

## Prediksi Return Saham Berbasis *Clustering* Menggunakan *K-Means Clustering* dan *Convolutional Neural Network*

Rizaldi Maulid Fathurachman<sup>1</sup>, Deni Saepudin<sup>2</sup>

<sup>1,2</sup>Fakultas Informatika, Universitas Telkom, Bandung

<sup>1</sup>rizaldimaulidf@students.telkomuniversity.ac.id, <sup>2</sup>denisaepudin@telkomuniversity.ac.id

### Abstrak

Prediksi harga saham di masa depan merupakan sebuah masalah yang menantang. Dalam penelitian ini telah diterapkan pendekatan *clustering* untuk mengelompokkan saham-saham yang menunjukkan kesamaan dalam prediksi harga. Proses ini membantu dalam mempermudah prediksi harga, karena saham yang ada dalam 1 cluster yang sama diharapkan mempunyai pola yang lebih mirip. Melalui *clustering*, saham-saham yang memiliki prediksi harga yang sama dikelompokkan berdasarkan cluster yang telah terbentuk. Kemudian, dilakukan training model menggunakan data rata-rata dari harga saham dalam setiap cluster untuk menghasilkan prediksi harga yang lebih tepat. Dalam penelitian ini metode *K-Means* digunakan untuk mengelompokkan saham berdasarkan rata-rata return dan standar deviasi, sedangkan *CNN* digunakan untuk memprediksi harga saham harian. Selanjutnya prediksi return dapat dihitung berdasarkan prediksi harga yang telah diperoleh. Dataset harga saham LQ45 dibagi menjadi dua bagian, yaitu training data dan test data. Data training digunakan untuk melatih model *CNN* untuk menganalisis data harga saham historis yang dapat membantu memprediksi return saham di masa depan, sedangkan data test digunakan untuk mengevaluasi kinerja model yang telah dilatih. Hasil penelitian ini menunjukkan bahwa, dengan melihat kriteria elbow maka *clustering* dengan metode *K-Means* akan diambil dengan jumlah clusternya adalah sebanyak 4 cluster. Dan setelah itu di lakukan prediksi melalui *CNN* dan hasilnya mampu menghasilkan prediksi harga saham harian dengan baik, dibuktikan dengan nilai *RMSE* prediksi untuk 1 hari ke depan pada *cluster* 1 adalah 0,09, Untuk *cluster* 2, *RMSE* yang diperoleh adalah 0,05, Sementara itu, *cluster* 3 memiliki *RMSE* sebesar 0,07, dan untuk *cluster* 4 memiliki *RMSE* yang diperoleh yaitu 0,57. Setelah itu, dihasilkan dari prediksi saham-saham di setiap *cluster* yang menghasilkan prediksi *return* tertinggi di setiap harinya. Kemudian dibuat prediksi kinerja portofolio menggunakan *equal weight* dengan hasil rata-rata *return* sebesar 0,0009 dan standar deviasi sebesar 0,0209 yang lebih baik dibandingkan dengan indeks LQ45.

**Kata kunci:** Prediksi *return* saham, *clustering*, *K-means*, *Convolutional Neural Network (CNN)*, portofolio investasi.

### Abstract

*Predicting future stock prices is a challenging problem. In this research, a clustering approach will be applied to group stocks that show similarities in price movements. This process helps in making price prediction easier, as stocks in the same cluster are expected to have more similar patterns. Through clustering, stocks that have similar price movements are grouped based on the clusters that have been formed. Then, a training model is performed using average data from stock prices in each cluster to produce more precise price predictions. In this research, the K-Means method is used to group stocks based on average return and standard deviation, while CNN is used to predict daily stock prices. Furthermore, return predictions can be calculated based on the price predictions that have been obtained. The LQ45 stock price dataset is divided into two parts, namely training data and test data. Training data is used to train CNN models to analyze historical stock price data that can help predict future stock returns, while test data is used to evaluate the performance of the model that has been trained. The results of this study indicate that, by looking at the elbow criteria, clustering with the K-Means method will be taken with the number of clusters is 4 clusters. And after that, predictions are made through CNN and the results are able to produce daily stock price predictions well, as evidenced by the RMSE value of predictions for the next 1 day in cluster 1 is 0.09, For cluster 2, the RMSE obtained is 0.05, Meanwhile, cluster 3 has an RMSE of 0.07, and for cluster 4 has an RMSE obtained which is 0.57. After that, the predictions of stocks in each cluster are generated which produce the highest return predictions on each day. Then a portfolio performance prediction is made using equal weight with an average return of 0.0009 and a standard deviation of 0.0209 which is better than the LQ45 index.*

**Keywords:** Stock return prediction, clustering, K-means, Convolutional Neural Network, investment portfolio.

## 1. Pendahuluan

### 1.1 Latar Belakang

Dalam penelitian ini dibahas mengenai tantangan dalam memprediksi harga saham di masa depan sebagai masalah utama dalam prediksi return saham. Masalah ini menarik karena dapat membantu investor

memahami karakteristik saham tertentu untuk keputusan investasi yang lebih baik. Untuk mengatasi tantangan tersebut, dalam penelitian ini diterapkan pendekatan *clustering* untuk mengelompokkan saham-saham yang menunjukkan kesamaan dalam prediksi harga. Proses ini membantu dalam mempermudah prediksi harga, karena saham yang ada dalam 1 cluster yang sama diharapkan mempunyai pola yang lebih mirip. Melalui *clustering*, saham-saham yang memiliki prediksi harga yang sama dikelompokkan berdasarkan *cluster* yang telah terbentuk. Kemudian, dilakukan *training model* menggunakan data rata-rata dari harga saham dalam setiap *cluster* untuk menghasilkan prediksi *return* yang lebih tepat.

Banyak penelitian telah dilakukan untuk mengatasi masalah ini. Penelitian Hongxiang He dan tim pada tahun 2018, Dimana dalam penelitian ini menemukan bahwa CNN mampu menghasilkan prediksi yang lebih akurat dibandingkan metode lain seperti regresi linier dan model ARIMA. Dalam eksperimen, CNN menunjukkan akurasi prediksi yang signifikan dengan akurasi 73,2% hingga 75,2% setelah 500.000 iterasi pada berbagai set data. Selain itu, F1-score tertinggi mencapai 73,0%, yang menegaskan kemampuan CNN untuk menangkap pola kompleks dalam data harga saham, memberikan keunggulan dalam memproyeksikan prediksi harga saham dibandingkan dengan model tradisional lainnya[4].

Penelitian yang dilakukan oleh Zheng Fang dkk. 2021, menggunakan algoritma K-means yang dioptimalkan dengan algoritma Artificial Fish Swarm (AFSA) untuk mengelompokkan saham berdasarkan karakteristik tertentu. Hasilnya menunjukkan bahwa metode KAFSA efektif dalam mengelompokkan saham menjadi dua kategori: saham berkinerja tinggi dan rendah, dengan koefisien siluet sebesar 0,99. Saham dalam kelompok B (berkinerja tinggi) memiliki harga penutupan, rasio harga-pendapatan, laba per saham, dan pengembalian aset bersih yang lebih tinggi dibandingkan dengan saham dalam kelompok A (berkinerja rendah), menawarkan panduan yang baik bagi investor dalam membuat keputusan investasi[2].

Adapun paper yang menjadi referensi acuan adalah Sarfaraz Hashemkhani Zolfani dkk pada tahun 2023, menggunakan metode PSO-CNN+MVF dan *K-means clustering* untuk mengelompokkan saham-saham berdasarkan mean dan variance dari fitur input dan return aset selama periode pelatihan. Hasil menunjukkan bahwa model ini unggul dalam return dan pengelolaan risiko. Model LSTM memberikan return harian 0,0071, sementara CNN+MVF mencatat 0,0046 setelah biaya transaksi. CNN+1/N mencatat risiko terendah, dan secara tahunan, metode yang diusulkan mencapai pengembalian tertinggi 4,9007 dengan rasio Sortino 9,0543, mengungguli model lain dalam efisiensi dan risiko yang lebih rendah, menunjukkan potensi besar dalam strategi investasi berbasis data[1].

Dalam penelitian ini, dilakukan *clustering* saham dengan menggunakan metode *k-means*. Fitur-fitur yang digunakan dalam clustering ini adalah rata-rata *return* dan standar deviasi. Setelah *return* saham terbentuk, semua saham telah diclusterkan berdasarkan karakteristik tersebut. Selanjutnya, untuk memprediksi tren pasar, digunakan algoritma *Convolutional Neural Network* (CNN) untuk memprediksi *return* saham harian. Hasil dari prediksi ini kemudian digunakan sebagai dasar dalam memilih saham untuk portofolio investasi. Saham-saham dengan *return* tertinggi kemudian dimasukkan dalam portofolio dan dengan menggunakan metode *equal weight*, kemudian portofolionya diuji dibandingkan dengan kinerja dari indeks LQ45.

Dengan demikian, penelitian ini bertujuan untuk menghadirkan solusi baru dalam analisis pasar saham, membantu para investor membuat keputusan investasi yang lebih informasional dan terinformasi.

## 1.2 Topik dan Batasannya

Topik yang dibahas dalam penelitian ini adalah bagaimana penerapan clustering dan penggunaan CNN dapat mempengaruhi prediksi return saham untuk seleksi awal dalam optimasi portofolio.

Adapun batasan dari penelitian ini adalah dataset indeks LQ45 yang diperoleh dari situs yahoo finance. Dataset yang digunakan adalah data harga harian selama 10 tahun mulai dari 1 Januari 2013 sampai 1 Januari 2023. Dataset ini terdiri dari 39 saham yang tergantung dalam indeks LQ45.

## 1.3 Tujuan

Tujuan dari penelitian ini adalah menerapkan metode *K-means clustering* untuk mengelompokkan saham berdasarkan rata-rata *return* dan standar deviasi, membangun model prediksi menggunakan *Convolutional Neural Network* (CNN), serta mengembangkan metode pengukuran kinerja portofolio yang terbentuk dari saham-saham yang dipilih berdasarkan prediksi return. Setelah itu menerapkan metode prediksi untuk menyeleksi saham dalam portofolio, Tujuannya adalah untuk membantu investor membuat keputusan investasi yang lebih baik dengan menemukan saham dengan potensi pertumbuhan terbaik.