Data Mining Approaches to Analyze Hang Seng Index and Hong Kong Stock Market

By Group 25

HE Hongliang 58631330

WU Shipeng 59047700

ZHOU Haitao 58324678

22 April 2025

Contents

Abstract

The Hong Kong Stock Market, particularly the Hang Seng Index (HSI), presents significant challenges for predicting movements and identifying behavioral patterns due to its volatility. Existing approaches typically rely on statistical models, which struggle with non-linear dynamics, or complex machine learning methods, which often lack interpretability and risk overfitting. To address these issues with simpler, interpretable models, this study proposes an integrated data mining framework using decision trees for HSI movement prediction, K-means clustering for seasonal trend analysis, hierarchical clustering for stock segmentation, and association rule mining for feature relationships. Leveraging Tushare API data, our approach achieves a mean prediction accuracy of 0.6649, uncovers seasonal trends linked to market instability (2020–2022), and identifies stocks with favorable risk-return profiles. Validated through cross-validation and real-world correlations, this framework demonstrates that simpler models can provide actionable, interpretable insights for investment decision-making.

1 Introduction

1.1 Motivation

The stock market plays a crucial role in global economies, and understanding its dynamics is essential for investors and policymakers. Predicting stock price movements is a complex challenge that requires careful analysis of historical data and the identification of underlying patterns. This project focuses on the Hong Kong Stock Market, specifically the Hang Seng Index (HSI), which serves as a barometer for the performance of the Hong Kong stock market. Given the volatility and unpredictability of stock prices, it is vital to develop methods that can assist in forecasting future trends and help investors make informed decisions.

1.2 Problem Definition

The main problem addressed in this project is predicting the movements of the HSI and gaining deeper insights into the patterns of stock market behavior. The project aims to tackle several specific challenges:

a) Predicting Index Movements: The objective is to predict whether the HSI will rise or fall based on historical data and market features.
b) Segmenting Stocks: Stocks within the HSI are not homogeneous, and their behavior can vary significantly. Clustering stocks into groups of similar behavior can help in identifying patterns and improving stock selection.
c) Uncovering Association Rules: Identifying dependencies among different stock features, such as price changes, volatility, and trading volume, can reveal hidden relationships and provide insights into market behavior.

1.3 Objectives

The key objectives of this project are as follows:

a) Predict the Movement of the Hang Seng Index: Using decision trees, we aim to predict the rise and fall of the HSI based on past market data, including features like opening prices and daily changes.
b) Analyze Seasonal Growth Trends: By applying K-means clustering, the project aims to identify seasonal trends in the HSI, helping to forecast patterns and understand the market's cyclical behavior.
c) Segment Stocks into Behaviorally Similar Groups: Using hierarchical clustering, stocks are grouped into clusters based on their behavioral patterns, providing a structured way of selecting stocks for investment.
d) Discover Hidden Relationships Using Association Rules: The project uses the Apriori algorithm to uncover dependencies among stock features, offering a data-driven approach to understanding the connections between different aspects of the stock market.

This comprehensive approach integrates different data mining techniques to create a holistic analysis of stock market data, which can ultimately support investment decision-making by revealing trends, correlations, and potential opportunities.

2 Literature Review

The application of data mining techniques to stock market prediction and analysis has been a focal point of research for many years. Several methodologies, including machine learning, clustering, and association rule mining, have been utilized to predict stock market trends, segment stocks, and discover patterns in stock data. This literature review will explore the key studies and approaches that have influenced the methods used in this project.

2.1 Stock Market Prediction Using Machine Learning

Stock market prediction remains one of the most challenging tasks in data science due to the inherently volatile and non-linear nature of financial data. Several machine learning models have been applied to predict stock prices and index movements. Decision trees, a popular machine learning algorithm, have been widely used in predicting stock prices due to their interpretability and ability to handle both numerical and categorical data. The work of Tsung-Sheng Chang[1] explored the use of decision trees to predict stock price movements, finding that decision trees, when combined with historical price data, could offer reasonable predictions for stock price direction.

2.2 Clustering Techniques in Stock Market Analysis

Clustering is another widely used technique for analyzing stock market data. K-means clustering and hierarchical clustering are often used to group stocks based on similarities in their price movements, trading volumes, or other financial indicators. Clustering can be valuable in identifying stocks with similar characteristics, enabling investors to segment the market and make more informed investment decisions.

K-means clustering, for example, is effective in segmenting stocks into groups based on their performance or volatility patterns, helping identify periods of stability or growth. Kuo-Ping Wu et al.[2] showed that K-means clustering could uncover seasonal patterns in stock data by grouping stocks that exhibit similar trends over time. This is consistent with the approach taken in this project, where K-means clustering is applied to identify seasonal growth trends in the HSI.

Hierarchical clustering is another popular method for stock analysis. Doherty et al.[3] utilized hierarchical clustering to segment stocks, helping investors identify opportunities based on the behavioral similarities between stocks. The hierarchical approach is particularly valuable in this project for selecting stocks by using dendrograms, which visualize the hierarchical relationships between stocks, providing a clear picture of their similarity.

2.3 Association Rule Mining for Stock Market Insights

Association rule mining, particularly using the Apriori algorithm, has been employed to uncover relationships between different features of stocks. This technique identifies frequent patterns and dependencies within the stock market, such as the correlation between stock prices and other indicators like volume, price change, or market sentiment.

Srisawat[4] applied association rule mining to discover patterns in the relationships between stock movements and external factors such as interest rates and market sentiment. His work showed that association rules could uncover hidden dependencies, leading to better predictions of stock price movements. The use of the Apriori algorithm in this project aligns with these findings, as it helps discover meaningful patterns in the stock data that may not be immediately obvious through traditional analysis methods.

2.4 Challenges in Stock Market Prediction

Despite the success of various data mining techniques, stock market prediction remains a challenging task due to the noisy and non-stationary nature of financial data. The problem of overfitting is prevalent when using complex models, which can lead to high accuracy on training data but poor generalization to unseen data. Regularization techniques and cross-validation are commonly used to mitigate this issue. In this project, cross-validation was applied to the decision tree model to ensure that it remained reliable across different data subsets.

Additionally, feature selection remains a critical challenge in stock market prediction. Identifying the right set of features that capture the essential characteristics of the market is key to building an effective model. In this project, feature selection was carefully done for both the decision tree and clustering models to ensure that the most informative features were used for prediction and clustering.

3 Methodology

This project employs a range of data mining techniques to analyze the Hang Seng Index (HSI) and stock market data, with the objective of predicting stock index movements, identifying seasonal growth patterns, clustering stocks based on behavioral similarities, and uncovering hidden relationships between stock features. The methodology is divided into data collection and preprocessing, followed by the application of machine learning models and clustering techniques.

3.1 Data Collection and Preprocessing

Data for this analysis was obtained via the Tushare API[5], which provides pre-cleaned stock market data, including daily stock prices, trading volumes, and key market indicators such as the opening and closing prices, percentage changes (pct_chg), and others. Preprocessing steps were carried out to ensure that the data was structured properly and free from inconsistencies or missing values.

a) Data Acquisition Strategy

Data was retrieved iteratively to avoid API rate limits and organized by stock codes to maintain data consistency. The dataset was aggregated according to stock codes, and cleaning operations were performed to ensure that only valid data points were included in the analysis.

b) Feature Selection

Key features, such as percentage change (pct_chg) and volatility, were selected for further analysis. All the features are shown in Table 3.1 and 3.2. These features were used for predicting stock price movements and for segmenting stocks based on their behavioral patterns.

Table 3.1 Features of Hang Seng Index Data

Feature	Explanation
ts_code	Index code
trade_date	Trading date
open / close	Opening / closing points
high / low	Highest / lowest points
pre_close	Previous closing point
change / pct_chg	Point change / percentage change
swing	Amplitude
vol / amount	Trading volume / trading amount

Table 3.2 Features of HKEX Stock Data

Feature	Explanation
ts_code	Index code
trade_date	Trading date
open / close	Opening / closing points
high / low	Highest / lowest points
pre_close	Previous closing point
change / pct_chg	Point change / percentage change
vol / amount	Trading volume / trading amount

3.2 Prediction of the Hang Seng Index Movement Using Decision Trees

Decision trees are a widely used model in machine learning for classification tasks. A decision tree model was employed to predict the movement of the Hang Seng Index (HSI), aiming to classify whether the HSI would rise or fall based on historical data.

3.2.1 Principle of Decision Trees

A decision tree is a hierarchical structure consisting of nodes and branches. Each internal node represents a decision based on a feature, and each leaf node represents a

classification (in this case, rise or fall of HSI). The tree splits the data into subsets based on feature values that lead to the most homogeneous groups. This process is repeated recursively until stopping conditions are met, such as a maximum tree depth or a minimum number of data points in each node.

The primary objective of the decision tree algorithm is to find the optimal splits that minimize impurity, typically using criteria such as Gini impurity or Information Gain. In this project, Gini impurity was used to evaluate the quality of each split, with the goal of creating the most homogeneous subsets possible.

3.2.2 Feature Selection for Decision Tree

Feature selection for the decision tree was performed by evaluating the importance of each feature in predicting the HSI movement. Chosen key features are shown in Table 3.3.

Table 3.3 Features selected for Decision Tree

Feature	Explanation
next_open	Opening point for the next day
close	Closing point from the previous day
open_EMA_5	EMA of the previous 5 days' opening points
close_EMA_5	EMA of the previous 5 days' closing points

Features were selected based on their ability to increase model accuracy, and cross-validation was applied to ensure that the chosen features provided consistent results across different subsets of the data. The model expects to use features from earlier time points to predict the price change of the day, so semantically, the statistical time of the feature will not be earlier than the day's opening.

3.2.3 Model Training and Validation

Cross-validation was employed to train and validate the decision tree model. This technique splits the data into multiple folds, where each fold is used as a testing set while the others are used for training. This ensures that the model is evaluated on all data points, providing a robust estimate of its performance.

3.3 Analyze the seasonal growth trend of HSI using K-Means Clustering

K-Means clustering is an unsupervised learning algorithm used to group data points into K clusters based on similarity. It aims to minimize the within-cluster variance, making each cluster as internally homogeneous as possible.

3.3.1 K-Means Clustering

The K-Means algorithm works by:

a) Initializing K cluster centroids randomly.
b) Assigning each data point to the nearest centroid.
c) Updating the centroids based on the mean of the points assigned to each centroid.
d) Repeating the process until the centroids no longer change significantly.

The number of clusters (K) needs to be specified before running the algorithm. One popular method to determine K is the Elbow Method, which is used to identify the point where adding more clusters results in diminishing returns in variance reduction.

3.3.2 Features Selection for K-Means Clustering

The dataset was split into amounts of 30-days chunks to be clusters. Our target to select features that may benefit to the chunk clustering. The chosen features are shown in Table 3.4.

A combination with 1023 lists from the ten features can be generated. But we should realize that when the feature list gets longer, the cluster score may probabalely decrease. We expected to reach a balance between silhouette score and feature length. So we manually choose $n = 4$.

Table 3.4 Features selected for K-Means Clustering

Feature	Explanation
skewness	Measure of the asymmetry of the probability distribution
kurtosis 错误!未找到引用源。	Measure of the “tailedness” of the probability distribution
ADX	Average Directional Index
ATR	Average True Range

3.3.3 Application in Stock Market Analysis

In this project, K-Means clustering is used to identify seasonal growth trends in the HSI by grouping stocks based on their price movements. This helps to understand how certain groups of stocks perform during specific periods and to identify similar patterns across stocks. The feature selected for clustering primarily includes percentage change (pct_chg), representing the daily stock movement.

3.3.4 Elbow Method for Determining Optimal K

To determine the optimal number of clusters, the Elbow Method was employed. The method involves plotting the sum of squared distances from each point to its assigned cluster centroid (within-cluster sum of squares). The elbow point is identified by locating the K value where the rate of decrease in this sum starts to level off, indicating the ideal number of clusters. As shown in Figure 3.1, the optimal k seems to be $k = 3$.

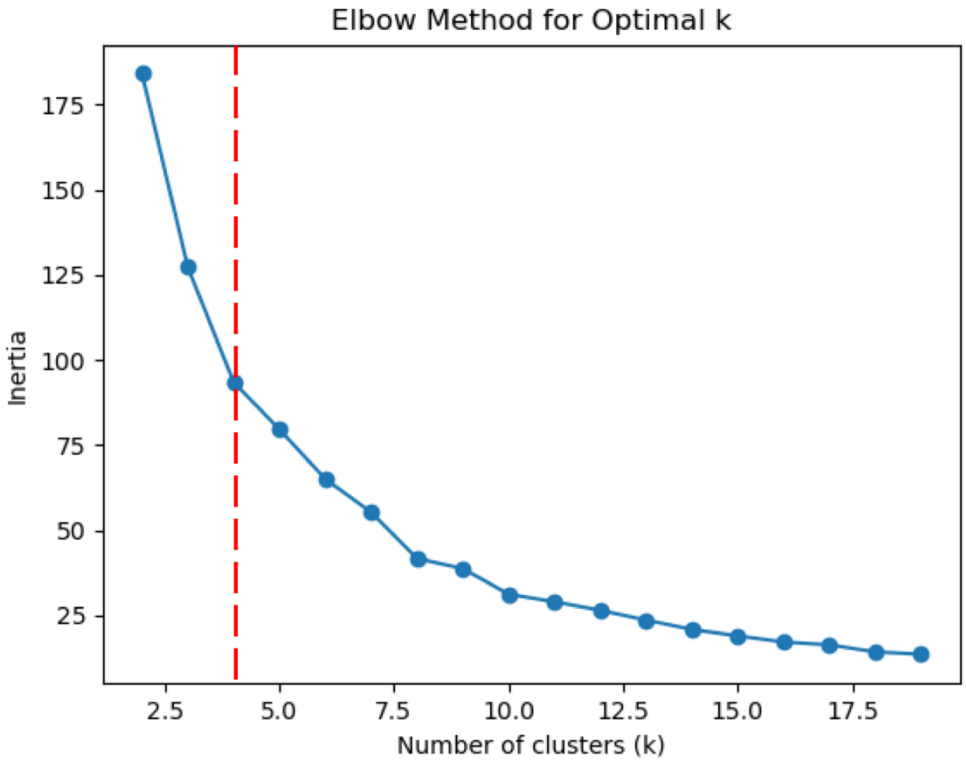

Fig 3.1 Elbow Method for Optimal k

3.3.5 Evaluation of Clustering Quality

The Silhouette Score was used to evaluate the quality of the clusters. The silhouette score quantifies how similar an object is to its own cluster compared to other clusters, with a score close to +1 indicating well-separated and well-defined clusters.

3.4 Implement stock selection Using Hierarchical Clustering

K-Means clustering is an unsupervised learning algorithm used to group data points into K clusters based on similarity. It aims to minimize the within-cluster variance, making each cluster as internally homogeneous as possible.

3.4.1 Hierarchical Clustering Algorithm

Hierarchical clustering proceeds in two primary steps:

a) Agglomerative Clustering: Starts with each data point as its own cluster and iteratively merges the closest clusters until all points belong to a single cluster.
b) Divisive Clustering: Begins with all data points in one cluster and iteratively splits them into smaller clusters.

In this project, the Agglomerative Clustering method was used to group stocks into behaviorally similar clusters based on their historical price movements.

3.4.2 Features Selection for Hierarchical Clustering

Considering that we want to select similar stocks, in this project, we choose the rate of increase or decrease (pct_chg) and volatility, and use the risk and return performance of stocks as clustering indicators. The indicators and their meanings are shown in Table 3.5.

Table 3.5 Features selected for Hierarchical Clustering

Feature	Explanation
avg_pct_chg	Average percentage change of stocks
volatility	Standard deviation of the percentage change of stocks

3.4.3 Dendrogram Construction and Cluster Selection

A dendrogram was constructed to visualize the hierarchical clustering process. By cutting the dendrogram at a certain level, the number of clusters can be determined. This visual approach provides a clear insight into how stocks are related based on their behavior.

The distance threshold for cutting the dendrogram was chosen based on a balance between the number of clusters and the within-cluster similarity, ensuring meaningful groupings. As shown in Figure 3.2, we'd choose optimal cluster $num = 153$.

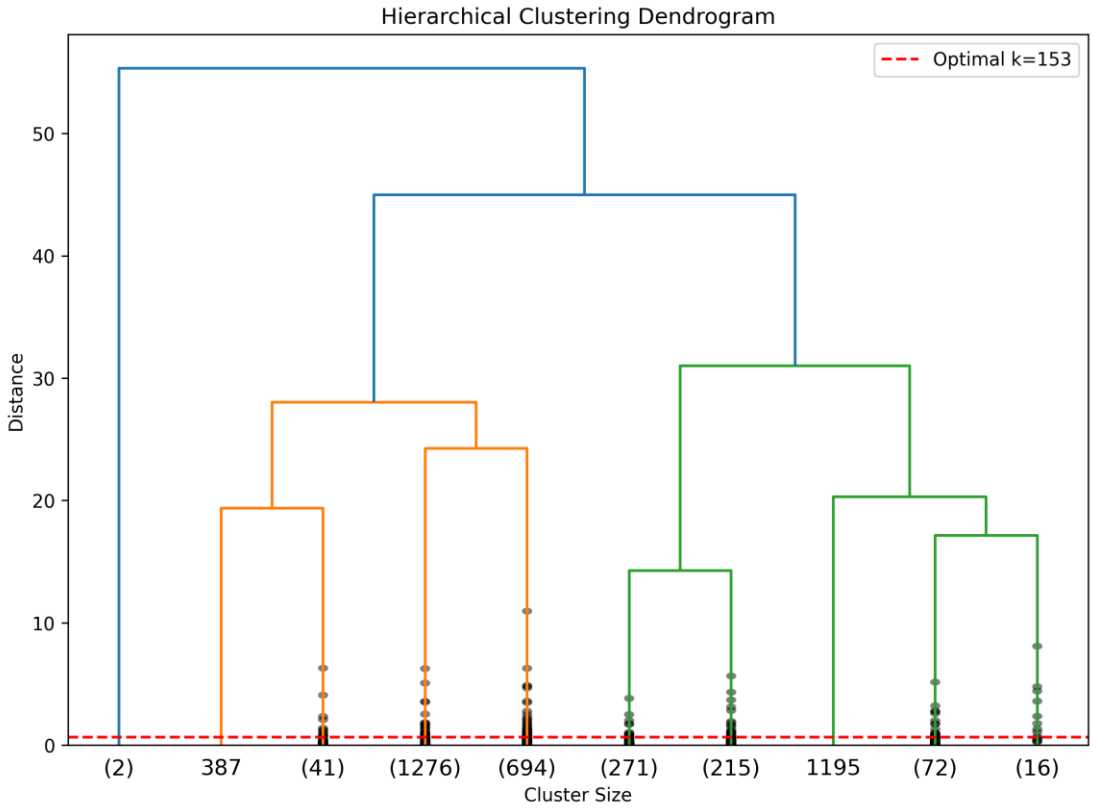

Fig 3.2 Hierarchical Clustering Dendrogram

3.4.4 Evaluation of Hierarchical Clustering

Hierarchical clustering does not rely on a specific metric like K-Means (which uses silhouette scores for validation). Instead, the quality of the resulting clusters is evaluated through the dendrogram, ensuring that the identified clusters make sense in terms of stock behavior patterns.

3.5 Evaluation for Frequent Pattern

The Apriori algorithm was used for discovering frequent itemsets and association rules within the stock market dataset. This technique helps identify relationships between various stock features, such as stock price changes, trading volume, and volatility.

3.5.1 Principle of the Apriori Algorithm

The Apriori algorithm works by identifying frequent itemsets in the dataset. An itemset is considered frequent if it occurs in a sufficiently high number of transactions (or data points). The algorithm then generates association rules, which describe how the presence of one feature (e.g., an increase in stock price) implies the presence of another feature (e.g., an increase in trading volume).

3.5.2 Features Selection for Frequent Pattern

Table 3.6 Features selected for Frequent Pattern

Feature	Explanation
Quantile Binning	Bin the continuous numerical features (pct_chg, swing, and vol) into categorical variables using equal-frequency binning (quantiles).
Bins	The pct_chg (percentage change) divided into Low, Medium, and High categories.

3.5.3 Rule Evaluation Metric

The association rules were evaluated using the following metrics:

a) Support: The proportion of data points that contain both items in the rule.
b) Confidence: The probability that the second item in the rule occurs given the first item.
c) Lift: The ratio of the observed support to the expected support if the items were independent.

These metrics help evaluate the strength and usefulness of the discovered rules in making predictions about stock market movements.

3.6 Integration of Models

One of the strengths of this project is the integration of various data mining techniques, which allows for a comprehensive analysis of stock market data. By combining the results of decision trees, K-Means clustering, hierarchical clustering, and association rule mining, this approach ensures that insights from multiple perspectives are synthesized to form a more accurate and holistic view of the market.

4 Result

In this section, we will evaluate the data mining models and represent the result.

4.1 Decision Tree Results

Table 4.1 shows the evaluation result of our decision tree model with 10-fold cross validation. Although stock markets are a process filled with noise, we have minimized

the influence of noise on our decision tree by using time period statistics, reaching an acceptable accuracy outcome. Figure 4.1 shows how our decision tree model looks like.

Table 4.1 Cross validation result of Decision Tree model

Evaluation Metric	Value
Mean accuracy	0.6649
Standard deviation of accuracy	0.0241

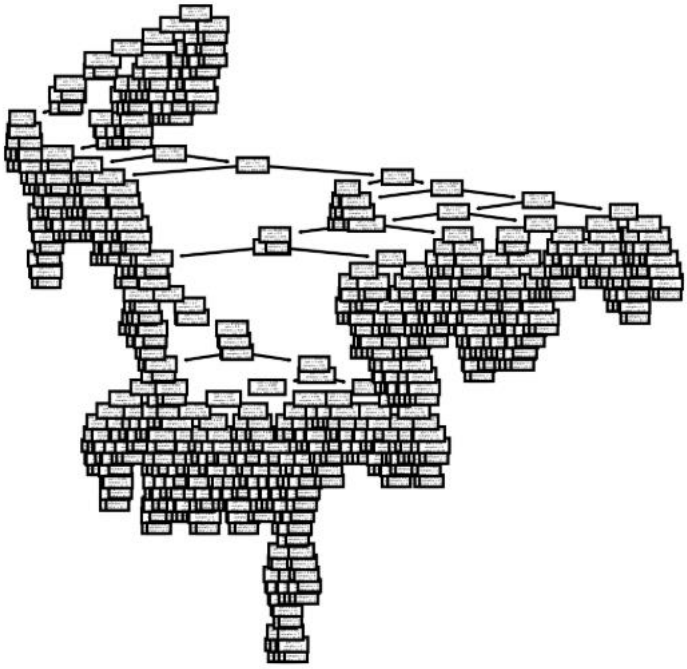

Fig 4.1 Visualized Decision Tree

4.2 K-Means Clustering Results

As is shown in Figure 4.2 , we calculated and showed Principal Component Analysis aka PCA[6] of the features we chose. This diagram can partially reflect the correctness of the algorithm. The PCA points are clustered together. We evaluate the significance of K-Means clustering from another perspective. Since our clustering model utilizes time chunks as instances, we could plot using time along with the mean and standard deviation of features that were not selected. Figure 4.3 shows that between 2020 and 2022, the instability of the market significantly increased. This can be correlated with real-world events, and to some extent, it validates the correctness of the clustering.

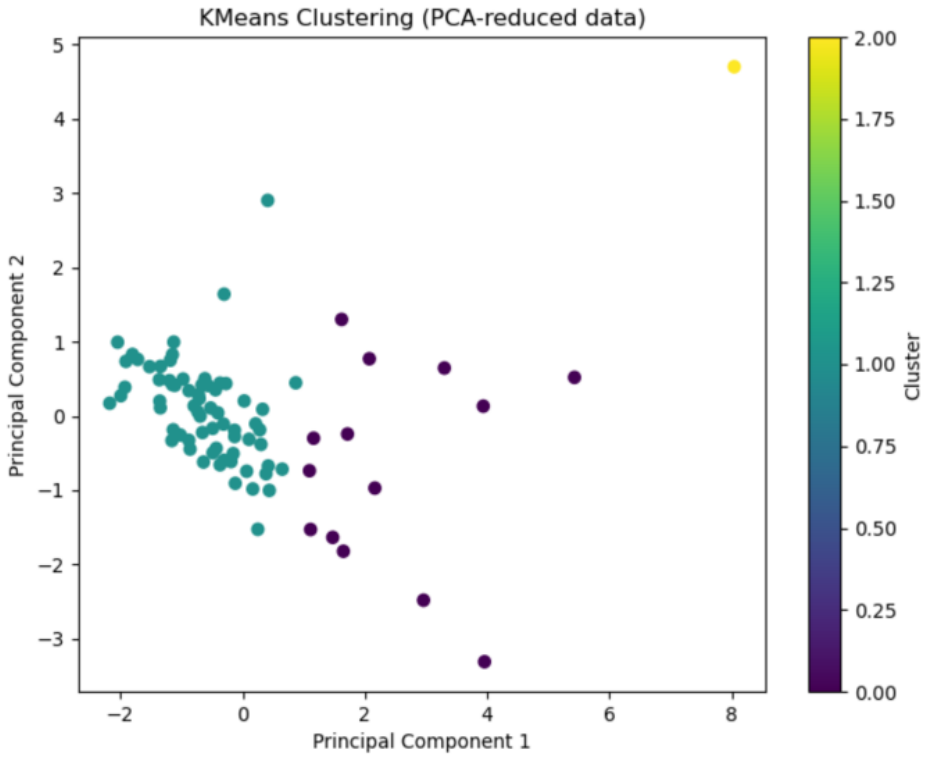

Fig 4.2 Feature distributions of clustered chunks

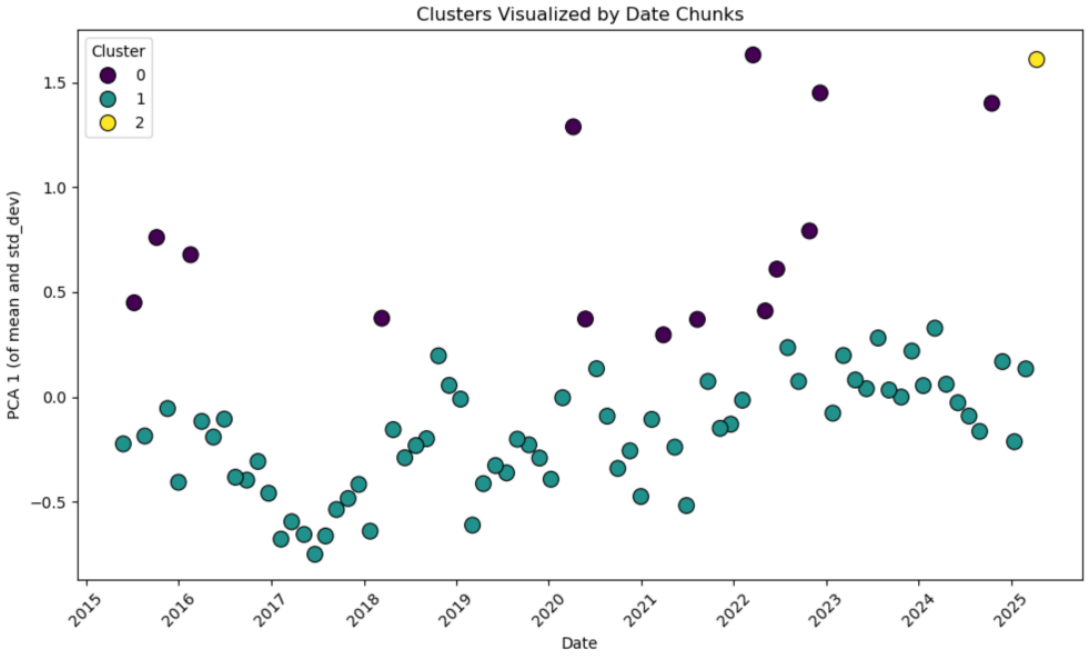

Fig 4.3 Temporal distributions of clustered chunks

4.3 Hierarchical Clustering Results

In stock data hierarchical clustering, we aim to identify stocks with high returns (higher average pct_chg) and low volatility (lower standard deviation of pct_chg). The ideal scenario would be to achieve high returns with low risk, but clearly, such a result is unrealistic because high returns are typically associated with higher risks. Therefore, stocks with low risk can only generate low returns.

From this, we draw two conclusions: First, we can never achieve a result that exactly matches our expectations. Instead, we can only identify data that tends towards this ideal, such as medium-risk stocks that trend towards high returns while also exhibiting lower risk. Hence, we use the normalized difference between returns and volatility as a score to evaluate all stocks. The higher the score, the closer the stock is to our ideal.

Second, we need to design some thresholds to select more stocks that meet the criteria. For example, we can define stocks with returns greater than the overall return of the top 60% and volatility lower than the overall volatility of the bottom 60% of stocks.

The selected stocks are shown in Table 4.4. From the clustering results shown in Figure 4.5, it is evident that the selected stocks not only ensure a return greater than zero, which is higher than most stocks, but also exhibit lower volatility. This confirms that we have successfully selected stocks that better meet the desired conditions.

Table 4.2 Chosen Stocks

ts_code	name
01759.HK	Sino Gas Holdings Group Limited
02097.HK	MIXUE Group
02295.HK	Maxicity Holdings Limited
06693.HK	Chifeng Jilong Gold Mining Co., Ltd.
08030.HK	Fengyinhe Holdings Limited
08112.HK	Cornerstone Financial Holdings Limited
08291.HK	Hong Kong Entertainment International Holdings Limited
08370.HK	Zhi Sheng Group Holdings Limited

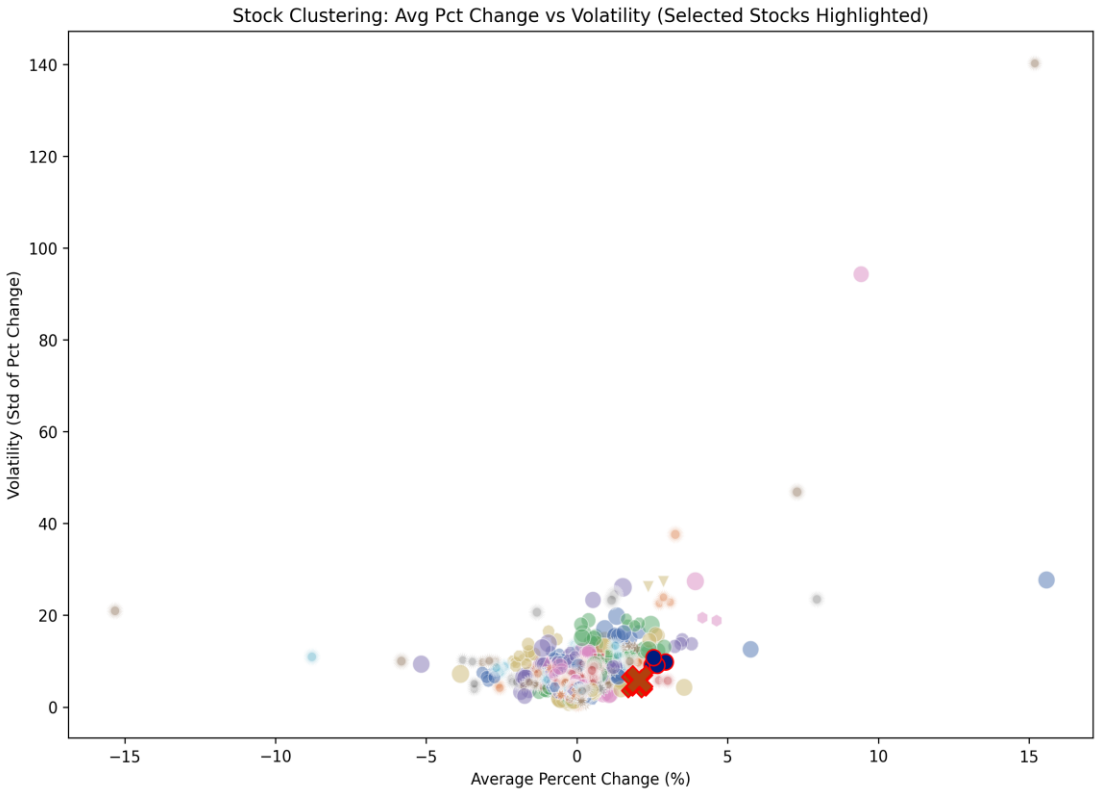

Fig 4.4 Stock Clustering Results with Selected Stocks Highlighted

4.4 Frequent Pattern

The association rules highlight critical interactions between price change, swing, and volume. For instance, high market volatility (swing) is often associated with high trading volumes, especially when the percentage change is significant. Conversely, low market volatility occurs during moderate price changes and low trading volumes.

The test results are shown in Table 4.3. From the results we can see that $Rule(\{'vol\_Low', 'pct\_chg\_Medium'\} -> \{'swing\_Low'\})$ gets the highest value of lift. From the result we can infer that when the trading volume is small and the daily average rise and fall rate (pct_chg) is in the middle, the swing of HSI is lower. These rules provide insights into how different market conditions are interrelated, offering a deeper understanding of market dynamics that could be useful for traders and analysts in making predictions or identifying trading opportunities.

Table 4.3 Results of Frequent Pattern

antecedents	consequents	support	confidence	lift
frozenset({'swing_Low'})	frozenset({'pct_chg_Medium'})	0.17	0.50	1.51
frozenset({'pct_chg_Medium'})	frozenset({'swing_Low'})	0.17	0.51	1.51

frozenset({'vol_Low'})	frozenset({'swing_Low'})	0.19	0.58	1.70
frozenset({'swing_Low'})	frozenset({'vol_Low'})	0.19	0.57	1.70
frozenset({'vol_High'})	frozenset({'swing_High'})	0.19	0.57	1.71
frozenset({'swing_High'})	frozenset({'vol_High'})	0.19	0.57	1.71
frozenset({'pct_chg_Low', 'vol_High'})	frozenset({'swing_High'})	0.08	0.65	1.95
frozenset({'pct_chg_Low', 'swing_High'})	frozenset({'vol_High'})	0.08	0.55	1.64
frozenset({'vol_Low', 'swing_Low'})	frozenset({'pct_chg_Medium'})	0.11	0.55	1.65
frozenset({'vol_Low', 'pct_chg_Medium'})	frozenset({'swing_Low'})	0.11	0.70	2.05
frozenset({'swing_Low', 'pct_chg_Medium'})	frozenset({'vol_Low'})	0.11	0.62	1.86
frozenset({'pct_chg_High', 'vol_High'})	frozenset({'swing_High'})	0.08	0.63	1.91
frozenset({'pct_chg_High', 'swing_High'})	frozenset({'vol_High'})	0.08	0.65	1.95

5 Conclusion

The application of data mining techniques to the Hang Seng Index and Hong Kong Stock Market has demonstrated significant potential in enhancing investment strategies. Decision trees provided a reliable prediction of HSI movements with a mean accuracy of 0.6649, despite market noise, through careful feature selection and cross-validation. K-means clustering effectively identified seasonal trends, validated by temporal correlations with market instability between 2020 and 2022. Hierarchical clustering successfully grouped stocks by risk and return profiles, enabling the selection of stocks with favorable characteristics, such as higher returns and lower volatility. Association rule mining, using the Apriori algorithm, uncovered meaningful relationships, notably linking low trading volume and moderate price changes to reduced market swing. The integration of these models offers a comprehensive perspective on market dynamics, empowering investors with data-driven insights.

Future work could explore more features and their derived values and try to employ advanced algorithms such as machine learning or big models approach and try to explore real-time data to further improve prediction accuracy and applicability.

References

[1] Tsung-Sheng Chang,A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction, Expert Systems with Applications,Volume 38, Issue 12,2011, Pages 14846-14851, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2011.05.063.

[2] Kuo-Ping Wu, Yung-Piao Wu, Hahn-Ming Lee. Stock Trend Prediction by Sequential Chart Pattern via K-Means and AprioriAll Algorithm, 2012 Conference on Technologies and Applications of Artificial Intelligence, Tainan, Taiwan, 2012, pp. 176-181, doi: 10.1109/TAAI.2012.42.

[3] K. A. J. Doherty, R. G. Adams, N. Davey and W. Pensuwon, Hierarchical topological clustering learns stock market sectors, 2005 ICSC Congress on Computational Intelligence Methods and Applications, Istanbul, 2005, pp. 6 pp.-, doi: 10.1109/CIMA.2005.1662299.

[4] A. Srisawat, An application of association rule mining based on stock market, The 3rd International Conference on Data Mining and Intelligent Information Technology Applications, Macao, China, 2011, pp. 259-262.

[5] Tushare, "Tushare Pro: Financial Data Interface," Tushare Pro, [Online]. Available: https://tushare.pro/

[6] Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). Computers & Geosciences, 19(3), 303-342. https://doi.org/10.1016/0098-3004(93)90090-R

Appendix: Response to Peer Review Comments

This appendix summarizes our responses to peer review comments received for the project, detailing whether improvements were made based on the feedback and assessing the helpfulness of each comment. Comments with similar themes (e.g., suggestions for time series models) have been merged for clarity and conciseness.

Comment 1: Incomplete Model Evaluation

Reviewers noted that the decision tree evaluation only used mean accuracy and standard deviation, lacking metrics like precision, recall, F1 score, and ROC curves, and suggested that other model evaluations were similarly limited. They also recommended displaying feature and target distributions to assess dataset skewness or imbalance.

We respectfully disagree that the model evaluation is incomplete. Our decision tree evaluation (Section 4.1) focused on mean accuracy (0.6649) and standard deviation (0.0241) from 10-fold cross-validation, which are robust metrics for assessing classification performance in noisy stock market data. The project aimed to predict HSI directional movements (rise/fall) rather than optimize for imbalanced classes, and we verified during preprocessing (Section 3.1) that the dataset was not significantly skewed, making additional metrics like precision, recall, or ROC curves unnecessary.

The critique of incomplete evaluation was less applicable, as our chosen metrics aligned with the project's objectives. However, the comment prompted us to clarify the rationale for our evaluation choices, making it moderately helpful.

Comment 2: Lack of Time Series Models and Limited Model Variety

Reviewers suggested that stock market prediction is a classic time series problem and recommended using models like LSTM to capture long-term dependencies, arguing that decision trees may not perform well and that too few models were used.

We acknowledge the relevance of time series models like LSTM for stock prediction. However, our project intentionally focused on simple, interpretable models (decision trees, K-means clustering, hierarchical clustering, and association rule mining) to explore whether less complex methods could yield meaningful insights (Section 1.3). This choice was driven by the need for interpretability in investment decision-making, where decision trees and clustering provide clear rules and patterns. Our decision tree model achieved a mean accuracy of 0.6649 (Section 4.1), significantly outperforming random guessing (50% accuracy), demonstrating its predictive capability. Additionally, methods like K-means clustering revealed seasonal trends (Section 4.2). Incorporating LSTM would increase complexity without guaranteed improvements in our noisy stock data context. Due to time constraints and project scope, we did not add time series models but noted this as a future direction in Section 5.

These comments prompted a clearer justification of our model choices in Section 1.3, emphasizing simplicity and interpretability, and the random guessing comparison strengthened result credibility. However, the suggestion to adopt LSTM was not

implemented, as it diverged from our project's focus, making the feedback moderately helpful.

Comment 3: Concerns About Decision Trees and Overfitting

Reviewers questioned whether decision trees could capture meaningful patterns in noisy stock market data and how overfitting was addressed given market volatility.

Decision trees were chosen for their simplicity and ability to handle non-linear relationships, suitable for predicting HSI directional movements (Section 3.2). Compared to random guessing (50% accuracy), our decision tree model achieved a mean accuracy of 0.6649 (Section 4.1), indicating its ability to predict HSI rise/fall. To mitigate overfitting and market noise, we applied 10-fold cross-validation and limited tree depth, ensuring model stability across data subsets (standard deviation 0.0241). We clarified these anti-overfitting strategies in Section 3.2.3 to address the reviewer's concern, emphasizing the effectiveness of simple models in noisy data.

This feedback was highly valuable, as it led to a clearer explanation of decision tree predictive capability and overfitting mitigation, enhancing result credibility.

Comment 4: Lack of Comparison Between Methodologies

Reviewers suggested adding comparisons between the results of each methodology in the evaluation section.

We believe comparing methodology results is unnecessary. The project used two datasets (Hang Seng Index and Hong Kong stock information) and applied decision trees, K-means clustering, hierarchical clustering, and association rule mining to analyze the Hong Kong stock market from different perspectives (Section 1.3). These methods address distinct objectives (e.g., prediction, trend analysis, stock segmentation, feature relationship discovery) without strong interdependencies, aiming to provide multi-perspective market insights rather than direct comparisons. Thus, we did not add a comparison analysis in Section 4 but emphasized the complementary nature of the methods in Section 3.6 (Model Integration), highlighting their collective contribution to comprehensive analysis.

This comment prompted us to clarify the multi-perspective analysis goal in Section 3.6, enhancing the report's coherence. However, as comparisons were not aligned with the project's objectives, the suggestion was moderately helpful.

Overall, the peer review comments significantly improved the report's clarity and coherence. We sincerely appreciate the constructive feedback, which has strengthened the quality of our work.