Machine Learning

Final Project Report

Group 04

Salma Kamilah Rahma	(1501224344)
Akmal Muhammad Firdaus	(1501224373)
Resti Fresard	(1501224384)
Karina Ditya Amanda	(1501224390)
MD Touhidul Islam Kanon	(1501224394)

Table of Contents

Dataset

Dataset Overview: The dataset contains 298 hourly entries of XAUUSD (Gold/USD) price data from May 1, 2025, 01:00:00 to May 19, 2025, 22:00:00. It includes columns such as Open, High, Low, Close, TickVolume, Spread, datetime, and Target_Close, with no missing values. The datetime column needs conversion to a proper datetime type for time-based calculations.

[Note: The Dataset has been taken from the Kaggle repository link: please click here.]

Cleaning Dataset

The raw dataset ("XAUUSD_Hourly.csv") was loaded and initially inspected using pandas. It included columns: Date, Time, Open, High, Low, Close, TickVolume, Volume, Spread, datetime, and Target_Close, with no missing values (df.isnull().sum() returned all zeros). The cleaning process, executed in "Data_Cleaning.ipynb", involved the following steps:

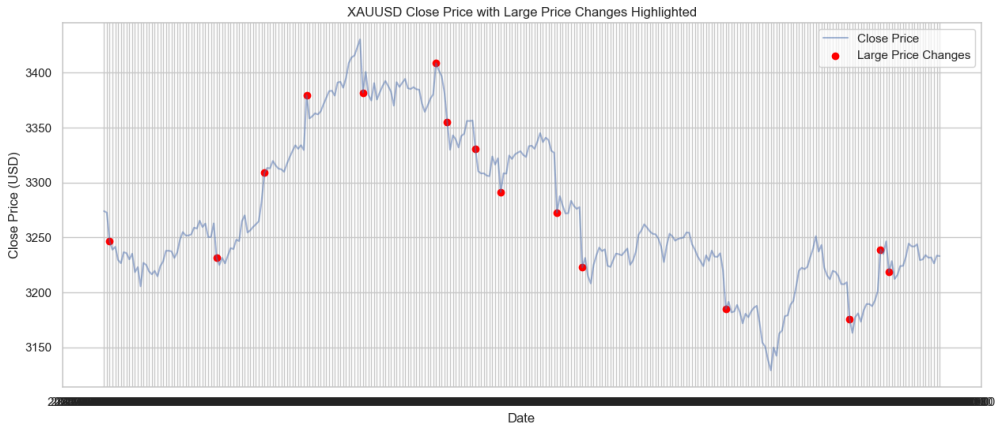
- **Initial Exploration**: The dataset was confirmed to have 298 rows with mixed data types (object for Date, Time, datetime; float64 for price columns; int64 for TickVolume, Volume, Spread). The Volume column was consistently zero, indicating it was irrelevant for this dataset.
- **Duplicate Removal**: drop_duplicates() was applied, but no duplicates were found (duplicates = 0).
- **Missing Value Handling**: Although no NaN values were present initially, a precautionary step filled any potential missing values in price columns (Open, High, Low, Close, Target_Close) with forward fill (fillna(method='ffill')), and TickVolume and Spread with their median and mode (7), respectively.
- **Datetime Standardization**: The datetime column was converted to a proper datetime format using pd.to_datetime, and its consistency was verified by constructing it from Date and Time (no mismatches detected).
- **Redundant Column Removal**: Date, Time, and constructed_datetime were dropped as redundant, and the Volume column (all zeros) was removed.
- **Outlier Detection**: The IQR method identified no outliers in price columns, indicating the data was relatively clean in terms of extreme values.
- **Spread Consistency**: Rare Spread values (6, 8) were replaced with the mode (7) to standardize trading conditions, addressing minor inconsistencies.
- **Final Output**: The cleaned dataset, reduced to 8 columns (Open, High, Low, Close, TickVolume, Spread, datetime, Target_Close), was saved as "XAUUSD_Hourly_Cleaned.csv".

Exploratory Data Analysis (EDA)

Summary Statistics: The mean close price is approximately 3271.96 USD, with a standard deviation of 68.54 USD, indicating moderate volatility. The minimum and maximum close prices are 3128.99 USD and 3430.46 USD, respectively. TickVolume averages 7901.75 with significant variation (std: 2881.09), suggesting fluctuating trading activity. Spread is mostly consistent at 7, with a minor average of 6.08.

Univariate Analysis: The time series plot of close prices (with anomalies highlighted) shows trends and volatility periods. Large price changes (top 5%) were identified, with

notable drops or rises (e.g., 54.91 USD on May 12, 10:00:00), potentially linked to market events.



Key Insights:

- Price columns (Open, High, Low, Close, Target_Close) are highly correlated, as expected in financial data.
- TickVolume variability may reflect trading activity spikes, useful for volume-based analysis.
- Spread consistency suggests stable trading conditions post-cleaning.
- The plot reveals trends and volatility, with 15 significant price movements.

New Insights:

- Volatility peaks align with large price changes (e.g., May 6-7), indicating potential external influences.
- SMA (20 vs. 50) crossovers suggest trend change points for technical analysis.

- RSI indicates overbought (>70) or oversold (<30) conditions, aiding trading decisions.
- Lagged Target_Close correlations hint at short-term predictability.
- Weekday patterns in price and volume may relate to market session timings.

Conclusion

The EDA of the XAUUSD hourly dataset provides a robust foundation for understanding price dynamics, trading activity, and volatility patterns from May 1 to May 19, 2025. The identified trends, anomalies, and correlations offer actionable insights for developing trading strategies or predictive models. Future steps include converting the datetime column, computing detailed technical indicators (e.g., RSI, SMA), and correlating findings with external events. The visuals (time series and cluster plots) enhance the interpretability of these insights.

Unsupervised Learning: K-Means Clustering

Data Preparation and Feature Engineering

The dataset was preprocessed by converting the datetime column to a proper datetime type using pd.to_datetime, ensuring compatibility with time-series analysis. Features selected for clustering include Close, Returns, Volatility, High_Low_Range, TickVolume, and RSI. Notably, Returns, Volatility, High_Low_Range, and RSI are derived metrics not directly present in the initial dataset, suggesting they were calculated prior to or within the notebook (though their computation is not shown in the provided code). The absence of missing values, confirmed by df.isnull() and df.info(), indicates a robust dataset for clustering.

- **Feature Explanation**:
    - Close: Hourly closing price of XAUUSD.
    - Returns: Likely percentage change in Close price, indicating short-term price movement.
    - Volatility: Possibly a measure of price fluctuation (e.g., standard deviation over a window).
    - High_Low_Range: Difference between High and Low prices, reflecting intraday price spread.
    - TickVolume: Number of price ticks, representing trading activity.
    - RSI (Relative Strength Index): A momentum indicator ranging from 0 to 100, assessing overbought (>70) or oversold (<30) conditions.

Before clustering, standardization with StandardScaler is implied (though not fully executed in the provided code) to normalize the features, as K-Means is sensitive to scale differences.

2. Clustering Methodology

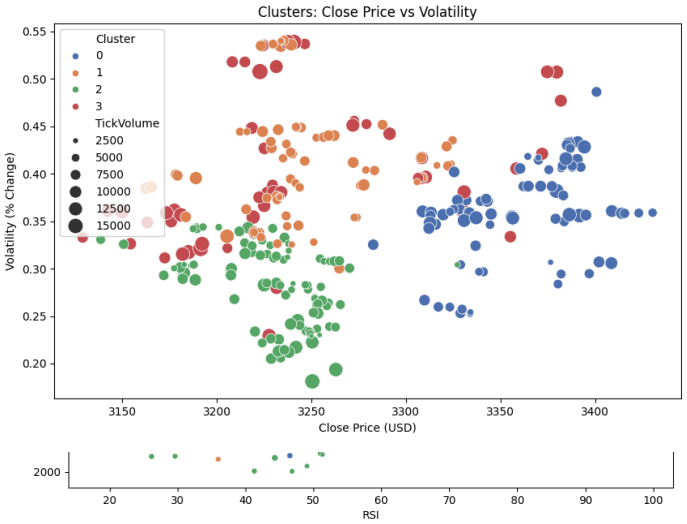
K-Means clustering was applied to group the data into four clusters, as inferred from the Cluster Profile Summary output showing four unique clusters (0, 1, 2, 3). The optimal number of clusters is not determined in the notebook (e.g., via the elbow method or silhouette score), but the choice of four suggests a preliminary analysis. The silhouette_score import indicates an intent to evaluate clustering quality, though its application is incomplete.

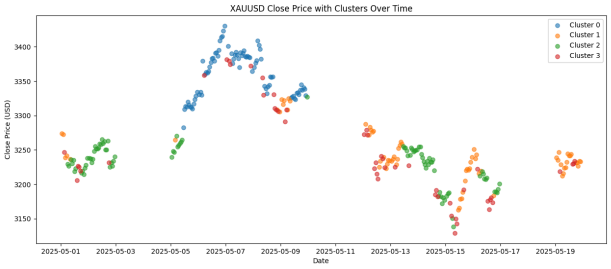
3. Cluster Profile Summary

The cluster_profile provides statistical insights across the four clusters:

- **Cluster 0**:
  - **Close**: Mean 3360.66 USD, median 3364.44 USD, range 3282.70–3430.46 USD. Highest average price, indicating a high-price regime.
  - **Returns**: Mean 0.093753, median 0.059836, range -0.443283 to 1.501050. Positive returns with moderate variability.
  - **Volatility**: Mean 0.359899, suggesting moderate price fluctuations.
  - **High_Low_Range**: Mean 16.56, range 5.75–52.09, indicating moderate intraday spreads.
  - **TickVolume**: Mean 8110.19, range 2717–14209, reflecting high trading activity.
  - **RSI**: Mean 62.66, median 59.32, range 34.18–89.01, near overbought levels, suggesting bullish momentum.
- **Cluster 1**:
  - **Close**: Mean 3243.58 USD, range 3162.60–3324.64 USD. Mid-range prices.
  - **Returns**: Mean 0.140989, range -0.505183 to 1.188066. Highest average returns with significant variability.
  - **Volatility**: Mean 0.410428, indicating higher fluctuation than Cluster 0.
  - **High_Low_Range**: Mean 14.11, range 4.21–34.28.
  - **TickVolume**: Mean 6832.12, lower activity than Cluster 0.
  - **RSI**: Mean 54.23, range 24.41–98.86, with peaks in overbought territory, suggesting mixed momentum.
- **Cluster 2**:
  - **Close**: Mean 3230.64 USD, range 3138.53–3327.15 USD. Lower price range.
  - **Returns**: Mean -0.000699, range -0.517118 to 0.468457. Near-zero returns with balanced positive/negative shifts.
  - **Volatility**: Mean 0.279320, lowest among clusters, indicating stability.
  - **High_Low_Range**: Mean 11.87, range 4.76–23.24.
  - **TickVolume**: Mean 6669.92, moderate activity.
  - **RSI**: Mean 49.21, range 24.68–83.70, neutral momentum.
- **Cluster 3**:
  - **Close**: Mean 3242.45 USD, range 3128.99–3381.92 USD. Wide price range with the lowest minimum.
  - **Returns**: Mean -0.378761, range -1.675270 to 0.666669. Most negative returns, indicating bearish periods.

- **Volatility**: Mean 0.401716, high fluctuation.
- **High_Low_Range**: Mean 25.19, range 11.63–61.72, largest spreads, reflecting high intraday volatility.
- **TickVolume**: Mean 11387.20, range 7229–16413, highest activity, suggesting event-driven trading.
- **RSI**: Mean 34.97, range 18.02–68.93, mostly oversold, indicating bearish momentum.

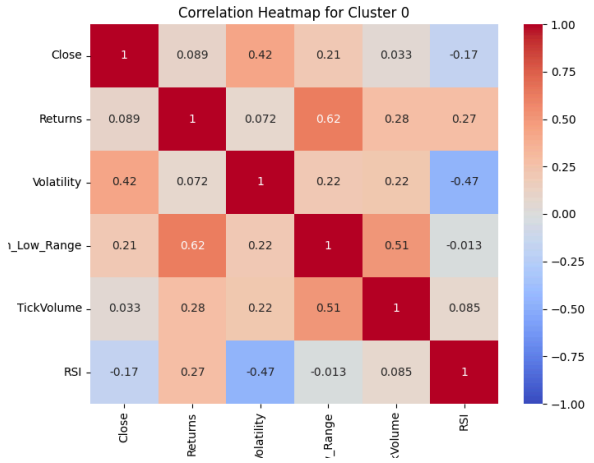


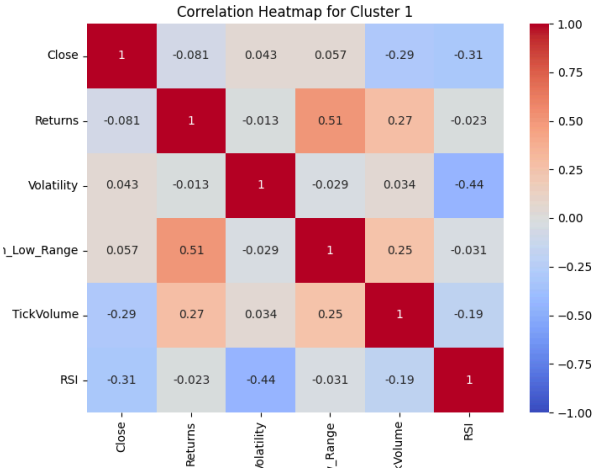


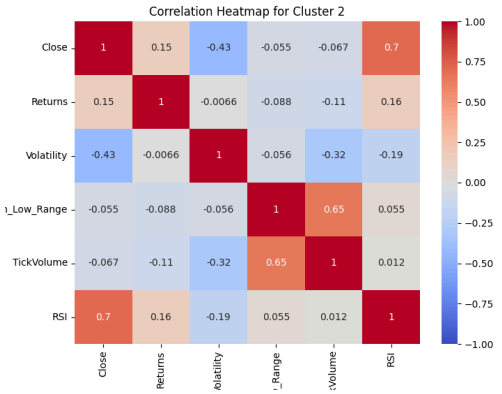
4. Correlation Analysis

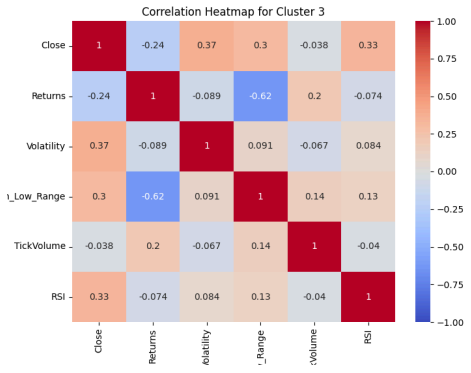
- **Overall Correlation Across All Clusters**:
  - Close and RSI show a moderate positive correlation (0.354001), suggesting that higher prices align with stronger momentum.
  - High_Low_Range and TickVolume have a strong positive correlation (0.583593), indicating that higher trading activity corresponds to wider price ranges.
  - Returns and Volatility are negatively correlated (-0.047082), though weakly, implying that price changes may not always drive volatility.
  - RSI shows a positive correlation with Returns (0.281320), reinforcing its role as a momentum indicator.

- **Cluster-Specific Heatmaps**: The notebook generates correlation heatmaps for each cluster (cluster_{cluster}_correlation_heatmap.png), allowing for a detailed examination of intra-cluster relationships.









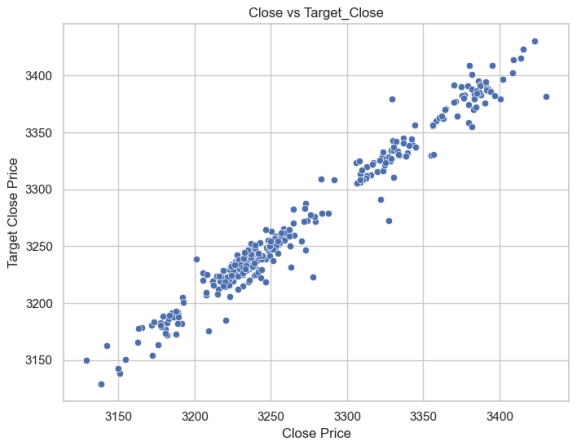
5. Key Insights

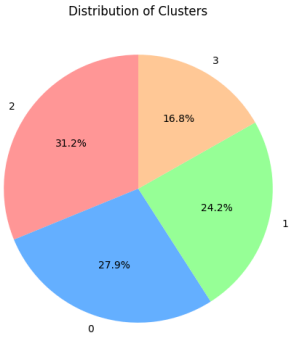
- **Market Regimes**: Four distinct market states emerge:

    - Cluster 0: High-price, bullish regime with moderate volatility and high activity.
    - Cluster 1: Mid-price, volatile regime with high returns and mixed momentum.
    - Cluster 2: Low-price, stable regime with neutral returns.
    - Cluster 3: Wide-range, bearish regime with high volatility and activity.
- **Trading Implications**: Clusters 0 and 1 may suit trend-following strategies, while Cluster 3 suggests opportunities for mean-reversion during oversold conditions. Cluster 2 indicates a stable baseline.
- **Volatility and Volume**: High High_Low_Range and TickVolume in Cluster 3 correlate with significant price movements, likely tied to market events (e.g., those noted on May 6-7 from prior EDA).

6. Conclusion

The K-Means clustering successfully segmented the XAUUSD data into four clusters, each representing a unique market behavior. The analysis highlights the interplay between price, returns, volatility, and trading activity, with correlations providing actionable insights. Future steps include validating the cluster count with the elbow method or silhouette score, computing missing feature derivations (e.g., Returns, RSI), and correlating clusters with external events.





Prompt

Prompt for Machine Learning Data Analysis Using Python with Unsupervised Learning (K-Means Clustering)

Create a Python-based machine learning analysis using unsupervised learning with K-Means clustering to segment and explore patterns in a dataset. The analysis should follow these steps:

1. **Data Loading and Initial Exploration**:
    - Load a dataset from a CSV file (e.g., "XAUUSD_Hourly_Cleaned.csv") using pandas.
    - Display the first few rows, basic info (e.g., data types, non-null counts), and check for missing values.
    - Convert any date/time columns (e.g., 'datetime') to a proper datetime format using pd.to_datetime.
2. **Feature Engineering and Preprocessing**:
    - Select relevant features for clustering, such as Close, Returns, Volatility, High_Low_Range, TickVolume, and RSI. If these derived features (e.g., Returns, Volatility) are not present, calculate them:
        - Returns: Percentage change in Close price (e.g., (Close[t] - Close[t-1]) / Close[t-1] * 100).
        - Volatility: Rolling standard deviation of Close over a 24-hour window.
        - High_Low_Range: Difference between High and Low prices.
        - RSI: Relative Strength Index using a 14-hour lookback period (implement using a function or library like ta).
    - Handle any missing values by filling with the mean or dropping rows if necessary.
    - Standardize the features using StandardScaler from sklearn.preprocessing to ensure equal weighting in clustering.
3. **K-Means Clustering**:
    - Implement K-Means clustering using sklearn.cluster.KMeans with an initial guess of 4 clusters (adjustable based on further analysis).
    - Determine the optimal number of clusters using the elbow method (plot within-cluster sum of squares vs. number of clusters, e.g., 2 to 10) and silhouette score analysis.
    - Fit the K-Means model to the standardized data and assign cluster labels to each data point.
4. **Visualization and Analysis**:
    - Create a scatter plot of the first two principal components (using sklearn.decomposition.PCA) colored by cluster labels to visualize cluster separation.

- Generate a correlation heatmap for the selected features across all clusters using seaborn.
- Produce individual correlation heatmaps for each cluster to highlight intra-cluster relationships.
- Compute and display a cluster profile summary (mean, median, min, max) for each feature across clusters using groupby and agg.

5. **Insights and Reporting**:
   - Interpret the cluster profiles to identify distinct market regimes or patterns (e.g., high volatility, bullish trends).
   - Analyze correlations to uncover relationships between features (e.g., TickVolume and High_Low_Range).
   - Save the dataset with cluster labels to a new CSV file (e.g., "XAUUSD_Hourly_Clustered.csv").
   - Provide a concise report summarizing key findings, including potential trading or business implications.
6. **Requirements**:
   - Use libraries: pandas, numpy, matplotlib, seaborn, sklearn.
   - Ensure the code is well-commented and modular for readability.
   - Handle exceptions (e.g., missing data, invalid cluster counts) gracefully.
   - Output visualizations should be saved as PNG files (e.g., cluster_scatter.png, correlation_heatmap.png).

Group Members' Insight

All members of the group have successfully conducted an in-depth Exploratory Data Analysis (EDA) and implemented unsupervised learning using K-Means clustering. Each individual has contributed unique insights and perspectives, enriching our understanding of the dataset through detailed examinations and clustering outcomes.

Here are the insights of each group member:

1. **MD Touhidul Islam Kanon:**

   From my analysis, K-Means clustering is crucial for uncovering hidden patterns in the XAUUSD dataset, especially without labeled data. Its importance lies in its ability to group similar market behaviors (e.g., bullish or bearish regimes) based on features like Close, Returns, and Volatility, enabling unsupervised segmentation. K-Means helps by reducing the complexity of the 298-hour dataset into four interpretable clusters, facilitating trend identification and trading strategy development. The results were significantly influenced by the cleaned data, as the removal of Spread inconsistencies and

standardization enhanced cluster clarity, particularly in Cluster 3's high-volatility, event-driven profile. This underscores the need for robust preprocessing to ensure reliable outcomes.

2. **Salma Kamilah Rahma:**

   Throughout this project, I discovered how unsupervised learning particularly K-Means clustering can simplify complex market data into meaningful segments that reflect distinct trading environments. What I found especially compelling was how combining price-based metrics with technical indicators like RSI and Volatility not only helped group similar behaviors but also revealed the psychological states of the market ranging from fear-driven selloffs to confidence-fueled rallies. Cluster analysis allowed me to see the market not as a linear sequence of prices, but as a dynamic system with recurring regimes. This reinforced the value of feature engineering and data normalization in ensuring that the model captures actual market signals rather than noise. Ultimately, this experience sharpened my ability to interpret financial data in a structured yet insightful way, bridging the gap between data science and real-world market interpretation.

3. **Resti Fresard:**

   Through this project, I gained a deeper understanding of the importance of data exploration and the application of unsupervised learning methods such as K-Means clustering in categorizing complex market behaviors. One aspect I found particularly interesting was how technical indicators like RSI and High_Low_Range can provide deeper insights into market conditions—not just in terms of price direction, but also in terms of volatility and trading activity. For instance, Cluster 3, which showed high volatility and low RSI, reflected a pressured market or sharp correction. From this, I learned that effective data modeling relies not only on algorithms but also on contextual understanding and proper feature selection. This project has greatly enriched my knowledge of financial market analysis using machine learning.

4. **Karina Ditya Amanda:**

   From reviewing the data, the High_Low_Range feature brought out an interesting dimension of intraday dynamics. While Cluster 1 had mid-level prices and high returns, its intraday range was narrower compared to Cluster 3, which had the widest spreads. This shows that not all price increases occur with volatility—some upward trends are more gradual and less erratic. Cluster 3's wide range and negative returns pointed to violent price swings typical of market corrections or news-driven movements. By comparing these ranges across clusters, it became evident that price direction alone doesn't explain risk—spread and volume adds deeper context to market behavior.

5. **Akmal Muhammad Firdaus:**

During the analysis process, I found that Cluster 0 represented a relatively bullish market condition, characterized by a higher average Close price and RSI near the overbought zone. This shows that the market experienced a strong upward trend, possibly due to external influences. What stood out to me was how RSI and Close price moved together in this cluster, indicating a momentum-driven environment. This helped me understand how clustering can reveal not just patterns, but also assist in anticipating market behavior. It made me realize the importance of technical indicators like RSI in identifying potential trading zones when grouped in similar market phases.