FinalProject-CMPSC-445-Tariff-ML

Made by AnatolyBarabanov and Matthew Danese

Professor: Janghoon Yang

# Description of the Project

This project develops a machine learning-based Tariff Impact Dashboard that enables users to:  
  
- Forecast commodity prices  
- Cluster commodities based on their price behavior  
- Classify commodity risk levels

The system uses the World Bank's 'CMO-Historical-Data-Monthly.xlsx' dataset and integrates three trained models (Regression, Clustering, Classification) into an interactive Streamlit web application.

Live Website: [Streamlit App](<https://finalproject-cmpsc-445-tariff-ml-ka3ig4hwa93cmwpuusycyv.streamlit.app>)

# Significance of the Project

Understanding commodity market behavior is crucial for traders and businesses. Our dashboard provides:  
  
- Forecasting of future prices to support better economic decisions  
- Clustering of commodities for comparative analysis  
- Risk classification to highlight commodities with high volatility risks

The project integrates multiple ML techniques into one web platform, offering real-time insights into complex tariff-related economic trends.

# Project Goals

- Predict future commodity prices  
- Group commodities with similar behavior  
- Classify commodities into risk categories  
- Provide visual insights via a web dashboard

# Data Collection

Source: World Bank – Commodity Markets <https://www.worldbank.org/en/research/commodity-markets>  
File: CMO-Historical-Data-Monthly.xlsx  
Monthly commodity prices across agriculture, energy, and metals sectors  
Data Characteristics:  
 — Covers over 80 commodities  
 — Time range: 1980s–2025s  
 — Data points: Monthly price observations  
Metadata Fields:  
 — Commodity name  
 — Units  
 — Monthly price values

# Project Structure

FinalProject-CMPSC-445-Tariff-ML/

├── models/                           # Trained machine learning models (.pkl)

│ └──kmeans\_model.pkl

│ └── price\_forecast\_model\_commodities.pkl

│ └── kmeans\_scaler.pkl

│ └── risk\_classification\_commodities.pkl

├── data/                             # Dataset (World Bank CMO data)

│   └── CMO-Historical-Data-Monthly.xlsx

├── scripts/                          # All training scripts

│   ├── tariff\_models.py              # Shared helper functions

│   ├── train\_regression.py            # Trains price forecasting model

│   ├── train\_kmean.py                 # Trains commodity clustering model

│   └── train\_classifier.py            # Trains risk classification model

├── requirements.txt                  # Python libraries to install

└── README.md # Project description

└── tariff\_dashboard1.py

# Scripts Overview

1. tariff\_models.py — First to Run: Shared Model Functions  
 - Has helper functions for training and saving machine learning models.  
  
2. train\_regression.py — Price Forecasting Model Trainer  
 - Trains a Linear Regression model to predict future commodity prices.  
  
3. train\_kmean.py — Clustering Model Trainer  
 - Trains a KMeans clustering model to group commodities based on their price trends.  
  
4. train\_classifier.py — Risk Classification Model Trainer  
 - Trains a Decision Tree Classifier to predict commodity risk levels.  
  
5. tariff\_dashboard.py — Streamlit Web Application  
 - Combines all three trained models into an interactive Streamlit dashboard.

# Instructions for Web Usage

To access and use the web app:  
  
1. Open the website: [Streamlit App](<https://finalproject-cmpsc-445-tariff-ml-ka3ig4hwa93cmwpuusycyv.streamlit.app>)  
2. Choose a feature from the sidebar:  
- Price Forecasting: Select a commodity to predict future prices.  
- Commodity Clustering: View clustering results.  
- Risk Classification: Select a commodity and get risk predictions (High / Medium / Low).

# Functionalities and Test Results

Task | Model | Description  
Price Forecasting | Linear Regression, ARIMA | Predicts future commodity prices  
Industry Clustering | KMeans + PCA | Clusters commodities by trends  
Risk Classification | Decision Tree | Categorizes commodities into risk classes  
  
Test Results:  
- Models load correctly from /models/.  
- Forecasted prices align with known trends.  
- Clusters are distinct and logically grouped.  
- Risk classifier produces consistent predictions.

# Data Processing

- Dropped commodities with excessive missing values.  
- Selected the most recent 60 months (5 years) for consistency.  
- Scaled features (StandardScaler) for KMeans clustering.  
- Created lag-based features for forecasting.  
- Computed price volatility for risk categorization.  
- Applied PCA for cluster visualization.

# Model Development

Model | Inputs | Output  
Linear Regression | Last 5 months of prices | Next month price  
KMeans Clustering | Scaled monthly price trends | Cluster assignment  
Decision Tree Classifier | Price volatility | Risk level (High/Medium/Low)  
  
Algorithms Justification:  
- Linear Regression: Effective for short-term forecasting.  
- KMeans: Suited for unsupervised commodity behavior grouping.  
- Decision Tree: Efficient and interpretable for risk prediction.

# How to Run

1. Clone the repo:  
   `git clone [https://github.com/AnatolyBarabanov/FinalProject-CMPSC-445-Tariff-ML.git`](https://github.com/AnatolyBarabanov/FinalProject-CMPSC-445-Tariff-ML.git%60)
2. Navigate to the folder:  
   `cd FinalProject-CMPSC-445-Tariff-ML`
3. Install dependencies:  
   `pip install -r requirements.txt`
4. (Optional) Retrain models:  
   `python train\_regression.py`  
   `python train\_kmean.py`  
   `python train\_classifier.py`
5. Launch the dashboard:  
   `streamlit run tariff\_dashboard.py`

# Discussion and Conclusions

This was a very interesting project. Especially it was very fun as a group. We were able to work together to make this project exactly the way we wanted it. We quickly figured out what we would do and found a data source. Working together allowed to add more to the project.

Conclusions:

* Price forecasting works well for most stable goods.
* Clustering identifies logical groups such as precious metals, agriculture, and energy.
* Risk classification effectively identifies commodities with high volatility.

Project problems:

* Initially there were problems with regression due to improper cleaning, but after better cleaning, everything went well.
* Creating the site was also difficult because there is truly little experience working with python sites. For a long time, Price Forecasting simply did not load on the site, we were able to solve the problem by adding some edits to the tariff\_dashboard1.py code.
* And unstable commodities occasionally caused forecasting errors.