISSN 2622-3090

STOCK PRICE MOVEMENT PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHM: The Konstanz Information Miner (KNIME)

Dina Anggraeni[1*], Kris Sugiyanto[2], M. Irwan Zam Zam[3], Harry Patria[4]
[1]Department of Accounting, Universitas Indonesia, Jakarta
[2]Department of Accounting, Universitas Indonesia, Jakarta
[3]Department of Accounting, Universitas Indonesia, Jakarta
[4]School of Business and Management, Institut Teknologi Bandung, Bandung
[*]dina.anggraeni@ui.ac.id

ABSTRACT

What happened to the money that was invested? Especially in light of the obscene wealth. Obviously, it refers to financial instruments. Let us start with the most basic investment instruments: firm stocks traded on stock exchanges. We have gathered field data from 30 manufacturing companies over the last five years (2015 –2019), including financial and non-financial data, as well as stock price variance over each year. The objective of our research is to find the best models for predicting the movement of the stock price in a given year based on the parameters provided using KNIME. The finding in our research is that Support Vector Machine (SVM) with 90.7% perform better than Naïve Bayes with 80% followed by Decision Tree, Tree Ensambles, and Random Forest with 75%, 74%, and 74%, respectively, in terms of classification accuracy. From our research, we expect the algorithms to be able to predict which company will return a capital gain for the investor.

Keywords: KNIME, Earning Per Share, Customer Focus, Sustainability, R&D Intensity, Corporate Governance, Organizational Culture, Asset Size, Stock Price, Decision Tree, Naïve Bayes, SVM, Random Forest, and Tree Ensemble

Jurnal Akun Nabelo:
Jurnal Akuntansi Netral, Akuntabel, Objektif
Volume 4/Nomor 2/Januari 2022
Jurusan Akuntansi FEB Universitas Tadulako

A. INTRODUCTION

The stock market is a fascinating subject to research these days. Despite its inherent volatility and unpredictability, the stock market may be a good way for investors to grow their money. Many investors have tried to uncover the easiest technique to predict the stock price movement in order to decide their next move. If investors know which company has a higher chance of increasing its stock price, they may maximize their profit while minimizing their risk. The primary research from Ravikumar and Saraf, for example, uses past data to forecast the stock prices of numerous companies. As a result of this phenomena, we are thrilled to consider machine learning as a feasible method of predicting stock price movement.

Provost and Fawcett (2013) state that Machine Learning emerged as an area of Artificial Intelligence, concerned with methods for enhancing an intelligent agent's knowledge or performance over time, in response to the agent's experience in the environment. This sort of change regularly involves dissecting information from the environment and making expectations around obscure amounts, and the information examination portion of machine learning has become progressively critical within the industry over time. Provost and Fawcett (2013) also state Machine learning approaches have the capacity to uncover patterns and insights, and help to predict the stock price movement.

The objective of our study is to identify the most effective algorithm to predict stock price movement classification (up or down trend) that will assist the investors to manage the risk and return in their investment. We used data from manufacturing companies throughout a five-year period (2015 – 2019) in this research. We considered seven variables: earning per share, customer focus, sustainability, R&D intensity, corporate governance, organizational culture, asset size. In which, the data are concatenated and pre-processed using filtering and stemming. Finally, measure the classification accuracy using Decision Tree, Naïve Bayes, SVM, Random Forest, and Tree Ensemble to assess classification accuracy. This study's contribution is to help readers comprehend the workflow of document classification using KNIME.

This paper is organized as follows: In Introduction, the background knowledge required to comprehend the categorization problem is described in the introduction, along with the benefits and limitations of the current study work. In Literature review, previous research work is examined along with the influence of current research activity. The work flow created with KNIME is described in depth in Methodology, along with all of the Nodes utilized in KNIME. As a result, the performance of Machine Learning Algorithms is compared, and a discussion of our research findings is given.

B. LITERATURE REVIEW

The Konstanz Information Miner (KNIME) is a modular computational environment that allows for simple visual assembly, interactive data analysis, and data processing. Feltrin (2016) explain that it's a predictive analytics open-source data tool. Built on top of the Eclipse Integrated Development Environment (IDE), KNIME is a modular visual software environment. Each node uses an algorithm to process data and can interact with other nodes, allowing for the creation and recording of sophisticated data processing workflows. KNIME is a powerful data integration and predictive analytics platform. The program is especially useful for documenting complex steps including pre-processing, statistical analysis, statistical modelling, and predictive analytics. Another benefit is its open-source collaborative ecosystem, which allows users to create new algorithms, tools, data manipulation, and visualization techniques.

KNIME nodes are available in a wide range of functionality; nodes are classified as follows:

- I/O: extracts information from files or databases and exports it in a variety of formats.
- Data Manipulation: pre-processes input data with filtering, grouping, pivoting, binning, normalization, aggregating, joining, sampling, partitioning, and other operations including missing data replacement.
- Views: allows for interactive data exploration by visualizing data and results in a variety of interactive views.
- Mining: performs common pattern recognition, dimensionality reduction, and feature reconstruction tasks using state-of-the-art data mining methods such as clustering, rule induction, decision tree, association rules, nave Bayes, neural networks, and support vector machines, to mention a few. Feltrin (2016)

We employed five supervised learning algorithms in our project: Decision Tree, Naïve Bayes, SVM, Random Forest, and Tree Ensemble.

1. Decision Tree

Decision Tree is one of the supervised learning algorithms that can be used for classification and regression problems. The decision tree approach is a robust statistical tool for categorization, prediction, interpretation, and data manipulation, as discussed in [3], and it has various potential uses in medical research. The following are some of the benefits of using decision tree models to summarize study findings:

1) Divides original input variables into important subgroups, simplifying complex interactions between input variables and target variables.
2) Simple to understand and interpret.
3) A non-parametric technique that makes no assumptions about distributions.
4) It's simple to deal with missing values without resorting to imputation.
5) Handles large amounts of skewed data without the requirement for data manipulation.
6) Resistant to outliers

The decision tree method, like all analytical tools, has limitations that users should be aware of. The key drawback is that it is susceptible to overfitting and underfitting, especially when working with tiny data sets. The generalizability and robustness of the resulting models may be limited as a result of this issue. Another issue is that significant correlations between potential input variables can lead to the selection of variables that increase model statistics but are not causally related to the desired outcome. As a result, it is required cautions when evaluating decision tree models and applying the outcomes of these models to develop causal hypotheses.

2. Naïve Bayes

Although Naive Bayes is a simplistic classifier, it nonetheless considers all feature evidence. In terms of storage space and calculation time, it is extremely efficient. As each example is seen, training consists solely of storing counts of classes and feature occurrences. As previously stated, p(c) can be calculated by counting the proportion of class c examples among all examples The proportion of examples in class c for which characteristic ei appears can be used to calculate p (ei|c). Provost and Fawcett (2013)

3. Support Vector Machine (SVM)

Support vector machines are linear discriminants. SVMs, like linear discriminants in general, classify cases based on a linear function of the features. Instead of thinking about separating the classes with a line, SVMs choose based on a simple, elegant idea: fit the fattest bar between the classes first. The SVM's objective function is based on the idea that a broader bar is better. The linear discriminant will then be the centre line through the bar once the widest bar has been found. The margin around the linear discriminant is the distance between the dashed parallel lines, and maximizing the margin is the goal. Provost and Fawcett (2013)

4. Random Forest

Denil and Freites (2014) stated Random Forest is an ensemble method for predicting outcomes by averaging the results of numerous separate base models. The predictions of numerous trees, each of which has been trained separately, are combined to form random forests. Unlike boosting, where the base models are trained independently and then combined using a complicated weighting system, the trees are often trained independently and then mixed using averaging. When building a random tree, there are three basic options to consider. These are (1) the leaf splitting method, (2) the type of predictor to utilize in each leaf, and (3) the tree randomness injection strategy.

5. Tree Ensemble

By integrating different models, ensemble learning improves machine learning results. When compared to a single model, this method offers for improved predictive performance. The basic concept is to learn a group of classifiers (experts) and then allow them vote. The advantage of an ensemble classifier is improved predicted accuracy, but the disadvantage is that an ensemble of classifiers is difficult to grasp. (www.geeksforgeeks.org)

ROC CURVE

Receiver Operating Characteristics (ROC) graph is a two-dimensional plot of a classifier with false positive rate on the x axis against true positive rate on the y axis. As such, a ROC graph depicts relative trade-offs that a classifier makes between benefits (true positives) and costs (false positives) (Provost and Fawcett, 2013).

1. Earning Per Share (EPS)

It is explained by Al-Othman (2019) EPS is the best index of the real price of a stock and the most common measure because it shows the share of each stock of the company's earnings after tax. Therefore, this ratio is one of the most important figures that investors look for in the financial statements. EPS is clearly calculated in financial statements using the following formula:

Earning Per Share = (Net Income after tax – dividends of preferred stocks)/
of common stocks.

2. Customer Focus (CF)

Adopted from previous research by Nwokah (2009), CF in this study to look at the long-term aspects of the company's value. Customers will have an ever-increasing worth for a company that draws attention to customers since they will become loyal to the company's products and services.

3. Sustainability (S)

Sustainability disclosure assumes a nature-effective approach. interpretative by analysing annual reports, SR, or websites company. The disclosure of sustainability in this study is based on the degree to which the information fits qualitative

features, with research instruments evaluated 1 (one) if performed by the company and 0 (zero) if not.

4. Research & Development (R&D) Intensity

R&D activity is a key factor that enables the company's achievement of growth in the future because it ultimately leads to an increase in company knowledge, better technological capacity and products, More innovative services and processes. R&D Intensity is formulated as total R&D expenses divided by total sales.

5. Corporate Governance (CG)

The Corporate Governance Score is used to assess CG. It uses a dichotomous method to content analysis, in which each CG item in the research instrument is assigned a score of 1 (one) if it is carried out by the company and 0 (zero) if it is not. The Board of Directors Index looks at the autonomy, structure, and effectiveness of boards of directors. The percentage of board independence is used to measure autonomy; attendance at meetings is used to measure effectiveness; and separation of authority, including board size, is used to measure structure (BOD Size).

6. Organizational Culture (OC)

The organizational culture (OC) is one of the most extensively used OC models in empirical research. Clan, Adhocracy, Market, and Hierarchy are Cameron and Quinn's classifications. By studying the Annual Report, Sustainability Report (SR), and Website of the manufacturing company, OC Disclosure takes an interpretative approach.

7. Assets Size

The assets size is calculated by Logarithm of Total Assets, adopting research conducted by Husna & Satria, (2019).

C. RESEARCH METHOD

To predict the stock price movement, we have taken data-sets from 30 manufacturing issuers of listed Company on Indonesian Stock Exchange and we used seven variables: Earning Per Share (EPS), Customer Focus, Sustainability, R&D Intensity, Corporate Governance, Organizational Culture, Asset Size.

Table 1
List of Companies

No	Code	No	Code	No	Code
1	HMSP	12	DVLA	23	SRSN
2	KINO	13	IGAR	24	STTP
3	LION	14	INAF	25	TFCO
4	MYOR	15	KAEF	26	TSPC
5	SRIL	16	KLBF	27	ULTJ
6	STTP	17	MERK	28	UNVR
7	ULTJ	18	MYOR	29	WIIM
8	UNVR	19	NIKL	30	WTON
9	AISA	20	PYFA		
10	AUTO	21	ROTI		
11	CEKA	22	SMBR		

In this research there are four steps to be done: Data reading, Data modelling, Data Evaluation and Visualization, and Deployment.

C.1. Data Reading

In effort to develop robust and accurate predictive analytics using KNIME machine learning capability, we initially take data from actual research on many local Indonesian companies listed on the Indonesia Stock Exchange (IDX) from 2015 to 2019. The data itself comprises of financial and non-financial information.
Based on the 5 years worth knowledge, we intend to provide accurate prediction on how would the stock price variance be in 2020. The result will give more confident for investments to put their money on the "positive" companies.
After load the completed data to KNIME workflow, we tested the association using Linear Correlation and then cleansed the data using Missing Value after loading the completed data into the KNIME workflow.

The most important step after loading the data is determining the partition, in which the Machine Learning will learn the input data patterns from. In this project we use 80% of the data for learning, and the rest (20%) for method testing based on Pareto Principle.

The data reading process is shown in the picture 3.1 below:

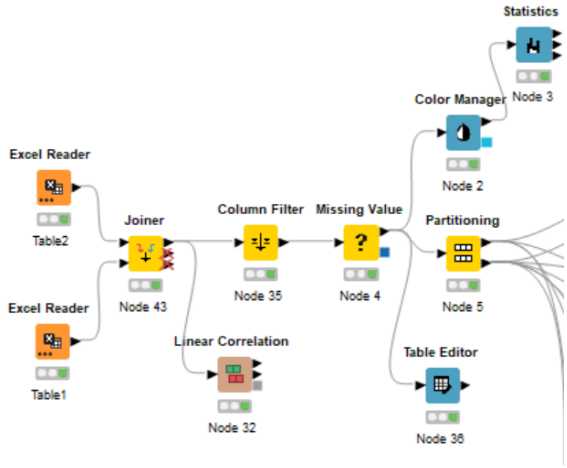


Figure 3.1
Data Reading

C.2. Data Modelling

After setting the partitioning, the next step would be selecting several Learner Models with which the data would be learned and predicted in the KNIME workflow. Below is the list of 5 Learner Models we selected in this project:

C.2.1. Decision Tree

This node creates a classification decision tree in main memory. The nominal attribute must be used as the target attribute. As the target attribute, the nominal attribute must be utilized. The other decision-making characteristics can be numerical or nominal. Numerical splits are always binary (two results) at a given split point, splitting the domain into two partitions. Nominal splits can be binary

(two outcomes) or multi-outcome (as many as nominal values). In the case of a binary split, the nominal values are divided into two categories.

The gini index and the gain ratio are two quality measurements provided by the algorithm for split calculation. A post-pruning procedure is also available to lower the tree's size and improve prediction accuracy. The pruning method is based on the minimum description length principle. The algorithm can be run in multiple threads, and thus, exploit multiple processors or cores. In this project, we use CLASS as class column, Gain ratio as quality measure, and no pruning for pruning method.

C.2.2. Naïve Bayes

The node creates a Bayesian model using the training data. It determines the Gaussian distribution for numerical attributes and predicts the number of rows per attribute value per class for nominal features. Using the naive Bayes predictor, the developed model might be utilized to predict the class membership of unclassified data.

In this project we used CLASS as the classification column, default probability is 0.0001 with 0.0001 minimum standard deviation, and 0.0 threshold standard deviation.

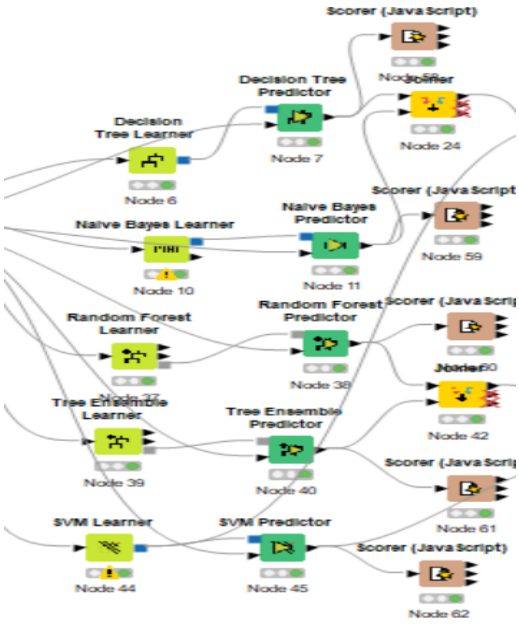


Figure 3.2
Data Modelling

C.2.3. Random Forest

Learns a random forest consist of a predetermined number of decision trees. Each of the decision tree models uses a distinct set of rows (records) and a different set of columns (describing attributes) for each split within the tree. Each decision tree's row sets are generated by bootstrapping and are the same size as the original input table. In a decision tree, the attribute set for an individual split is decided by selecting sqrt(m) attributes at random from the available attributes, where m is the total number of learning columns. The attributes can also be provided as bit (fingerprint), byte, or double vector. The output model describes a random forest and is applied in the corresponding predictor node.

C.2.4. Tree Ensemble

Learns a set of decision trees (for example, random forest* variations). Each tree is built with an unique set of rows (records) and/or columns (attributes). For further information, look at the Data Sampling and Attribute Sampling options. Bit (fingerprint), byte, or double vector attributes are also possible. The output model describes an ensemble of decision tree models and is applied to the associated predictor node, aggregating the votes of the individual decision trees using the selected aggregation mode.

C.2.5. SVM

This node trains a support vector machine on the input data. It supports a number of different kernels (HyperTangent, Polynomial and RBF). The SVM learner supports multiple class problems as well (by computing the hyperplane between each class and the rest), but note that this will increase the runtime.

Each learner model will be applied and its accuracy is measured using ROC. The model which provided the highest ROC will be selected for deployment.

C.3. Data Evaluation & Visualization

This step comprises of the ROC evaluation result of the selected learner and provide visual information.

C.4. Deployment

The last and final step in the workflow is the deployment. The selected learner model is applied on the target data of 2020, and expected to predict the variance of stock price for each designated company.

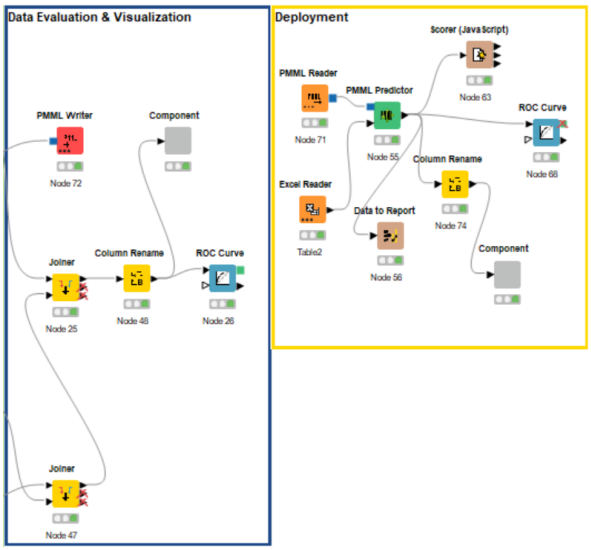


Figure 3.3
Data Evaluation & Visualization; Deployment

D. RESULT

We analysed the ROC Curves of the five machine learning models for measuring the classification of stock price changes as can be seen in figure 4.1, where we obtained a precision True Positive Rate (TPR) for 3 models with the highest precision rate being Support Vector Machine (SVM), Naïve Bayes (NB) and Decision Tree (DT) with the highest precision value is SVM followed by NB and DT.

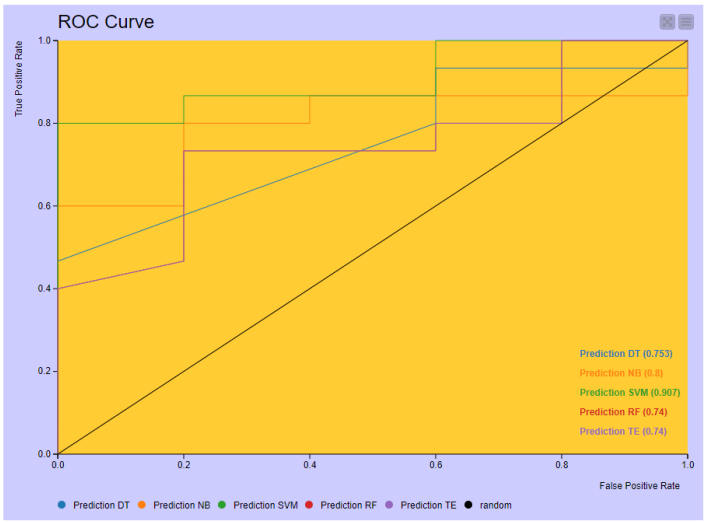


Figure 4.1
ROC Curves

The confusion matrix evaluation parameter for the SVM model that we use in our experimental machine learning predictions can be seen in Figure 4.2. Where we obtained a precision result of 75% for the measurement of the classification according to changes in stock prices as we have mentioned in the methodology section, we believe the results of this measurement have reflected the best results from several machine learning models that we have made and we hope that our research can contribute to the reader's understanding of the relationship between independent variables and the classification of changes in stock prices.

Scorer View

Support Vector Machine Predictor

Confusion Matrix

Rows Number : 20	Down (Predicted)	Up (Predicted)
Down (Actual)	0	5
Up (Actual)	0	15
	NaN%	75.00%

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
Down	0	0	15	5	0.00%	0.00%	0.00%	100.00%	0.00%
Up	15	5	0	0	100.00%	75.00%	100.00%	0.00%	85.71%

Overall Statistics

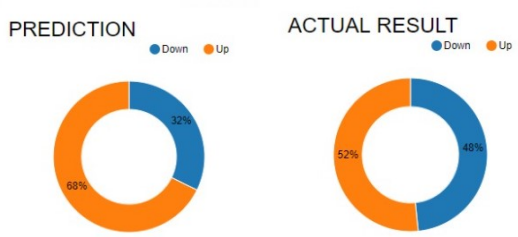
Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
75.00%	25.00%	0.000	15	5

Figure 4.2
Score View

Based on these results, we chose the SVM machine learning model as a predictive model for stock price changes in a sample of manufacturing companies with independent factors such as Earning Per Share (EPS), Customer Focus, Sustainability, R&D Intensity, Corporate Governance, Organizational Culture, Asset Size.

After experimenting with the 2020 data, we discovered a discrepancy between the prediction and the actual results. However, after studying circumstances that

occurred in 2020, we came to the conclusion that this occurred as a result of the Covid-19 pandemic, which resulted in a slew of unforeseen events for businesses. Furthermore, while the models and variables we employ will reflect the normal conditions, there will most likely be a significant divergence between prediction and actual if any unforeseen circumstances arise.



Graph 4.1
Prediction & Actual Result

E. CONCLUSION

In comparison to the other four models (DT, NB, RF, and TE), our research suggests that the SVM model is the best supervised machine learning technique for producing classification predictions. Our findings demonstrate that SVM achieves a score of 90.7 percent, which is higher than the other two machine learning models that achieve the closest results, NB at 80 percent and DT at 75.3 percent.

The hurdles for our current study include more testing of new datasets under extreme situations, like as the Covid-19 pandemic, which may yield different precision outcomes; nonetheless, we are confident that the SVM method will deliver better results than other models even in extreme conditions.

In our opinion, stock price movement is influenced by both historical and future influences. In this research, we solely use historical variables. As a result, some of the predictions are incorrect. We recommend that next researchers include future variables and unusual characteristics like Covid-19 in their studies. So that it can better represent the stock price movement.

REFERENCES

Provost, F and Tom Fawcett. (2013). Data Science for Business. O'Reilly Media Inc.

Feltrin. L. (2016). KNIME an Open Source Solution for Predictive Analytics in the Geosciences. Department of Earth Sciences, University of Western Ontario, 28-29

Song, Y and Lu, Y. (2015). Decision tree methods: applications for classification and prediction, Shanghai Arch Psychiatry, 133-134

Denil, M., Matheson, D., and Nando de Freites. (2014). Narrowing the Gap: Random Forests In Theory and In Practic. University of Oxford and University of Coulmbia

Ensemble Classifier | Data Mining - GeeksforGeek. accessed on September 19, 2021

Leqaa Al-Othman (2019). Income smoothing in banks and insurance companies and its impact on earnings per share – evidence from Jordan. Banks and Bank Systems, 14(4), 126-132. doi:10.21511/bbs.14(4).2019.12

N. Gladson Nwokah. (2009).Customer-focus, competitor-focus and marketing performance, Measuring Business Excellence, Vol. 13

Husna, A. and Satria, I. (2019). Effects of Return on Asset, Debt to Asset Ratio, Current Ratio, Firm Size, and Dividend Payout Ratio on Firm Value. University Persada of Indonesia

Ravikumar, Srinath and Prasad Saraf. (2020). Prediction of Stock Price using Machine Learning (Regression, Classification) Algorithms. Vishwakarma Institute of Technology. India

KNIME hub. https://hub.knime.com. accessed on September 19, 2021.