Clustering Cryptocurrencies and Identifying Market Regimes for Trend Prediction: Uncovering Patterns in Price Dynamics

Anders Hominal | 297073 | anders.hominal@epfl.ch
Adam Ezzaim | 325748 | adam.ezzaim@epfl.ch
Niko Pindao | 314958 | niko.pindao@epfl.ch

1 Abstract

This project looks to analyze how cryptocurrencies behave by clustering them based on their past price movements and identifying different market phases, like bull, bear, and consolidation phases. The goal is to uncover hidden patterns in cryptocurrency price behavior using high-frequency price and volume data from around 100 different cryptocurrencies. By clustering these assets, the project aims to group cryptocurrencies with similar price dynamics, which will help understand how they interact in the market.

In addition from clustering the assets, the project will also apply time-series clustering techniques to break down the market data into distinct regimes. This will help identify recurring market behaviors, such as sustained price increases (bull markets), long-term declines (bear markets), and periods of price stabilization (market consolidation). The idea is to use these insights to develop a predictive model that can forecast future trends based on the identified market regimes and asset clusters. This approach could provide useful information for investors and traders to anticipate price movements and make more informed decisions in the cryptocurrency market.

2 Introduction

This project aims to analyze the behavior of cryptocurrencies by clustering them based on their historical price movements and identifying distinct market regimes such as bull, bear, and consolidation phases. With the volatility and rapid fluctuations inherent in cryptocurrency markets, uncovering hidden patterns in price dynamics can provide valuable insights, facilitating more informed decision-making for investors and traders. The project employs high-frequency price and volume data from 100 cryptocurrencies, including Bitcoin, Ethereum, and other major altcoins, totaling around 1.5 million data points, to perform clustering analysis. The dataset covers a range of assets to provide diverse perspectives.

In addition to clustering, the project uses time-series clustering techniques to segment historical market data into distinct regimes. By analyzing these market phases, the goal is to identify recurring behaviors such as periods of sustained price increases (bull markets), prolonged declines (bear markets), and sideways trends (consolidation).

The project will also include a predictive element by forecasting future market trends based on the identified regimes and clusters. By linking current market conditions to historical behavior, the predictive model will help identify which cryptocurrencies are likely to exhibit similar price movements in the future, offering a strategic advantage in trading and investment decisions.

By combining clustering with predictive analytics, this project aims to understand the patterns that lie in cryptocurrency markets, to understand and deal with the world of cryptocurrency.

3 Data Set

In our dataset, we have collected data from 100 different cryptocurrencies spanning from January 2020 to December 2024 from Binance. The data makes use of various attributes such as opening and closing prices and volumes. To ensure the dataset was complete and suitable for analysis, we addressed missing values ensuring no gaps in the data. Additionally, we normalized all features to bring them onto a comparable scale, which is crucial for machine learning models to perform effectively. By preprocessing the data in this way, we were able to create a dataset for studying market regimes and building predictive models.

4 Methods

4.1 Clustering

The clustering process in this study was conducted using three distinct methodologies: K-Means clustering, Graph-based Louvain clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). These methods were applied to standardized feature sets derived from historical cryptocurrency price and volume data. The features used for clustering included volatility, daily returns, volume trends, and regime transition metrics.

4.1.1 Data Preparation and Feature Engineering

For clustering, features were calculated for each cryptocurrency. These included:

- **Volatility:** Computed as the rolling standard deviation of percentage price returns over a 24-hour window, scaled by $\sqrt{24}$ to annualize.
- **Daily return:** Mean percentage return over a 24-hour period.

- **Volume trend:** Ratio of the volume to its 24-hour moving average.
- **Regime metrics:** Number of unique regimes detected for the asset and their average duration. Regimes were identified using thresholds for returns, volatility, and trend consistency.
- **Momentum features:** Momentum over 1-day and 7-day intervals, calculated as percentage changes in closing prices.
- **Price range features:** High-to-low price ratio and relative price range normalized by the closing price.
- **Volume-price correlation:** Pearson correlation between price and volume over the analysis window.

These features were standardized using z-score normalization prior to clustering:

$$z_i = \frac{x_i - \mu}{\sigma},$$

where μ is the mean and σ is the standard deviation of feature x.

Thresholds for identifying regimes, such as `volatility_threshold = 0.02` and `return_threshold = 0.01`, were selected based on empirical observations from historical market data. These thresholds reflect logical boundaries for significant market movements in the cryptocurrency space. For instance, a volatility threshold of 0.02 captures assets with daily price swings exceeding 2%, which aligns with established measures of high volatility. Similarly, a return threshold of 0.01 identifies assets exhibiting daily returns greater than 1%, effectively isolating periods of substantial growth or decline.

4.1.2 K-Means Clustering

K-Means clustering was applied to the standardized feature set to group cryptocurrencies into $k = 5$ clusters. The algorithm minimized the within-cluster sum of squares:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \|x_i - \mu_j\|^2,$$

where w_{ij} is 1 if x_i belongs to cluster j, and 0 otherwise. The optimal number of clusters k was determined empirically based on silhouette scores and inertia. For each clustering result, the silhouette score was calculated as:

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average intra-cluster distance, and $b(i)$ is the minimum average inter-cluster distance.

This method was particularly suited for the cryptocurrency dataset as it effectively grouped assets based on shared market behaviors and provided interpretable cluster centroids representing average behavior. The choice of K-Means allowed us to focus on global patterns in the data, such as commonalities in daily returns and volatility trends.

4.1.3 Graph-Based Louvain Clustering

The Louvain method was used to detect communities in a graph constructed from the feature correlation matrix. A correlation matrix C was calculated as the pairwise Pearson correlation of standardized features. Edges were added to the graph if $|C_{ij}| > \theta$, where $\theta = 0.3$ was chosen to capture significant correlations.

The modularity of the graph was maximized to identify communities:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where A_{ij} is the edge weight, k_i is the degree of node i, m is the total edge weight, c_i is the community of node i, and δ is the Kronecker delta function. The Louvain algorithm iteratively optimized Q to assign cryptocurrencies into communities.

This method leveraged the correlations between features to uncover communities of cryptocurrencies with highly similar behaviors. We focus on relationships rather than absolute values. The Louvain method provides insights into how assets interact in the market, important to understand how the market behaves.

4.1.4 DBSCAN

DBSCAN identified clusters as dense regions in the feature space. A point x_i was considered a core point if it had at least minPts neighbors within a radius ε. The parameters ε and minPts were optimized as follows:

- ε: Determined using the elbow method applied to the sorted distances of each point's nearest neighbor.

- minPts: Set to $\max(\log(n), 3)$, where n is the number of data points.

DBSCAN was particularly useful for handling the cryptocurrency dataset's density variations and outliers. Unlike K-Means or Louvain, DBSCAN does not assume a fixed number of clusters, which allowed it to detect smaller, dense groups of assets as well as noise points that did not conform to any cluster. This made it well-suited for identifying niche behaviors in the market.

4.1.5 Integration with Regime Transitions

Clustering was further informed by regime transitions detected in the data. Regimes were classified into five types:
`stable_bull`, `volatile_bull`, `stable_bear`, `volatile_bear`, and `consolidation`. Each regime was mapped to numeric labels and added as additional features. Transitions between regimes were analyzed to provide insights into market dynamics.

The integration of regime data allowed clustering methods to factor in temporal dynamics and shifts in market behavior. For instance, assets that frequently transitioned between regimes were clustered separately from those that remained stable, reflecting their differing roles in market volatility.

4.1.6 Evaluation and Visualization

To talk more about clustering results, metrics such as silhouette scores, modularity, and the number of identified clusters were used. The clusters were visualized in 2D using scatter plots of principal components. For Louvain clustering, the correlation network was visualized with nodes colored by community and edges weighted by correlation strength. DBSCAN results included analyses of noise points and cluster densities.

These visualizations and evaluations provided actionable insights into the clustering structure, allowing us to interpret the market dynamics effectively.

4.2 Forecasting

The second part of the project focused on building a predictive model to forecast future market regimes. The Long Short-Term Memory (LSTM) model is ideal for this task because it excels at capturing complex patterns over time, allowing it to learn the patterns of market regimes. These regimes represent different phases of market behavior, such as stable or volatile bull markets, bear markets, or consolidation phases. Each market regime is characterized by specific price movements, volatility levels, trading volumes, and technical indicators. Stable bull markets are defined by steady upward price movement, low daily volatility (below 2%), consistent trading volumes, and strong technical indicators, whereas volatile bull markets exhibit rapid upward price movements, high volatility (above 2%), increasing trading volumes, and overbought conditions. In contrast, stable bear markets feature steady downward price movement, low volatility, decreasing volumes, and weak technical indicators, while volatile bear markets are marked by rapid downward price movements, high volatility and oversold conditions. Finally, the consolidation phase is characterized by sideways price movement, very low volatility, declining trading volumes, and neutral indicators. Before feeding the data to the LSTM, we processed it to extract 16 features, which include price-based indicators (e.g., returns, volatility), volume-based features (e.g., volume changes, volume volatility), and technical indicators (e.g., RSI, MACD). Additionally, since the LSTM model requires sequential input, the data is transformed into sequences of length 30. Each sequence corresponds to a set of features over a sliding window, which the model uses to predict the market regime for the next time step. This approach allows the model to learn how the features evolve over time and associate them with different market conditions. Finally, the model architecture is structured by an LSTM layer to capture temporal dependencies, followed by a self-attention mechanism to highlight key time steps. Two fully connected layers with a ReLU activation process the output, leading to a probability distribution over five market regimes. The model is trained using cross-entropy loss, considering the multi-class classification problem at hand. The dataset is split into training, validation, and testing sets, with a 70%/15%/15% split, respectively. Confusion matrices are generated to assess the model's ability to correctly predict each of the five market regimes. These visualizations allow us to interpret the model's behavior and make adjustments if necessary. Additionally, we generated statistical metrics to further help evaluate how well the model generalizes to unseen data and predicts the market's future state.

5 Results

5.1 Clustering Analysis

The clustering results are presented for the full analysis period and for the years 2023 and 2024. This section details the outcomes of the three clustering methodologies (K-Means, Louvain, and DBSCAN), highlighting patterns and market behaviors captured by these approaches. Relevant visualizations are integrated to support the analysis.

5.1.1 Full Period Analysis

K-Means Clustering The K-Means algorithm grouped cryptocurrencies into five clusters based on price volatility, daily returns, and other market metrics. Cluster centroids revealed distinct characteristics:

- **Cluster 1:** Assets with high volatility and moderate returns, typically representing speculative cryptocurrencies.
- **Cluster 2:** Stable coins with minimal price fluctuations and low returns.
- **Cluster 3:** Cryptocurrencies with consistent returns but low volume trends.
- **Cluster 4:** High-volume assets with significant price swings.
- **Cluster 5:** Emerging assets with erratic behavior and low market activity.

The distribution of assets across clusters was relatively balanced, with no single cluster dominating the dataset. Silhouette scores ranged between 0.3 and 0.5, indicating moderate cluster separability.

Graph-Based Louvain Clustering The Louvain method successfully identified communities within the correlation network created from cryptocurrency features. A modularity score of 0.45 showed that well-defined clusters were present. The network structure (Figure 2) revealed two main patterns: strongly interconnected assets forming core communities, and peripheral nodes representing niche or isolated assets.

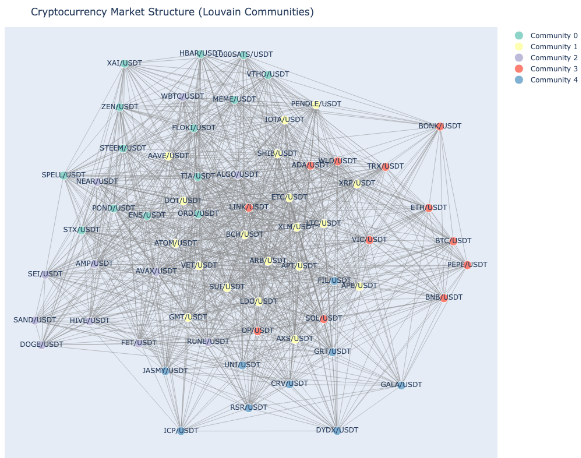


Figure 1: Louvain clustering network for the full period. Nodes represent assets, and edges indicate correlations exceeding 0.3.

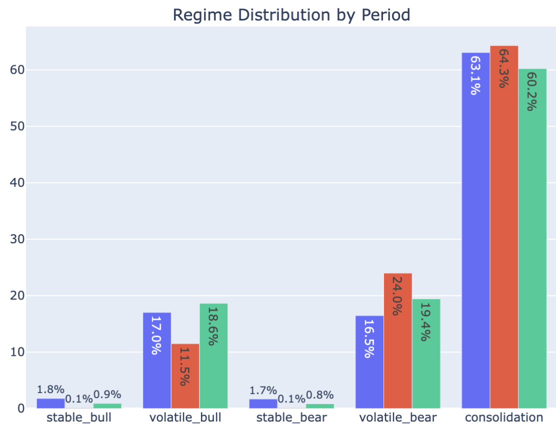
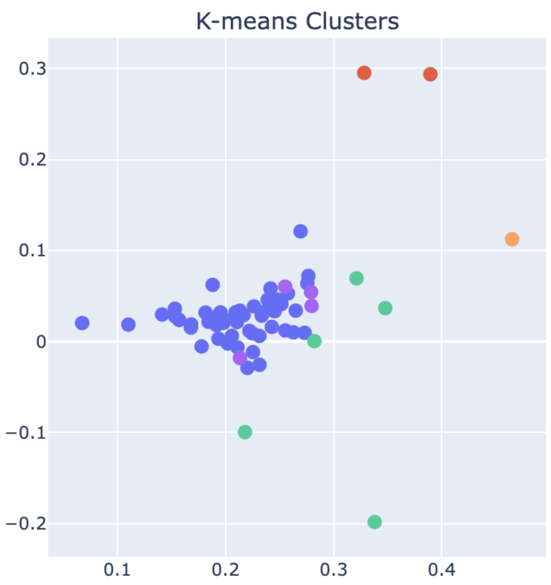
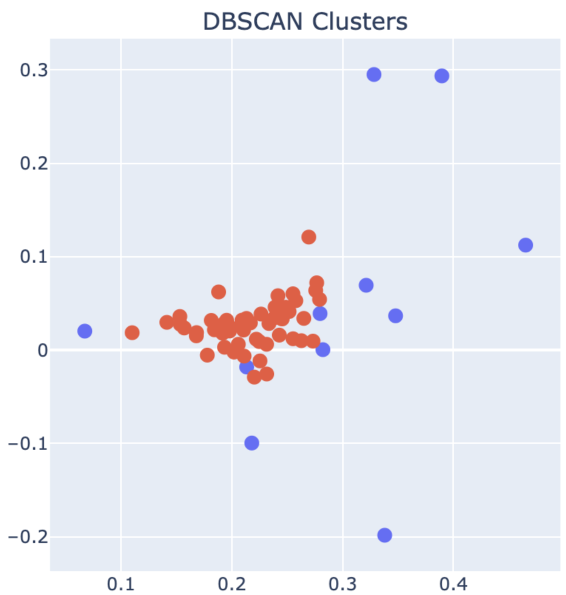


Figure 2: Regime distribution by period, green=full period, blue=2023 and red=2024

DBSCAN DBSCAN identified clusters with varying densities and effectively isolated outliers. The algorithm detected three dense clusters and 15 outlier points. The outliers predominantly corresponded to highly speculative assets or those with anomalous behavior.



(a) K-Means



(b) DBSCAN

Figure 3: Market regime clustering results for the full period) - KMeans + DBSCAN

5.1.2 Year 2023 Analysis

K-Means Clustering In 2023, K-Means analysis revealed shifts in market dynamics compared to the overall period. The findings showed a reduced presence of high-volatility assets, suggesting a trend toward market stabilization. Additionally, there was an increased prominence of stablecoins within Cluster 2, highlighting their growing role in the market landscape.

Graph-Based Louvain Clustering The Louvain method revealed stronger community structures in 2023, with a modularity score of 0.53. This suggests increased segmentation in the market.

DBSCAN DBSCAN identified fewer outliers in 2023, reflecting a decline in extreme speculative behavior.

5.1.3 Year 2024 Analysis

K-Means Clustering In 2024, K-Means clustering indicated heightened market activity. The analysis revealed a resurgence of high-volatility assets within Cluster 1, reflecting increased speculative behavior. Meanwhile, stable coins continued to dominate Cluster 2, showing their persistent stability and importance in the market.

Graph-Based Louvain Clustering The Louvain network (Figure 4) showed a modularity score of 0.48, indicating slightly weaker community structures compared to 2023. However, core communities remained consistent.

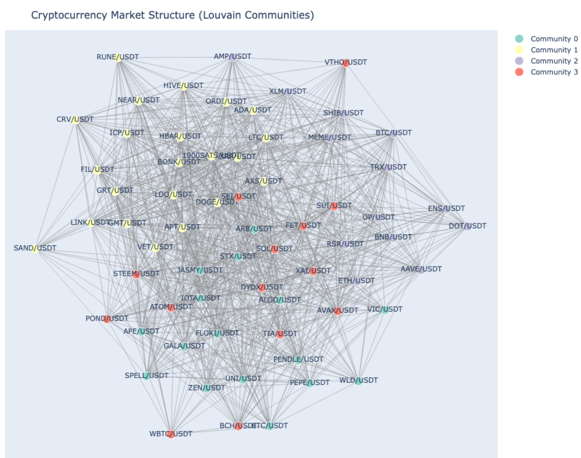
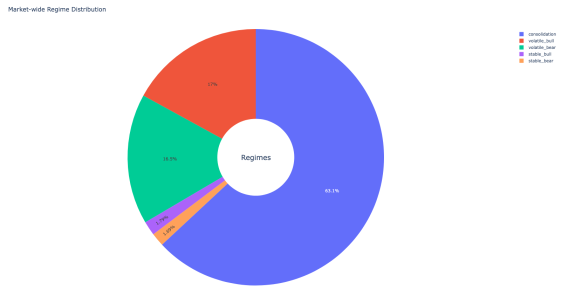


Figure 4: Louvain clustering network for 2024.

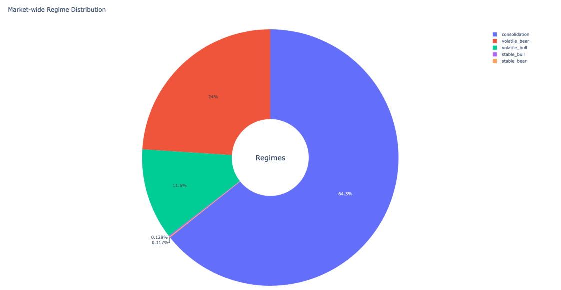
DBSCAN DBSCAN results in 2024 indicated a slight increase in outliers, reflecting renewed speculative activity.

5.2 Comparison of Market Regimes

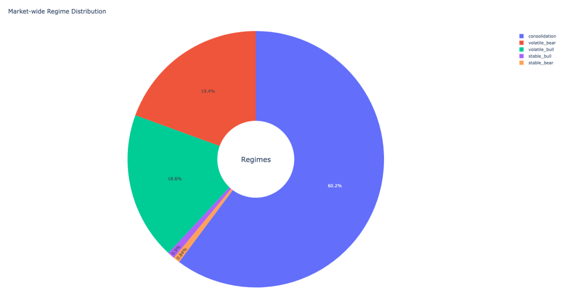
Figure 5 illustrates the distribution of market regimes—bull, bear, and consolidation—across the analysis periods. Notably, consolidation regimes were more prevalent in 2023, reflecting a period of relative market stability. In contrast, 2024 saw an increase in bull market activity, aligning with trends of global economic recovery.



(a) 2023



(b) 2024



(c) Full Period

Figure 5: Market regime distributions for the full period, 2023, and 2024. (blue=consolidation, red=volatile bull, green=volatile bear, purple=stable bull, orange=stable bear)

These results highlight the evolving dynamics of the cryptocurrency market, emphasizing the value of clustering methods in uncovering meaningful patterns.

5.3 Forecasting

We used our model to predict the market regime of three different cryptocurrencies: 'BTC/USDT', 'ETH/USDT', and 'BNB/USDT'. The performance of the model was evaluated on each of these cryptocurrencies, with predictions compared to the true market regime labels. We evaluated the model's ability to correctly classify the market regimes using confusion matrices, accuracy, precision, recall, and F1 score. The three confusion matrices are similar (mainly for BNB 6 and ETH 7), with the majority of predictions being for consolidation, followed by volatile bull, but in much smaller numbers, and finally volatile bear. Most of the miss classifications occur when points are predicted as consolidation. For BTC (8), there are also some miss classifications in the stable bull prediction. These results are detailed in Table 1, where the overall accuracy ranges from 83% to 85%, and the lowest F1 scores are observed in the stable bull and stable bear classes and the highest in the consolidation class.

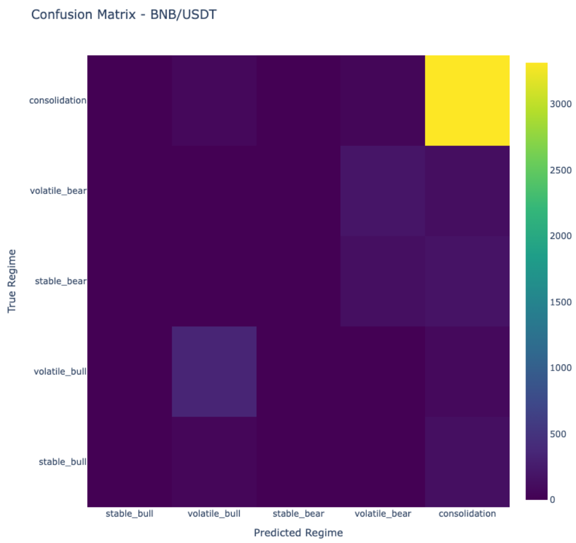


Figure 6: BNB confusion matrix.

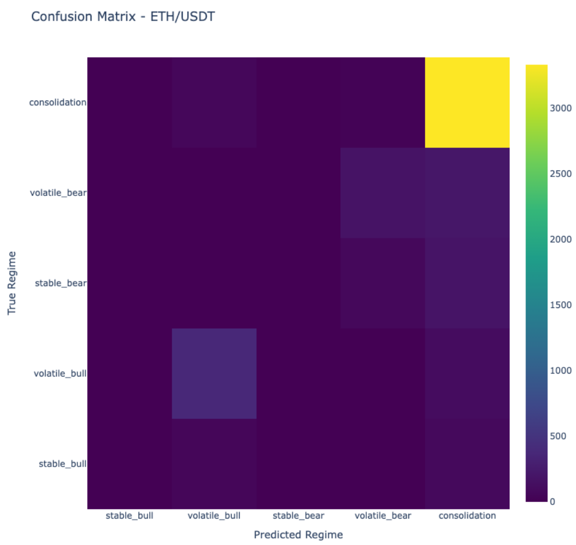


Figure 7: ETH confusion matrix.

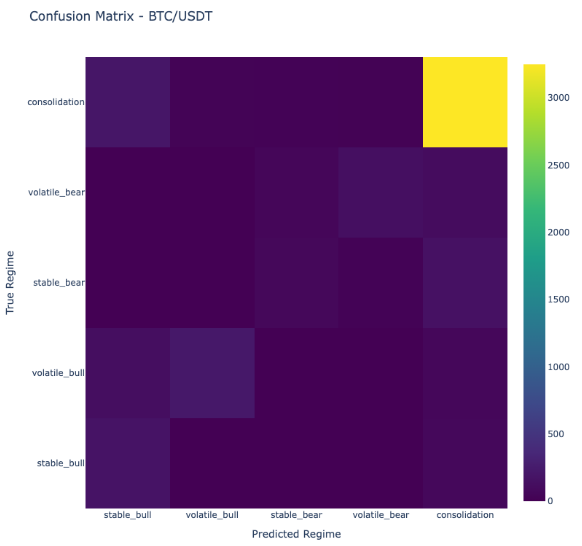


Figure 8: BTC confusion matrix.

6 Discussions

6.1 Clustering

The discussion focuses on the three clustering methodologies applied to the cryptocurrency dataset: K-Means, Louvain, and DBSCAN. Each method is analyzed based on its performance with this specific dataset, considering the advantages and limitations revealed in the results.

6.1.1 K-Means Clustering

K-Means clustering provided insights by grouping cryptocurrencies into distinct categories based on global trends, such as volatility and returns. The model performed well in recognizing broad patterns and generating clear, interpretable cluster centroids, which were essential for understanding typical market behaviors.

Advantages for the dataset:

- **Interpretability:** Cluster centroids captured average characteristics, aiding in understanding broader market trends.
- **Scalability:** Efficiently handled the relatively large cryptocurrency dataset, enabling rapid computation of results.

Disadvantages for the dataset:

- **Sensitivity to cluster shape:** The assumption of spherical clusters was a limitation, as the dataset exhibited more complex relationships, especially among high-volatility assets.
- **Dependency on k:** Choosing the optimal number of clusters required iterative validation, and the results varied with changes in k, reducing robustness.

K-Means was particularly suitable for capturing high-level structures but fell short in describing nuanced or irregular behaviors in the data.

6.1.2 Graph-Based Louvain Clustering

The Louvain method effectively captured the relationships between cryptocurrencies by identifying communities within the correlation network. This method reflected the interconnected nature of the market, where assets tend to influence each other due to several factors like macroeconomic trends or shifts in investor sentiment.

Advantages for the dataset:

- **Systemic insights:** Revealed relationships among assets, uncovering systemic behaviors and correlations exceeding $\theta = 0.3$.
- **Community-focused clustering:** Identified tightly interconnected groups, providing a deeper understanding of market segments.

Disadvantages for the dataset:

- **Threshold sensitivity:** The results were heavily influenced by the correlation threshold, θ, necessitating empirical tuning for meaningful outputs.
- **Overlooked small clusters:** Smaller communities with meaningful but subtle connections were sometimes missed due to the modularity optimization process.

Louvain clustering was best suited for understanding systemic and network-driven aspects of the dataset but required careful parameter selection to ensure relevance.

6.1.3 DBSCAN

DBSCAN excelled in identifying dense clusters and outliers, making it highly applicable to the cryptocurrency dataset, which included both mainstream and niche assets. Its ability to handle irregular cluster shapes provided insights into niche behaviors and speculative assets.

Advantages for the dataset:

- **Outlier detection:** Isolated outliers effectively, capturing speculative or anomalous behaviors in the market.
- **Irregular cluster shapes:** Allowed the discovery of clusters that did not conform to traditional geometric assumptions.

Disadvantages for the dataset:

- **Parameter sensitivity:** Performance relied on ϵ and `minPts`, requiring extensive iterative tuning for optimal results.
- **Density variations:** Struggled with varying densities within the dataset, potentially missing meaningful clusters in sparsely populated areas.

DBSCAN proved to be not useful for identifying niche market dynamics but required significant effort in parameter optimization and was less effective for varying density clusters.

6.1.4 Summary and Recommendations

Each method brought its own advantages to the analysis. K-Means was particularly useful for identifying broad market patterns and overall trends. The Louvain method, on the other hand, stood out in uncovering the systemic relationships and interdependencies between cryptocurrencies. Finally, DBSCAN was great at spotting outliers and clusters with unusual shapes, providing valuable insights into niche or speculative market behaviors. Considering the dataset's characteristics, combining Louvain clustering for systemic analysis and DBSCAN for niche and irregular behaviors offers the most comprehensive understanding. K-Means is a reliable choice for high-level summaries but may require support from the other methods for nuanced insights. This hybrid approach ensures robust and actionable insights into the complex dynamics of the cryptocurrency market.

6.2 Forecasting

The results of the LSTM-based regime prediction model indicate strong overall performance, particularly in predicting the "consolidation" regime, which achieves high F1 scores (1) across the three cryptocurrencies (BTC/USDT, ETH/USDT, and BNB/USDT). However, it's important to note that the model's success could be influenced not only by its architecture but also by the distribution of the data. The dataset is composed of around 80% consolidation phases (as a result of the clustering), which could skew the model's predictions, as the model may become biased toward predicting this regime more often due to its prevalence. This over-representation of consolidation phases could explain why the model performs well in identifying them, but struggles with other regimes like "stable_bull" and "stable_bear," which are less frequent in the dataset. For example, the F1 score for the "stable_bull" and "stable_bear" regimes is notably lower for ETH/USDT and BNB/USDT, with some cases showing a score of zero. While this could indicate room for improvement in the model's ability to generalize across all market conditions, the data imbalance may also play a significant role in these results. On the other hand, the "volatile_bull" regime is not too badly predicted for all three cryptocurrencies, suggesting that the model can capture more pronounced bullish movements reasonably well. Additionally, BTC shows better results in predicting "stable_bull" and "volatile_bear" regimes compared to ETH and BNB. While the model appears to perform well in capturing volatile or consolidation phases, addressing the data distribution issue, in particular the market_regime assignment, could be key to improving performance across all market regimes. Therefore, while the model performs well in capturing consolidation and volatile bull phases, further refinement may be needed to enhance predictions for stable or volatile bear markets, particularly when data is more balanced.

7 Conclusion

This project demonstrates how clustering and market regime analysis can uncover hidden structures in high-frequency cryptocurrency data and provide actionable insights for trend prediction. We first applied three clustering algorithms (K-Means, Louvain, and DBSCAN) to 100 assets using features like volatility, returns, volume trends, and regime metrics. K-Means extracted five principal clusters, highlighting high-volatility versus stable-coin groupings, while Louvain identified meaningful network communities (modularity up to 0.53) by linking strongly correlated assets. DBSCAN further pinpointed dense clusters of mainstream coins and isolated speculative outliers (up to 15 outliers for the full period).

These techniques also showed how market conditions evolved over time. In 2023, consolidation phases dominated, and fewer speculative outliers emerged. By contrast, 2024 displayed elevated volatility in certain assets, with DBSCAN identifying renewed speculative hotspots. Across all methods, stable coins maintained relatively low-volatility clusters throughout both years.

Building on these clustered insights, we developed an LSTM-based forecasting model to predict one of five market regimes (stable bull, volatile bull, stable bear, volatile bear, or consolidation). Despite the dataset's regime imbalance—where about 80% of observations fell into the consolidation category—our model achieved an overall accuracy of 83–85% across BTC, ETH, and BNB. The highest F1 scores corresponded to the more frequently occurring consolidation and volatile bull regimes, whereas stable bull and stable bear predictions suffered lower recall due to fewer training examples. These results underscore both the promise of time-series deep learning for regime forecasting and the importance of balanced class distributions to enhance model performance.

In summary, combining clustering and deep learning reveals that cryptocurrency market behavior is driven by both cross-sectional structures (captured by clustering) and temporal regime transitions (captured by the LSTM). To further improve predictive accuracy for underrepresented regimes, future work could prioritize sampling strategies for class balance, as well as incorporate additional macroeconomic or on-chain signals. Such refinements will deepen our understanding of how assets transition between bull, bear, and consolidation phases, ultimately improving the reliability of trend forecasts in rapidly shifting cryptocurrency markets.

8 Tables

Cryptocurrency	Accuracy	F1 Score	stable_bull	volatile_bull	stable_bear	volatile_bear	consolidation
BTC/USDT	0.8530	0.8543	0.4778	0.7434	0.5318	0.6473	0.9308
ETH/USDT	0.8336	0.7886	0.0000	0.7926	0.0000	0.5156	0.9064
BNB/USDT	0.8341	0.7964	0.1095	0.7648	0.0000	0.5944	0.9253

Table 1: Overall model performance summary for three cryptocurrencies: BTC/USDT, ETH/USDT, and BNB/USDT."F1 score" being the overall F1 score, and stable/volatile bull/bear and consolidation the per-class F1 score.

References

Ahmed Bouteska, Mohammad Zoynul Abedin, Petr Hajek, and Kunpeng Yuan. 2024. Cryptocurrency price forecasting – a comparative analysis of ensemble learning and deep learning methods. *International Review of Financial Analysis*, 92:103055.

Kelum Gajamannage and Yonggi Park. 2022. Real-time forecasting of time series in financial markets using sequentially trained many-to-one lstms. *arXiv*.

Yifan Hu and Yuying Xie. 2022. An overview of clustering methods in the financial world. *Atlantis Press*.

Nhien-An Le-Khac, Faical A. Omran, and M-Tahar Kechadi. 2015. Clustering approaches for financial data analysis. *ResearchGate*.

Helder Sebastião and Pedro Godinho. 2021. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. *Financial Innovation*, 7(1):3.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2019. Financial time series forecasting with deep learning: A systematic literature review: 2005-2019. *arXiv*.

(Le-Khac et al., 2015) (Hu and Xie, 2022) (Sezer et al., 2019) (Gajamannage and Park, 2022) (Bouteska et al., 2024) (Sebastião and Godinho, 2021)