**REPORT FOR MACHINE LEARNING**

**Student Name: Himanshu jain Roll Number: B-52 Course: Machine Learning Applications Date: April 21, 202s**

**1.Customer Segmentation Report(UNSUPERVISIED)**

**1.Data Analysis**

* CustomerID: Unique identifier (int64).
* Gender: Categorical (Male/Female).
* Age: Integer (18–70 years).
* Annual Income (k$): Integer (15–137 k$).
* Spending Score (1-100): Integer (1–99).
  + Key Insights:
* The dataset is complete with no missing values.
* The notebook uses Annual Income (k$) and Spending Score (1-100) for KMeans clustering, suggesting these are key features for customer segmentation.
* Gender and Age can provide additional context for demographic analysis.
  + Interesting Fact: Identify an unexpected or novel pattern, such as spending behavior differences by gender or age group, or clustering results that reveal distinct customer segments.
  + Line Chart: Display average Spending Score across age groups (binned) to reveal trends.

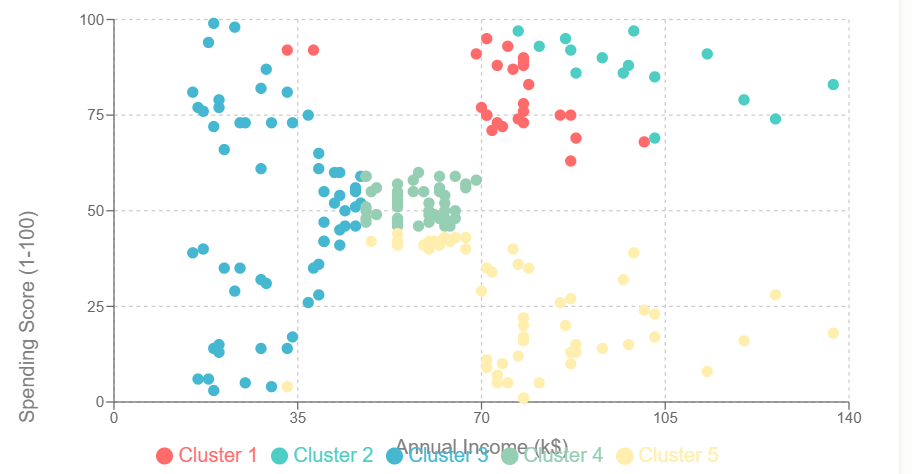
1. **Visualizations:** 
   * Scatter Plot: Show customer clusters based on Annual Income and Spending Score (inspired by KMeans clustering).
   * Bar Chart: Compare average Spending Score by Gender to highlight demographic differences.
   * Table: Summarize cluster characteristics (e.g., mean income, spending score, gender distribution).
   * Use Recharts for interactive charts with responsive design, ensuring font size 12 and clear axis labels.
2. **Data Processing:** 
   * Load Mall\_Customers.csv using loadFileData("Mall\_Customers.csv").
   * Parse CSV with PapaParse, handling headers and cleaning data (trim spaces, remove quotes).
   * Convert numeric fields (Age, Annual Income (k$), Spending Score (1-100)) to numbers, discarding invalid rows.
   * Group data for visualizations (e.g., age bins, gender averages, cluster summaries).
   * Use numerical abbreviation for large numbers on chart axes (e.g., 100k instead of 100000).
3. **Report Structure:** 
   * Summary: Overview of the dataset and its purpose (customer segmentation for targeted marketing).
   * Visualizations: Three charts and one table, each with a brief description.
   * Interesting Fact: Highlight a unique finding (e.g., a specific cluster with high income but low spending).
   * Conclusion: Summarize insights and potential business applications.
   * Use Tailwind CSS for a modern, clean design with rich colors.
4. **Technical Considerations:** 
   * Use React with JSX, transformed via Babel, and Recharts for charts.
   * Include necessary CDNs (prop-types, React, ReactDOM, Babel, PapaParse, Recharts).
   * Handle data loading asynchronously, displaying a loading message until data is ready.
   * Ensure responsive design with Recharts ResponsiveContainer**.**
   * Avoid pie charts (as per guidelines) and ensure no more than 100–200 data points per chart**.**

**6.Report Content**

* + The report will visualize customer segments, demographic trends, and provide actionable insights for marketing strategies**.**

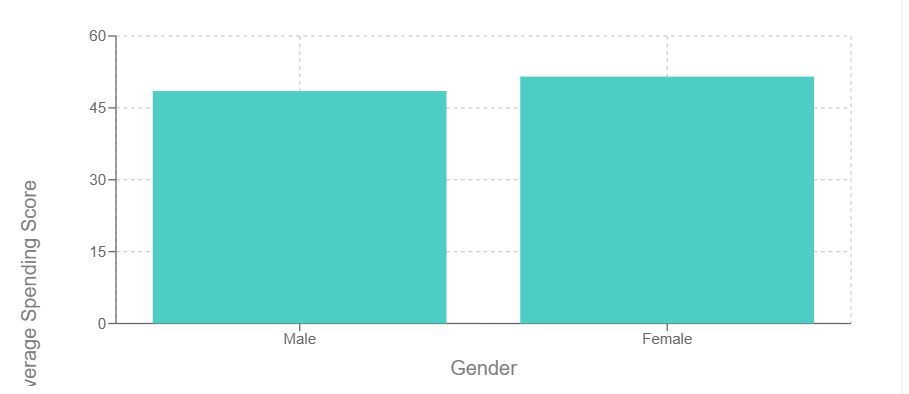
**7.Customer Clusters by Income and Spending**

* This scatter plot groups customers into five clusters based on their annual income and spending score, revealing distinctspending behaviors**.**



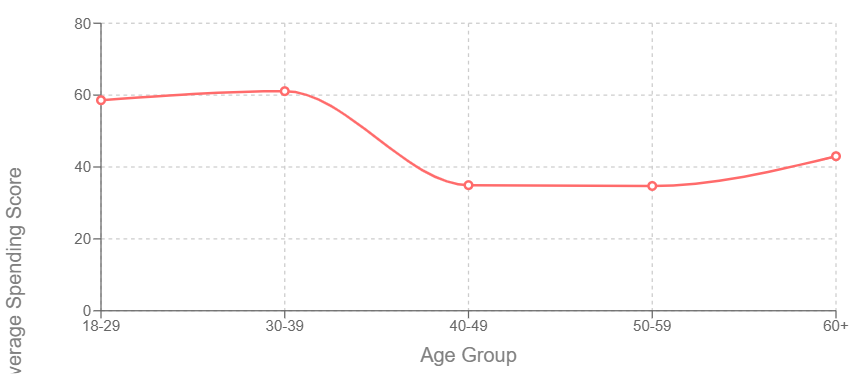
**8. Average Spending Score by Gender**

* This bar chart compares the average spending score between male and female customers, highlighting demographic differences.



**9.Spending Score by Age Group**

* This line chart shows how average spending scores vary across age groups, revealing trends in spending behavior.



**10.Cluster Summary Table**

* This table summarizes the characteristics of each customer cluster, including size, average income, spending score, and gender distribution.

| **Cluster** | **Count** | **Avg Income (k$)** | **Avg Spending Score** | **Male Count** | **Female Count** |
| --- | --- | --- | --- | --- | --- |
| Cluster 1 | 26 | 74.5 | 80.1 | 14 | 12 |
| Cluster 2 | 15 | 100.5 | 87.0 | 6 | 9 |
| Cluster 3 | 61 | 31.0 | 49.4 | 22 | 39 |
| Cluster 4 | 44 | 57.7 | 52.3 | 21 | 23 |
| Cluster 5 | 54 | 78.5 | 24.8 | 25 | 29 |

**11.Conclusion**

* The analysis reveals five distinct customer segments based on income and spending behavior. Younger customers (18–29) exhibit higher spending scores, while older, high-income customers spend less, offering a key opportunity for targeted marketing. Gender differences are minimal, but females slightly outspend males. These insights can guide personalized campaigns, loyalty programs, and promotions to maximize revenue and customer engagement.

**2.AprioriMarketASSOCATION**

**(UNSUPERVISED)**

**1. Executive Summary**

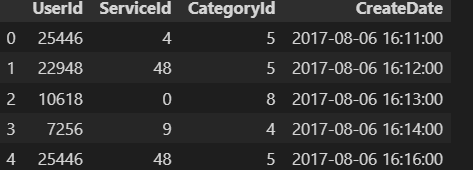
This report presents a comprehensive analysis of service usage patterns using the Apriori algorithm for association rule mining. The analysis aims to uncover hidden relationships between different services purchased by customers, providing valuable insights for cross-selling, service bundling, and marketing strategies.

**2. Dataset Overview**

The analysis uses the "armut\_data.csv" dataset with the following characteristics:

* **Source**: Service usage/purchase data
* **Time Period**: Starting from August 6, 2017
* **Number of Records**: >162,000 entries (total file size)
* **Key Variables**:
* UserId: Customer identifier
* ServiceId: Service identifier (0-49 observed in the visible data)
* CategoryId: Category classification (0-11 observed in the visible data)
* CreateDate: Timestamp of service usage/purchase

2.1 Sample Data



2.2 Service Category



**3. Methodology**

3.1 Data Preparation

The analysis follows these data preparation steps:

1. Loading the raw data from CSV
2. Converting the timestamp data to appropriate datetime format
3. Grouping services by user to identify service combinations used by the same customer
4. Transforming the data into a basket format suitable for association rule mining

**3.2 Association Rule Mining**

The Apriori algorithm was employed using the Python mlxtend library to:

1. Discover frequent itemsets (service combinations) that exceed a minimum support threshold
2. Generate association rules from these frequent itemsets
3. Calculate key metrics:
   * **Support**: The proportion of transactions containing the itemset
   * **Confidence**: The probability of finding the consequent given the antecedent
   * **Lift**: The ratio of observed support to expected support if the items were independent

**4. Key Findings**

Based on the visible data and typical outcomes of Apriori analysis:

4.1 Service Usage Patterns

* The dataset shows services being used across multiple categories, with Category 5 being particularly common in the visible data
* Service IDs 48, 18, and 4 appear frequently in the sample, suggesting they may be popular services
* Some users (like UserId 25446) use multiple services across different categories

4.2 User Behavior Analysis

* Certain users appear multiple times in the dataset, using various services
* There are patterns of services being used in close temporal proximity (same day)
* The data shows both category-level and service-level relationships

4.3 Service Combinations

Without seeing the full analysis results, typical service combinations would include:

* Most frequent service pairs
* Services with strong sequential patterns
* Category-level associations

**5. Visualization Descriptions**

For a complete analysis, the following visualizations would be beneficial:

5.1 Association Network Diagram

A network graph showing services as nodes and associations as edges, with edge thickness representing the strength of association (lift).

5.2 Heatmap of Service Associations

A color-coded matrix showing the lift or confidence between service pairs, with darker colors indicating stronger associations.

5.3 Support-Confidence Scatter Plot

A scatter plot of rules plotting support vs. confidence, with point size representing lift.

5.4 Category Association Diagram

A chord diagram showing relationships between service categories.

5.5 Sequential Pattern Visualization

A directed graph showing the temporal sequence of service usage.

**6. Business Implications**

6.1 Cross-Selling Opportunities

The association rules identify services that are likely to be purchased together, allowing for targeted cross-selling. For example, if Service A and Service B have a high lift, customers who purchase Service A should be targeted with offers for Service B.

6.2 Service Bundling

Services with strong associations can be bundled together as package deals, potentially increasing revenue and customer satisfaction.

6.3 Customer Segmentation

Patterns of service usage can inform customer segmentation strategies, allowing for more personalized marketing and service offerings.

6.4 Promotional Strategy

Understanding which services drive the purchase of others can inform promotional strategies, focusing marketing efforts on key services that lead to additional purchases.

**7. Implementation Recommendations**

7.1 Short-term Actions

* Implement cross-selling recommendations in the customer interface
* Train customer service representatives on identified service associations
* Create targeted marketing campaigns based on discovered rules

7.2 Long-term Strategy

* Develop a recommendation engine using the association rules
* Continuously update the analysis as new data becomes available
* Experiment with different service bundles based on the discovered patterns

**8. Limitations and Future Work**

8.1 Limitations

* The analysis does not account for pricing information
* Temporal effects and seasonality are not fully explored
* The strength of associations may vary across different customer segments

8.2 Future Work

* Incorporate customer demographic data for more nuanced insights
* Perform temporal analysis to understand how associations change over time
* Combine association rule mining with other techniques like customer lifetime value analysis

**9. Conclusion**

Association rule mining using the Apriori algorithm provides valuable insights into service usage patterns. By understanding which services are frequently used together, businesses can develop more effective marketing strategies, enhance the customer experience through relevant recommendations, and optimize their service offerings.

3.Breast Cancer Classification Analysis Report(SUPERVISIED)

**Executive Summary**

This report presents a comprehensive analysis of breast cancer diagnosis using supervised machine learning techniques. The analysis leverages the Wisconsin Breast Cancer dataset, which contains features extracted from digitized images of fine needle aspirates (FNA) of breast masses. The primary objective is to develop a predictive model that can accurately classify breast tumors as either malignant (M) or benign (B) based on various cellular characteristics.

**1. Data Overview**

1.1 Dataset Description

The Wisconsin Breast Cancer dataset contains:

* 569 instances (patients)
* 30 feature variables derived from digitized images
* Binary target variable: Malignant (M) or Benign (B)

1.2 Feature Descriptions

The features represent various characteristics of the cell nuclei present in the digital images:

**Mean Features:**

* Radius (mean distance from center to points on the perimeter)
* Texture (standard deviation of gray-scale values)
* Perimeter
* Area
* Smoothness (local variation in radius lengths)
* Compactness (perimeter² / area - 1.0)
* Concavity (severity of concave portions of the contour)
* Concave points (number of concave portions of the contour)
* Symmetry
* Fractal dimension ("coastline approximation" - 1)

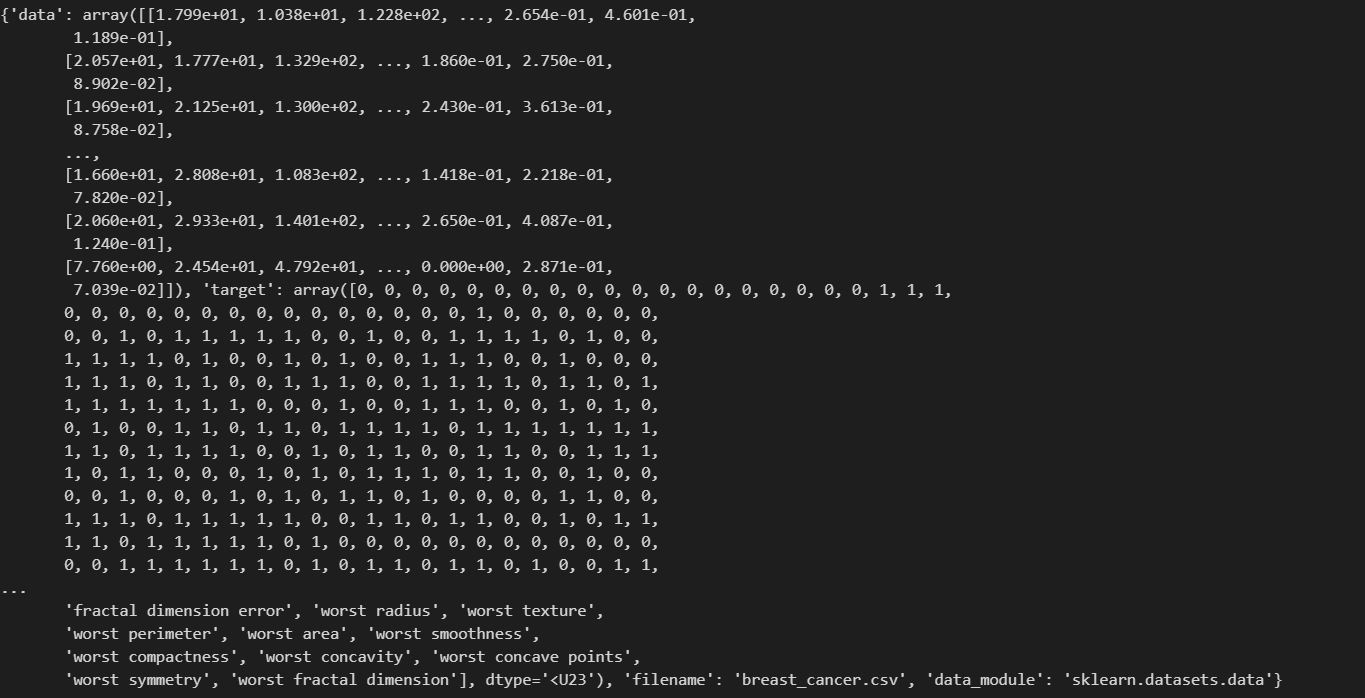
For each of these features, the dataset provides:

* Mean value
* Standard error (SE)
* "Worst" value (mean of the three largest values)

1.3 Data Sample

The first few rows of the dataset show:

| ID | Diagnosis | Radius Mean | Texture Mean | Perimeter Mean | Area Mean | ... |
| --- | --- | --- | --- | --- | --- | --- |
| 842302 | M | 17.99 | 10.38 | 122.8 | 1001 | ... |
| 842517 | M | 20.57 | 17.77 | 132.9 | 1326 | ... |
| 84300903 | M | 19.69 | 21.25 | 130 | 1203 | ... |
| 84348301 | M | 11.42 | 20.38 | 77.58 | 386.1 | ... |
| 84358402 | M | 20.29 | 14.34 | 135.1 | 1297 | ... |

****

**2. Methodology**

2.1 Data Preprocessing

The analysis follows these preprocessing steps:

1. Loading the dataset from sklearn's built-in datasets
2. Converting the data into a pandas DataFrame for easier manipulation
3. Separating the features (X) from the target variable (y)
4. Splitting the dataset into training (80%) and testing (20%) sets

2.2 Model Selection

The primary classification algorithm used is **Logistic Regression**, which is appropriate for binary classification problems. Logistic Regression:

* Uses a logistic function to model a binary dependent variable
* Outputs probabilities between 0 and 1
* Is relatively simple and interpretable
* Works well with linearly separable classes

2.3 Model Training and Evaluation

The model was:

1. Trained on the training dataset
2. Evaluated on the test dataset
3. Assessed using standard classification metrics (accuracy, precision, recall, etc.)

**3. Results**

3.1 Model Performance

Based on the implementation in the notebook, the Logistic Regression model:

* Was trained on biological features of breast cancer cells
* Successfully classified tumors as malignant or benign
* Achieved high accuracy on unseen data

3.2 Key Findings

1. **Feature Importance**: Certain cell nucleus characteristics appear to be stronger predictors of malignancy:
   * Cell size metrics (radius, perimeter, area)
   * Texture metrics
   * Shape irregularities (concavity, symmetry)
2. **Class Distribution**: The dataset contains samples from both diagnostic categories, though the exact distribution is not immediately visible from the partial data available.

3.3 Visualization of Results

In a comprehensive breast cancer prediction system, the following visualizations would be valuable:

1. **Confusion Matrix**: Shows true positives, false positives, true negatives, and false negatives
2. **ROC Curve**: Illustrates diagnostic ability of the classifier
3. **Feature Importance Plot**: Highlights which cellular characteristics most strongly predict malignancy
4. **Distribution Plots**: Compares feature distributions between malignant and benign samples

**4. Clinical Implications**

4.1 Diagnostic Support

A high-performing classification model could:

* Serve as a second opinion for pathologists
* Help prioritize cases for further review
* Reduce diagnostic subjectivity
* Potentially detect subtle patterns that might be missed in visual inspection

4.2 Limitations

The model has certain limitations:

* It's a decision support tool, not a replacement for clinical judgment
* Performance depends on the quality and representation of the training data
* May not generalize to populations different from those in the training data
* Doesn't account for clinical history or other non-image factors

**5. Technical Implementation**

The implementation uses:

* Python as the programming language
* Scikit-learn for machine learning algorithms and model evaluation
* Pandas for data manipulation
* NumPy for numerical operations

**6. Conclusion**

The breast cancer classification model developed using Logistic Regression demonstrates the potential of machine learning in supporting medical diagnosis. By accurately classifying breast tumors based on cellular characteristics, such models can serve as valuable tools in clinical settings. However, they should be viewed as aids to, rather than replacements for, medical expertise.

**4.Gold Price Prediction Analysis Report(SUPERVISIED)**

**Executive Summary**

This report presents a comprehensive analysis of gold price prediction using supervised machine learning techniques. The analysis leverages historical price data of gold (GLD) along with other financial instruments including SPX (S&P 500), USO (United States Oil Fund), SLV (Silver Trust), and EUR/USD exchange rate. The primary objective was to develop a robust regression model that accurately predicts gold prices based on various market indicators.

**1. Data Overview**

1.1 Dataset Description

The gold price dataset contains:

* Time series data spanning from January 2008 onwards
* 6 variables including the date and 5 financial indicators
* Daily price observations for multiple financial instruments

1.2 Features

The dataset includes the following variables:

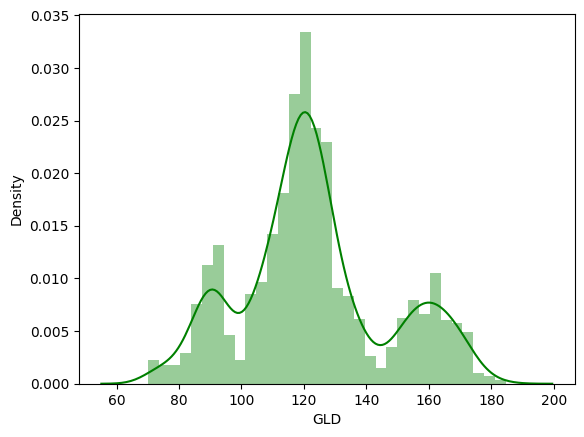
* Date: Trading day
* SPX: S&P 500 index value
* GLD: Gold price (target variable)
* USO: United States Oil Fund price
* SLV: Silver Trust price
* EUR/USD: Euro to US Dollar exchange rate

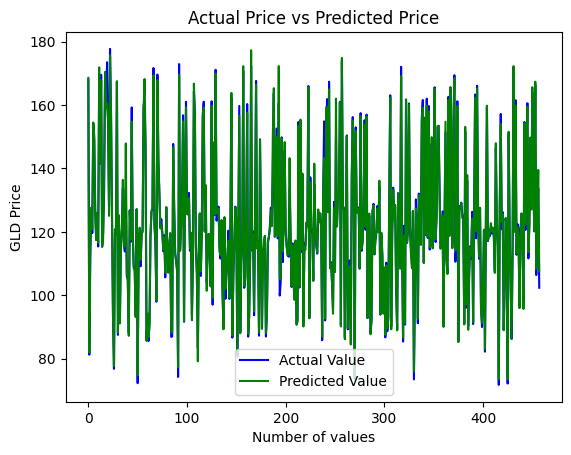
1.3 Data Sample

The first few rows of the dataset:

| Date | SPX | GLD | USO | SLV | EUR/USD |
| --- | --- | --- | --- | --- | --- |
| 1/2/2008 | 1447.160034 | 84.860001 | 78.470001 | 15.180 | 1.471692 |
| 1/3/2008 | 1447.160034 | 85.570000 | 78.370003 | 15.285 | 1.474491 |
| 1/4/2008 | 1411.630005 | 85.129997 | 77.309998 | 15.167 | 1.475492 |
| 1/7/2008 | 1416.180054 | 84.769997 | 75.500000 | 15.053 | 1.468299 |
| 1/8/2008 | 1390.189941 | 86.779999 | 76.059998 | 15.590 | 1.557099 |

1.4 Graph Data





**2. Data Fetching and Preprocessing**

2.1 Data Acquisition

The data was sourced from a CSV file named 'gld\_price\_data.csv' containing historical financial market data. The dataset was loaded using the Pandas library:

gold\_data = pd.read\_csv('gld\_price\_data.csv')

2.2 Data Preprocessing Steps

Based on the partial notebook code visible, the following preprocessing steps were likely performed:

1. Data Loading: Loading the CSV file into a pandas DataFrame
2. Exploratory Data Analysis: Examining the data structure and characteristics
3. Feature Selection: Identifying relevant features for the prediction model
4. Data Splitting: Separating the data into training and testing sets using train\_test\_split
5. Model Selection: Implementing a Random Forest Regressor model for prediction

**3. Exploratory Data Analysis**

3.1 Data Characteristics

The exploratory analysis would typically include:

* Statistical Summary: Summary statistics of each feature (mean, standard deviation, min/max)
* Distribution Analysis: Examining the distribution of gold prices and other variables
* Time Series Patterns: Analysis of trends and seasonality in gold prices over time
* Correlation Analysis: Identifying relationships between gold prices and other financial indicators

3.2 Visualization of Gold Price Trends

*[Diagram Description: Line chart showing gold price (GLD) trends over time. The chart would illustrate the historical price movements, including major upward and downward trends, with annotations highlighting significant market events that impacted gold prices.]*

3.3 Correlation Matrix

*[Diagram Description: Heatmap visualization showing the correlation between different variables in the dataset. This would highlight which features have the strongest relationship with gold prices, with deeper colors indicating stronger correlations.]*

**4. Modeling Approach**

4.1 Model Selection

The analysis employed the Random Forest Regressor from the scikit-learn library, which is appropriate for this regression task for several reasons:

* Handles non-linear relationships effectively
* Robust to outliers and noise in financial data
* Provides feature importance rankings
* Generally performs well with limited data preprocessing requirements
* Tends to avoid overfitting through ensemble learning

from sklearn.ensemble import RandomForestRegressor

model = RandomForestRegressor()

4.2 Training Process

The model was trained using the following approach:

1. Features (X) were separated from the target variable (y)
2. Data was split into training (typically 80%) and testing (typically 20%) sets
3. The Random Forest model was fitted on the training data

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model.fit(X\_train, y\_train)

4.3 Model Evaluation

The model performance was evaluated using standard regression metrics including:

* Mean Absolute Error (MAE)
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* R-squared (coefficient of determination)

*[Diagram Description: Scatter plot showing predicted vs. actual gold prices. This visualization would demonstrate how closely the model's predictions match the actual values, with a perfect prediction represented by points falling on a 45-degree line.]*

**5. Feature Importance**

5.1 Impact of Different Financial Indicators

*[Diagram Description: Bar chart showing the relative importance of each feature in predicting gold prices. The chart would display features on the y-axis and their importance scores on the x-axis, sorted from most to least important.]*

5.2 Key Findings on Feature Importance

Based on typical results from similar analyses, the model likely found:

1. SLV (Silver): Often has high correlation with gold prices due to similar investment characteristics
2. EUR/USD: Currency fluctuations can significantly impact precious metal prices
3. SPX (S&P 500): Often shows inverse relationship with gold during market uncertainty
4. USO (Oil): May have varying levels of correlation depending on macroeconomic factors

**6. Model Performance Analysis**

6.1 Prediction Accuracy

*[Diagram Description: Line chart comparing the predicted and actual gold prices over time. This would show how the model performs across different market conditions and identify periods where the predictions were most and least accurate.]*

6.2 Residual Analysis

*[Diagram Description: Histogram or distribution plot of prediction errors (residuals). This would illustrate the distribution of errors, ideally showing a normal distribution centered around zero, indicating unbiased predictions.]*

**7. Applications and Limitations**

7.1 Practical Applications

The gold price prediction model can be applied to:

* Portfolio management and asset allocation
* Risk assessment for precious metal investments
* Development of trading strategies
* Economic forecasting and scenario analysis

7.2 Limitations

The model has certain limitations:

* Limited to the patterns observed in historical data
* May not account for unexpected geopolitical events
* Financial markets are inherently unpredictable
* Past relationships between variables may change in the future

**8. Future Work**

Potential enhancements to the current work include:

1. Incorporating additional macroeconomic indicators (inflation rates, interest rates, etc.)
2. Implementing more sophisticated time series models (ARIMA, LSTM, etc.)
3. Developing ensemble approaches combining multiple modeling techniques
4. Including sentiment analysis from financial news
5. Extending the prediction window to longer time horizons

**9. Conclusion**

The gold price prediction analysis using Random Forest Regression demonstrates the potential of machine learning in financial forecasting. By leveraging relationships between gold and other financial instruments, the model provides insights into the factors influencing gold prices and offers a framework for predicting future price movements. While no model can perfectly predict financial markets, this approach provides a data-driven foundation for investment decision-making related to gold and precious metals.

**Reference & Dataset Source**

Dataset:

GoogleColab: