A Hidden Markov Model Approach to Daily Stock Return Dynamics of PT. Kimia Farma Tbk.

ABSTRACT

|  |
| --- |
| PT Kimia Farma Tbk is one of the enterprises owned by the government of Indonesia that is engaged in the pharmaceutical sector. The daily stock return price of PT Kimia Farma Tbk (KAEF) from 2012 to 2024 is constantly changing. These changes can be influenced by one of the conditions of the stock market that cannot be observed directly. In addition, this change moves up and down drastically, so stock price predictions cannot use just one model. In this study, a hidden Markov model (HMM) was chosen for modeling this data because it allowed changes from one model to another. Initially, a 3-state discrete HMM was used to model the data. The simulation showed that the model could not accurately follow the movement of the data. To overcome this, the discrete HMM was modified into a continuous HMM by replacing the parameters of the probability mass function with the probability density function of the continuous distribution for each hidden state. So, the daily stock return price of KAEF was modeled by a 3-state continuous HMM, and the data in each state follows a logistic distribution. To measure the accuracy of using the continuous HMM, we used the mean absolute error (MAE). The MAE value obtained from the training data simulation using discrete MHM is 0.03188 or 7.97% of the original data range, while the MAE from the training data simulation using continuous HMM is 0.029818 or 7.45% of the original data range. Furthermore, the MAE value of the test data simulation is 0.0338527 or 8.46% of the original data range. The results show that the HMM performs very well in modeling the dynamics of the KAEF data. |

*Keywords: Stock Return, Hidden Markov model, Mean Absolute Error, Continuous Hidden Markov Model*

1. INTRODUCTION

Stock price predictions are very much needed by investors to make decisions about buying or selling shares. The risk of falling prices (capital loss) and the risk of liquidation of the company issuing shares are some risks faced when buying shares (Sawidji, 2012). Shares considered good are shares that can provide a realized return that is not too far from the expected return (Husnan, 2005). Return is the result obtained from an investment (Jogiyanto, 2009). A positive return indicates that the investment results are profitable, and vice versa. A negative return indicates that the investment results are losing money. This decision-making depends on the pattern of the stock price movement (Troiano et. al. 2018). Studies related to predicting stock prices have been widely conducted. Cahyani and Mahyuni (2020) used the Moving Average model (MA) to identify the trend of LQ45 stock price movements. Furthermore, Rusyida and Pratama (2020) used the Autoregressive Integrated Moving Average (ARIMA) to predict non-stationary Garuda stock prices by first making them stationary using differencing. Pipin, et al. (2023) used the Recurrent Neural Network - Long Short -Term Memory (RNN - LSTM) model to predict stock prices because it can identify non-linear data and reduce complex trends in stock prices. Then, Catello et al. (2023) predicted the shares of Apple Inc., IBM Corp., and Dell Inc. from October 13, 2004, to January 21, 2005, using the hidden Markov model (HMM) and gave results that HMM has strong predictive capabilities compared to several models, such as ARIMA.

Currently, pharmaceutical companies are among the industrial companies that are growing quite significantly. Pharmaceutical companies are important contributors to economic growth (Sari & Yousida, 2022). In addition, pharmaceutical companies play an important role in realizing public health as researchers, developers, and distributors of medicines. Therefore, pharmaceutical companies will always be needed in any condition (Majid & Oktavina, 2016).

PT Kimia Farma Tbk is a state-owned enterprise in the pharmaceutical sector. It was founded in 1817 by the Dutch colonial government as Indonesia's first and oldest pharmaceutical company (Pakpahan & Rioni, 2024). Many people are interested in becoming stock investors in PT Kimia Farma Tbk, which can be seen from the number of shareholders, which is 90.025% by the government and 9.975% held by the public (Asnawi et al., 2021). In addition, based on data from finance.yahoo.com, PT Kimia Farma Tbk's daily stock return price index (KAEF) from 2012 to 2024 continues to fluctuate as in Figure 1 below.

**Figure 1. Daily stock returns of PT Kimia Farma Tbk (KAEF)**

Due to certain conditions, stock returns KAEF are constantly changing. Most fluctuate between -1 and 1, and several times experience drastic decreases and increases between -2 and 2. This makes stock price predictions unable to use just one model; in other words, a model is needed that allows changes from one model to another. In addition, stock return conditions that cause positive or negative stock returns cannot be observed directly, so a hidden Markov model (HMM) is expected to be able to predict them.

A Hidden Markov model is a pair of stochastic processes with the underlying stochastic process not directly observable (hidden) but can only be observed through a series of other stochastic processes that produce the observed data set (Rabiner & Juang, 1986). A hidden Markov model consists of two processes, the Markov process and the observation process (Setiawaty, 2002). The Markov process is a stochastic process that has Markov properties. This means that the next conditions only depend on the current conditions and do not depend on the conditions in the past (Hoek & Elliot, 2018).

This HMM is divided into two classes, i.e., discrete HMM for discrete observation sequences and continuous HMM for continuous observation sequences. In this study, this daily stock return will be modeled using discrete HMM first to obtain HMM parameters. Then, to fulfill the continuous nature of daily stock return data, the parameter containing the emission probability or probability mass function is replaced by a parameter containing the probability density function of each distribution per state. So, at the end we will work with a continuous HMM.

2. MATERIALS AND METHODS

**2.1 Markov Chains**

Suppose is a sequence of random events from the sample space . The sequence is a Markov chain if it satisfies the following conditions:

for each and all . This means that the Markov chain is a stochastic process that has a condition where the probability of a future event depends only on the current event (Hoek & Elliot, 2018).

The Markov chains are used to measure or estimate movements that occur at any time. This process involves the use of a Markov transition matrix where each value in the transition matrix is the probability of each movement from one state toanother (Syafruddin et al., 2014).

**2.2 HiddenMarkov** **Models**

A Hidden Markovmodel (HMM) consists of a pair of stochastic processes . The set of events is not observed directly; in other words, it is hidden and represents a collection of events that form a Markov chain. represents the observation process (Rabiner 1989). The following are the parameters of a hidden Markov models with discrete observations:

1. , the number of states of theMarkov chain with state space .
2. , the number of possible observations of for each state*,* with the observation space is .
3. , the transition probability matrix with and .
4. , the emission probability matrix with ,

, and .

1. , initial probability matrix with and .

Let be the representation of the discrete HMM.

There are two assumptions used in HMM. First, the Markov assumption states that the next state is only influenced by the current state. Second, the current observation only depends on the current state and is independent of previous observations (Jurafsky & Martin, 2009). The Illustration of dependency between variables in an HMM can be seen in Figure 2.

Markov Chain

Observation Process

**Figure 2. Illustration of the hidden Markov model**

There are three main problems in using HMM. First, the evaluation problem, namely calculating the probability of a sequence of observations, can be solved using the forward algorithm. Second, the decoding problem aims to find the optimal hidden sequence that maximizes the probability of the observation sequence and the hidden state. Third, the learning problem is finding the best parameters using the Baum-Welch and expectation maximization (EM) algorithms (Young et al., 2002). The following is a description of the solution to these three problems.

**2.2.1 Evaluation Problem**

Given an HMM and a sequence of observations where *T* is a number of observations, we will calculate the joint mass probability of the sequence of observations as follows (Young et. al*.* 2002).

 .

The above formulation has a repetitive nature, so a forward algorithm is needed to facilitate the calculation of the evaluation problem. Define the forward variable:

, for and .

The steps of the forward algorithm are:

* 1. Initialization:

, .

* 1. Induction: for

, .

* 1. Termination

.

**2.2.2 Decoding Problem**

Given the HMM and the observation sequence , we will find the optimal state sequence which maximizes

This problem can be solved using the Viterbi algorithm as follows (Rabiner 1989). For and , define

.

By induction we obtain:

, for .

with the following steps.

1. Initialization

,

, For .

1. Recursion for

, .

, .

1. Termination: best *state* at time

.

1. Backtracking*:* the best *state sequence* on

.

1. Output*:*

*.*

**2.2.3 Learning Problem**

Given HMM and observation sequence , we will obtain the best parameters that satisfies

.

This problem can be solved by the Baum-Welch algorithm and the expectation algorithm maximization(EM). The following are the steps for the Baum-Welch algorithm to search (Rabiner 1989).

1. Forwardand backwardvariables for and are defined as

.

1. For , define

 ,

, for

Note that .

Estimation is done using the Expectation algorithm Maximization(EM) which consists of two stages, namely stages E and M(Harpaz and Haralick 2006).

1. E-Step

For each with is the set of all HMM parameters, calculate

 It can be seen that the log-likelihood function becomes

,

implying

,

 since (Firmansyah et al., 2018).

1. M-Step

Find so that . Therefore,

in other words

(Shamsul et. al*.* 2009).

Solving the M-step will give the estimate parameter We can find the detail in Firmansyah et al. (2018).

**2.3 Mean Absolute Error**

According to Subagyo (1986) the mean absolute error (MAE) is one measure of accuracy in forecasting that shows how big the average prediction error is with the actual data . The smaller the MAE value*,* the more accurate the prediction results obtained. The equation used to obtain the MAE valueat a predicted value () with the actual data ()is:

*,*

where is the amount of data.

3. results and discussion

**3.1 Data Description**

The data used in this study is the daily closing stock return data of PT Kimia Farma Tbk with an observation period of January 2, 2012 - December 31, 2024. The 3208 data were obtained from the finance.yahoo.com site. It has an average of 0.000211347 and a variance of 0.00134296 with a range of values between -0.18 to 0.22. In this modeling, 90% of the return data was selected as training data and the rest as test data.

**3.2 Discrete HMM Modeling**

The return data of PT Kimia Farma Tbk used in this modeling is continuous data. The initial stage in determining the distribution of each *state* or condition in continuous HMM requires discrete HMM modeling first, so it is necessary to carry out discretization with clustering*.* This modeling uses the -means clustering method*,* which is a method of grouping data into data groups based on the center of the group closest to the data, or commonly called the centroid(Ong 2013). In this HMM, the value of *k* used is the number of possible observations (*.* Determination of the best model is done with the help *of* Akaike Information Criterion (AIC) according to the number of states and cluster with the AIC equation as follows.

with being the likelihood valueof the HMM model (Cavanaugh & Neath, 2019). Using the forward algorithm and training data, we calculate the value of the loglikelihood and the AIC values for The Table 1 and Table 2 show the results of calculating the log-likelihood and the AIC value for HMMs with the combination of states and cluster

**Table 1. Log likelihood Values in Discrete HMM**

****

From Table 1, we see that for the HMMs have almost the same loglikelihood value for every . So, we will focus only for HMMs with the value of

**Table 2. AIC Values on Discrete HMM**



For each , the number of is selected as the state that has the smallest AIC value. In Table 2, the smallest AIC values are colored yellow. This minimum AIC value converges to for . The minimum value of AIC for is obtained when . So, the best discrete HMMwas chosen when and *.* Using the Baum-Welch Algorithm,the values of the discrete HMM parameters with the combination of the number of states and are as follows.

 *;*  ; .

The matrix shows the transition probability between hidden states. This matrix indicates that the hidden state 1 tries harder to stay in that state with a probability of 0.848, while hidden states 2 and 3 both may transit to each other with a fairly large probability. The initial probability matrix shows a value of 0.9 in the hidden state 2 which indicates that the model will most likely start with hidden data state 2 if generated based on the model that has been obtained. The matrix shows probability of each hidden state to generate data.

Using the Viterbi algorithm, we can predict the state that generates every data. The hidden state sequence which generates the data can be seen in Figure 3.



**Figure 3. Visualization of hidden sequences discrete HMM state with 3 states and 3 clusters**

In Figure 3, state 1 is visualized in red, state 2 in blue, state 3 in green. State 1 tends to generate very large data changes, state 2 generates the data that fluctuations are not too large, while state 3 generates data that tends to be stationary.

Using the discrete HMM, we generate 50,000 random discrete data sets. Since our model is discrete, we choose the mean cluster to represent the cluster (. The best MAE value obtained from the simulation results on the training data was 0.03188, which means 7.97% of the length of the original data range. Figure 4 is a visualization of the simulation data against the original data on discrete HMM.



**Figure 4. Visualization of discrete HMM training data simulation with MAE 0.022.**

Based on Figure 4, although the discrete HMM provides a small MAE value, it visually provides a less accurate picture. The up and down movements of the simulated data do not look like the original data. This is because the model built does not match the characteristics of the data. Stock return data is continuous, while the projection data is obtained from a discrete model. Therefore, this discrete model needs to be modified according to the characteristics of continuous data so that a model can improve the accuracy of predictions from the data.

**3.3 Modifying Parameters of Discrete HMM**

In this section we will do some modification to the parameter of the discrete HMM. The parameter with

for and will be replaced by a parameter , with is the joint density function of for Using the new parameter , with and are the same parameter obtained in HMM discrete modeling, we have the joint density function of observations as follow

From the Viterbi Algorithm, data generated by each hidden state can be obtained. The parameter for each state can be estimated using the maximum likelihood estimation (MLE) method. In this modeling, two continuous distributions will be compared, i.e. the normal distribution and the logistic distribution.

Next, construct two modified HMMs, firstly, the continuous HMM with each state data normally distributed with

 *;*  ;

and secondly, the continuous HMM with each state data logistic distributed with

 *;*  ;  .

Table 3 shows the summary of distributions and their parameters for each hidden state.

**Table 3. Continuous distribution parameters of each hidden state**

|  |  |  |
| --- | --- | --- |
| Hidden State | Normal Distribution | Logistics Distribution |
|  |  |  |
|  |  |  |
|  |  |  |

The return price data of PT Kimia Farma Tbk which has been modeled with each modified HMM parameter will be simulated by generating random data as in discrete HMM. The decision on which parameter to choose depends on the five smallest MAE values in the simulation of the random data generator *i* of 50,000 data sets. Table 4 shows the five best MAE values from the simulation of 50,000 data sets.

**Table 4. Best MAE values from simulations of 50000 random data sets**

|  |  |
| --- | --- |
| Normal Distribution | Logistics Distribution |
| Random seed | MAE | MAE Percentage  | Random seed | MAE | MAE Percentage  |
| 540 | 0.03237 | 8.09258 | 338 | 0.03016 | 7.54108 |
| 4922 | 0.03239 | 8.09825 | 4616 | 0.03007 | 7.51820 |
| 9558 | 0.03193 | 7.98233 | 8382 | 0.02997 | 7.49123 |
| 11,005 | 0.03178 | 7.94425 | 12267 | 0.02982 | 7.45448 |
| 37071 | 0.03201 | 8.00235 | 30583 | 0.03002 | 7.50545 |

Based on Table 4, HMM modified using logistic distribution tends to have a smaller MAE value compared to HMM modified using normal distribution. From 50000 data sets, the smallest MAE value for HMM was obtained by the 12267th of the random data sets. The magnitude of this MAE value is 0.029818, which means 7.45% of the original data range. Visualization of the simulation data with an MAE value of 0.029818 can be seen in Figure 5.



**Figure 5. Visualization of HMM training data simulation modified with MAE 0.029818.**

Figure 5 shows that the simulation data (yellow) provides a movement that matches the movement of the original data (blue). The MAE value is also approximately 10% of the original data range. This means that the modified HMM modeling using the parameter can model the training data well and accurately.

The data that has been simulated on the training model, that is, data generated using random seed 12267, is then continued with the testing model. The simulation results show an MAE value of 0.0338527. Visualization of the testing data simulation can be seen in Figure 6.



**Figure 6. Visualization of modified HMM test data simulation with MAE 0.0338527.**

Based on Figure 6, the simulation data has a similar movement to the test data. In addition, the MAE value is comparable to 8.46% of the original data range, which means less than 10%. Although there is still an error in Figure 6, it can be concluded that the model can predict well and accurately.

4. Conclusion

Daily return data of PT Kimia Farma Tbk can be modeled using the modified discrete hidden Markov model. The size of the error in the simulation is calculated using the Mean Absolute Error (MAE). The MAE value obtained is <10% of the range of daily stock return data of PT. Kimia Farma Tbk. These results indicate that the modified discrete HMM modeling provides accurate simulation results for training data and test data. In addition, the depiction of the simulation data also shows movements similar to the original data.

Disclaimer (Artificial Intelligence)

Authors hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of this manuscript.

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

References

Asnawi MH, Hasanah SA, Gizela VM, Hasana AE, Toharudin T. 2021. Application of ARCH/GARCH in Stock Price Prediction of PT. Kimia Farma (Persero) Tbk. National Online Statistics Seminar, ISSN ONLINE (SNSO 2021).

Cahyani NNM, Mahyuni ​​LP. 2020. Accuracy of Moving Average in LQ45 Stock Prediction on the Indonesia Stock Exchange. E-Journal of Management. Vol: 9(7) pages 2769-2789

Cavanaugh JE, Neath AA. 2019. The Akaike Information Criterion: Background, Derivation, Properties, Application, Interpretation, and Refinements. Wiley Interdisciplinary Reviews: Computational Statistics. 11(3):1-11. doi:10.1002/wics.1460.

Catello L, Ruggiero L, Schiavone L, Valentino M. 2023. Hidden Markov Models for Stock Market Prediction. arXiv: 2310. 03775v1.

Firmansyah M, Setiawaty B, Purnaba IGP. 2018. Parameter Estimation of Exponential Hidden Markov Model and Convergence of Its Parameter Estimator Sequence. International Journal of Applied Mathematics. 31(1):58-59.

Harpaz R, Haralick R. 2006. The EM Algorithm as a Lower Bound Optimization Technique. The Graduate Centre, New York (US).

Hoek J, Elliot RJ. 2018. Introduction to Hidden Semi-Markov Models. Cambridge University Press, United Kingdom.

Husnan, Suad. 2005. Basics of Portfolio Theory and Securities Analysis, Fourth Edition. Yogyakarta: UPP STIM YKPN.

Jogiyanto. 2009. Portfolio Theory and Investment Analysis. Third Edition, First Printing. Yogyakarta, BPFE.

Jurafsky D, Martin JH. 2009. Speech and Language Processing. Second Edition. New Jersey (US), Prentice-Hall.

Majid MN, Oktavina M. 2016. Analysis of Individual Stock Performance in the Pharmaceutical Industry on the Indonesia Stock Exchange in 2016. Innovatio. Vol: 16(2) pages 109-118.

Ong JO. 2013. Implementation of the K-Means Clustering Algorithm to Determine President University's Marketing Strategy. Scientific Journal of Industrial Engineering. Vol: 12(1) pages 10-20.

Pakpahan AY, Rioni YS. 2024. Financial Report Analysis to Assess Financial Performance at PT. Kimia Farma Tbk. for the period 2017-2022. E: Scientific Journal of Local Political Studies and Development. Vol: 10(3) page 146-156.

Pipin SJ, Purba R, Kurniawan H. 2023. Stock Prediction Using Recurrent Neural Network (RNN-LSTM) with Adaptive Moment Estimation Optimization. Journal of Computer System and Informatics (JoSYC). Vol: 4(4) page 806-815.

Rabiner LR, Juang B. 1986. An Introduction to Hidden Markov Models. IEEE ASSP Magazine. Vol: 3(1) page 4-16.

Rabiner LR. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE. Vol: 77(2) page 257‐286.

Rusyida WY, Pratama VY. 2020. Garuda Indonesia Stock Price Prediction in the Midst of the Covid-19 Pandemic Using the ARIMA Method. SQUARE: Journal of Mathematics and Mathematics Education. Vol: 2(1) page 73-81.

Sari NNP, Yousida I. 2022. The Effect of Liquidity and Profitability and Company Size on Stock Prices in Pharmaceutical Companies Listed on the Indonesia Stock Exchange. Scientific Journal of Economics and Business. Vol: 8(2) page 209-220.

Sawidji S. 2012. How to Quickly Start Stock Investment for Beginners Revised Edition. Jakarta, Gramedia.

Setiawaty B. 2002. Equivalent Representations of Hidden Markov Models. Journal of Mathematics and Its Applications. Vol: 2(1) page 1-12.

Setiawaty B., Lestari A. Y., Lesmana D. C., Ardana N. K. K. 2024. Predicting the Conditions of Stock Market Using Hidden Markov Models. AIP Conference Proceedings. Vol: 3201:020013. doi.org/10.1063/5.0230595

Shamsul H, John Y, Roberto T. 2009. Pattern recognition letters. Elsevier Journal. Vol: 30(1) page 1301-1309.

Subagyo, Pangestu. 1986. Forecasting Concept and Application. Yogyakarta, BPPE UGM.

Syafruddin S, Irma S, Sukarna. 2014. Application of Markov Chain Analysis to Predict Patient Status at Barru District Hospital. Online Journal of Natural Science. Vol: 3(3) page 313-321.

Troiano L, Villa EM, Loia V. 2018. Replicating a Trading Strategy by Means of LSTM for Financial Industry Applications. IEEE Trans Industr Inform. Vol: 14(7) pages 3226-3234.

Young S. Evermann G, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P. 2002. HMM Toolkit (HTK) Book (for version 3.2.1). University Engineering Department, Cambridge