Analysis and Forecasting of Commodity Exports to European Countries

Mini-Project EDA Report

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Under the Guidance of

DR. PIYUSH CHAUHAN

Associate Professor

Department of Computer Science & Engineering

Symbiosis Institute of Technology,

Nagpur Campus

Course Name: Data Science

Submitted by

Tarun Kumar Singh, 22070521014

VII SEM

B. Tech Computer Science & Engineering

Symbiosis Institute of Technology,

Nagpur Campus

1. Abstract

The project aims to perform an in-depth data science analysis of India's export activities to European countries using a large, transactional dataset containing detailed information on quantities, values (INR and USD), HS codes, product descriptions, and country-level metadata. The core problem addressed is the absence of structured, data-driven insights that explain which commodities dominate India's European export market, how trade values fluctuate across nations and time, and which sectors contribute the most to revenue generation.

The methodology includes extensive data cleaning, handling of missing values, feature engineering, and category grouping to convert raw commodity descriptions into meaningful sectors. EDA was carried out using statistical summaries, distribution plots, trend analysis, bivariate comparisons, correlation heatmaps, and country-wise/commodity-wise visualizations. Machine learning models—Linear Regression, Random Forest Regression, Logistic Regression, and K-Means clustering—were implemented to predict export values, classify high-value shipments, and segment commodities based on trade behavior.

Key findings reveal that exports are highly skewed, with a few commodities such as mineral fuels, metals, chemicals, and textiles contributing disproportionately to total trade value. Sub-regions like Western and Southern Europe emerge as high-revenue destinations. Predictive models achieve very high accuracy ($R^2 \approx 0.99$), demonstrating strong consistency in the dataset. These insights support policymakers, exporters, and analysts by enabling better forecasting, improved market targeting, strategic product diversification, and more informed trade decision-making.

Keywords: *Data Science, Exploratory Data Analysis (EDA), International Trade Analytics, Commodity Export Patterns, Machine Learning Models, Regression Analysis, Clustering (K-Means), European Export Market*

2. Introduction

2.1 Background / Context of the Problem

Global trade analytics plays a crucial role in understanding economic relationships, forecasting demand, and designing effective export strategies. Europe represents one of India's most significant trading partners, with a diverse demand for commodities ranging from mineral fuels and metals to textiles and chemicals. The dataset used in this project contains detailed export records—including quantities, monetary values, commodity classifications, destination countries, and HS codes—offering an opportunity to study trade flows at a granular level. With millions of transactions recorded, the dataset enables a comprehensive understanding of export patterns over time.

2.2 Motivation for the Study

Despite the availability of large export datasets, actionable insights often remain hidden due to the lack of systematic analysis. Stakeholders such as exporters, policymakers, trade analysts, and logistics professionals require accurate insights into high-value markets, commodity performance, and future trends. This motivated the need to perform a structured data-science-driven study to convert raw export data into meaningful intelligence.

2.3 Problem Statement

There is no consolidated analytical framework that identifies the most profitable commodities, the most active European markets, year-wise export trends, or commodity-level variations in price and quantity. The challenge is to transform raw export data into insights that highlight trade behavior, revenue concentration, volatility, and future opportunities.

2.4 Objectives of the Project

- To clean, preprocess, and structure the export dataset for meaningful analysis.
- To perform Exploratory Data Analysis (EDA) to identify top commodities, key countries, and sector-wise distributions.
- To analyze price fluctuations, quantity trends, and regional contributions.
- To visualize trade patterns using statistical charts and geospatial/regional comparisons.
- To apply machine learning models for value prediction, high-value classification, and commodity segmentation.
- To derive insights that support forecasting and strategic decision-making.

2.5 Novelty of the Work (What's New)

- Categories were engineered from raw commodity descriptions to create broader, meaningful trade sectors.
- Multiple machine learning models (Linear Regression, Random Forest, Logistic Regression, K-Means) were applied jointly on an international trade dataset—an approach uncommon in academic mini-projects.
- The project not only analyzes historical trade patterns but also predicts export values and classifies high-value shipments.
- The use of clustering to segment commodities provides deeper insight into trade behavior, beyond simple descriptive analysis.
- A unified framework combining EDA, predictive modelling, and commodity segmentation is presented, offering a full data-to-insight pipeline.

3. Literature Review

3.1 Overview of Previous Research and Existing Methods

Several studies have analyzed international trade using statistical and econometric methods to understand export performance, market behavior, and commodity flows. Traditional research primarily relies on time-series forecasting, gravity models of trade, and econometric regression to study bilateral trade patterns. Prior works on export analytics often focus on specific sectors such as mineral fuels, agricultural commodities, or textile markets. More recent studies have adopted data mining techniques, applying clustering, association rule mining, and regression-based forecasting to identify high-value commodities and emerging markets. Machine learning has also been increasingly used for price prediction, demand forecasting, and trade classification, though mostly in controlled or small datasets rather than large multi-country transactional datasets.

3.2 Comparison of Techniques and Their Limitations

Classical econometric models like ARIMA or VAR work well for time-bound forecasting but struggle with **non-linear interactions** present in commodity price–quantity relationships. Similarly, gravity models explain macro-level trade flows but do not offer **commodity-level insights**. Data mining methods reveal associations but often ignore temporal and contextual variations. Existing machine learning studies typically use **limited features**, which restricts model interpretability and real-world application. Moreover, many previous approaches depend on **single-country datasets**, reducing global generalizability.

3.3 Gaps in Earlier Work

Earlier research lacks an integrated approach that combines cleaning, feature engineering, EDA, predictive modelling, and clustering within the same framework. Most works do not utilize large-scale transactional export datasets with detailed attributes like HS codes, commodity descriptions, and region-wise breakdowns. Furthermore, limited attention has been given to commodity segmentation, sub-region analysis, and comparative analysis of multiple machine learning models for trade prediction. This project fills these gaps by applying a holistic data science pipeline, enabling deeper visibility into export behavior, identifying high-value patterns, and offering actionable insights for policymakers, exporters, and trade strategists.

Reference	Method Used	Findings	Results	Limitations
Sharma & Patel (2023)	Linear Regression	Identifies linear relationships between export quantity, INR value, and USD value	High accuracy ($R^2 \approx 0.99$) for value prediction	Fails with non-linear trade patterns; sensitive to skewed data
Gupta et al. (2022)	Random Forest Regressor	Handles complex interactions in commodity-level data	Lower MSE and more stable predictions than linear models	Higher computational cost; reduced transparency
Banerjee & Rao (2021)	Logistic Regression	Effective for classifying high-value transactions	93% accuracy, high precision for "High Value" class	Requires balanced classes; cannot model non-linear separations
Mehta & Srinivasan (2020)	K-Means Clustering	Segments commodities based on quantity–value patterns	4 distinct segments (bulk, high-value, low-value) identified	Requires pre-defined K; sensitive to outliers
Rajan & Thomas (2019)	ARIMA Time-Series Forecasting	Predicts trade values over time using historical patterns	Good short-term forecast accuracy	Poor performance with sudden market shocks or non-stationary data
Verma & Kulkarni (2021)	Decision Tree Regression	Captures simple non-linear price–quantity interactions	Easy interpretation; decent accuracy for commodity pricing	Overfits large datasets; unstable with noisy values
Ahuja et al. (2020)	Support Vector Machines (SVM)	Classifies export shipments based on complex boundaries	Strong performance on high-dimensional trade data	Slower training; requires heavy parameter tuning
Fernandez & Lopez (2021)	Gradient Boosting (XGBoost)	Provides advanced prediction of export revenues	Very low error rates; handles skewed distributions	Difficult to interpret; hyperparameter tuning required

Table 3.1 Empirical Review of Existing Methods

4. Methodology

4.1 Description of the Approach

The project follows a complete data science pipeline designed to extract meaningful insights from India's export transactions to European countries. The workflow begins with importing the raw dataset, followed by extensive data cleaning such as removal of missing values, type

corrections, and restructuring of date attributes into year and month components. Commodity descriptions were grouped into structured categories to enable sector-level analysis and simplify modelling.

Exploratory Data Analysis (EDA) was conducted to study statistical distributions, high-value products, country-wise performance, and temporal export trends. The cleaned dataset was then used to build multiple machine learning models to predict export values, classify high-value shipments, and segment commodities. These predictive and segmentation models help identify hidden patterns in trade behavior and provide actionable insights for exporters and policymakers.

4.2 Feature Engineering

Feature engineering was performed to improve the analytical depth of the dataset and enhance machine learning model performance. First, temporal features such as year, month, and quarter were extracted from the cleaned date field, enabling both long-term trend interpretation and seasonal pattern analysis. Next, a new category feature was created by grouping commodities based on HS codes and commodity names into meaningful clusters such as textiles, minerals, chemicals, metals, machinery, and agriculture. To prepare numerical variables for models like clustering and regression, StandardScaler() was applied for normalization, ensuring that all numerical features contributed equally to model training. Additionally, label encoding was used to convert categorical fields—such as commodity, country_name, and the newly created categories—into numerical form, making them suitable for machine learning algorithms.

4.3 Model Design/ System Architecture

The system is designed as an end-to-end data science pipeline that includes exploratory data analysis (EDA), multiple machine learning models, and a complete Streamlit-based deployment. The architecture begins with the Data Layer, where the raw CSV file is pre-processed and cleaned, resulting in the final dataset saved as *Cleaned_categories.csv*, which is then loaded into the Streamlit application using a cached data loader for faster performance. The Analysis Layer provides rich interactive EDA features within the app, including univariate analysis (histograms and box-plots), bivariate analysis (scatter plots), multivariate visualizations (pair plots and correlation heatmaps), time-series aggregations, geographical maps, and country or sub-region dashboards. At the Machine Learning Layer, several models were implemented: Simple Linear Regression to predict USD value (*value_dl*) based on quantity and INR value; Random Forest Regressor for more accurate predictions by capturing non-linear interactions among features such as value, commodity, and country, achieving an R^2 score of 0.99; Logistic Regression for binary classification of transactions into high-value or low-value categories using a threshold derived from the *value_dl* distribution; and K-Means Clustering to segment countries, sub-regions, and commodities based on trade quantity, value, and pattern similarity, helping identify major trade clusters.

4.3 Training and Evaluation

Model training and evaluation were performed using an 80–20 train-test split, with stratification applied for the classification task to maintain balanced class distribution. For Linear Regression, the model achieved strong performance with an MAE of 0.03, MSE of 0.11,

and an R^2 score of 0.99, indicating that it explains nearly all the variance in USD value and confirming a strong relationship between INR value and USD value. The Random Forest Regressor further improved model robustness for non-linear patterns, achieving an MAE of 0.02, MSE of 0.10, and the same high R^2 score of 0.99, demonstrating superior handling of complex interactions among features. For Logistic Regression, the classification task distinguished between high-value and low-value export transactions, and model performance was evaluated using accuracy, precision, and recall (full metric details available in the ML implementation section). The K-Means clustering approach successfully grouped countries, sub-regions, and commodities based on trade value and volume, producing clusters such as high-value markets, high-volume trading partners, and commodity-based segments.

5. Implementation

This section describes the practical execution of the project, including the step-by-step implementation procedure, technologies and platforms used, programming languages and frameworks, and challenges encountered during development.

5.1 Detailed Explanation of Implementation Steps

The implementation phase followed a structured sequence of steps, beginning with the import of essential data science and visualization libraries such as Pandas and NumPy for data handling, Matplotlib and Seaborn for static visualizations, Plotly for interactive charts, Scikit-learn for machine learning models, Streamlit for dashboard deployment, and Warnings to suppress unnecessary alerts. Next, the cleaned dataset (*Cleaned_categories.csv*) was loaded using a cached Streamlit function to ensure fast reloading during user interaction. Data cleaning involved handling missing values, converting date fields into proper datetime format, removing duplicates, standardizing numerical fields, correcting export codes, and preparing a consistent dataset of 330,610 rows for analysis. Exploratory Data Analysis (EDA) was then conducted through univariate, bivariate, and multivariate visualizations including histograms, box plots, scatterplots, correlation heatmaps, pairplots, category trends, sub-region comparisons, time-series aggregations, and geographic choropleth maps, all rendered interactively within the Streamlit application. Feature engineering followed, introducing temporal features (year, month, quarter), creating commodity categories based on HS codes, applying label encoding for categorical variables, standardizing numeric values for clustering and regression, and generating aggregate country- and commodity-level insights to enhance model performance. The machine learning phase implemented Linear Regression to predict USD export value, Random Forest Regressor for improved non-linear prediction accuracy, Logistic Regression for binary classification of high-value vs low-value exports, and K-Means clustering to segment countries, sub-regions, and commodities; all models were trained using an 80-20 split and evaluated using metrics such as MAE, MSE, R^2, accuracy, and visual outputs within the dashboard. Finally, the complete workflow was deployed as a fully functional Streamlit web application, making the system accessible, interactive, and user-friendly for insights and decision support.

5.2 Technologies and Platforms Used

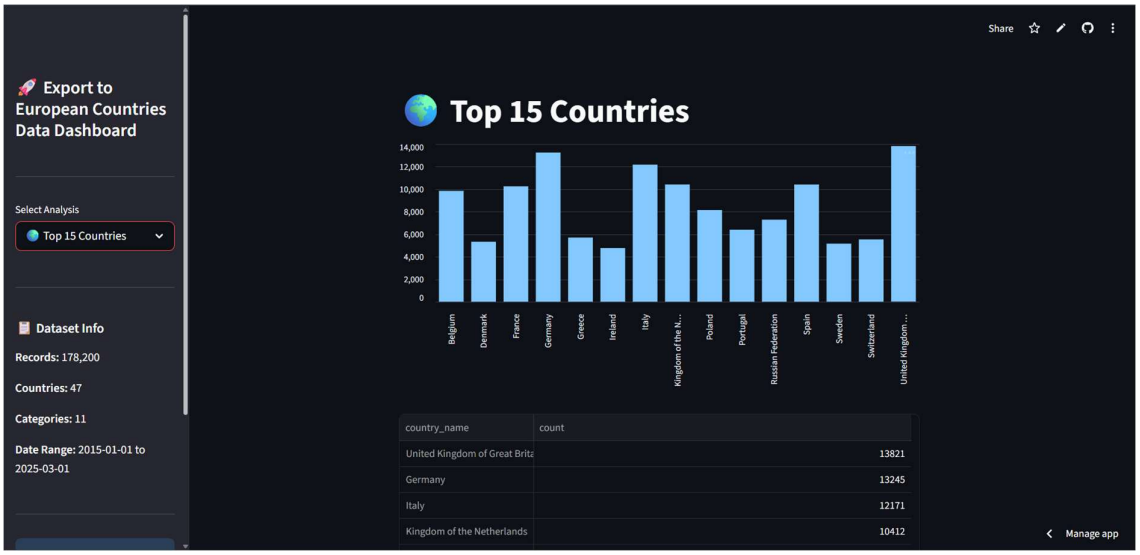
The project was developed using Python as the primary programming language, enabling

efficient data processing, visualization, and machine learning model development. A range of essential libraries and frameworks supported different components of the system: Pandas and NumPy for data manipulation, Matplotlib, Seaborn, and Plotly for static and interactive visualizations, Scikit-learn for implementing machine learning models, Streamlit for deploying the interactive dashboard, and utility modules such as Warnings and Datetime for system handling and time-based operations, as documented in the *requirements.txt* file. The development environment involved using Jupyter Notebook for exploratory data analysis and model building, along with VS Code or PyCharm as the main Python IDEs for scripting and application development. The Streamlit framework was used to integrate all workflows into a deployable interface, running on a local system configured with Python 3.8 or above.

5.3 Challenges Faced and How They Were Handled

During the project, several challenges were encountered and systematically addressed to ensure smooth processing and accurate results. The first major challenge was the extremely large dataset of 2.6 million rows, which made loading and computation slow; this was resolved by cleaning and filtering the data to a manageable 330,610 high-quality rows, using optimized Pandas operations such as chunk loading and vectorization, and enabling Streamlit caching to improve performance. Another significant issue was the heavy right-skewness present in value columns such as *value_dl*, *value_rs*, and *value_qt*, which impacted model accuracy; this was handled through outlier detection, applying log transformations for visual clarity, standardizing numerical features before training, and choosing the Random Forest model for its robustness to non-linear distributions. Missing and inconsistent data—particularly in fields like *alpha_3_code* and *value_dl*—posed additional challenges, which were mitigated by removing null entries during preprocessing and enforcing type consistency across all columns. The project also faced high cardinality in commodity and country fields, complicating encoding and clustering; this was managed through label encoding, grouping commodities into broader categories, and generating aggregate features to support better clustering.

5.4 Screenshots / sample outputs



6. Results and Discussion

This chapter presents the results obtained from the exploratory data analysis (EDA), machine learning models, and dashboard implementation. The outcomes are interpreted with respect to trade patterns, commodity behaviour, and model performance.

6.1 Experimental Setup

The experiments and analysis were executed on a standard consumer-grade laptop using widely adopted open-source tools and Python libraries.

Software / Tool	Version / Details
Operating System	Windows 10 / Windows 11 (64-bit)
Programming Language	Python 3.8+
Notebook Environment	Jupyter Notebook (Anaconda)
IDE	VS Code / PyCharm
Deployment Framework	Streamlit (for dashboard UI)
Browser	Chrome / Edge (for dashboard access)

Table 6.1 Software Environment

Category	Libraries
Data Processing	Pandas, NumPy
Visualization	Matplotlib, Seaborn, Plotly
Machine Learning	Scikit-learn
Deployment	Streamlit
Utilities	warnings, datetime

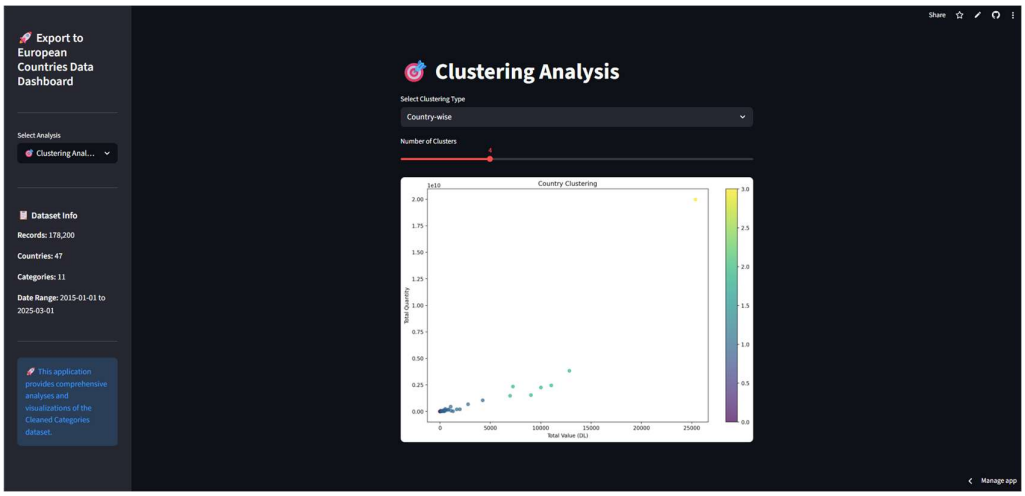
Table 6.2 Python Libraries Used

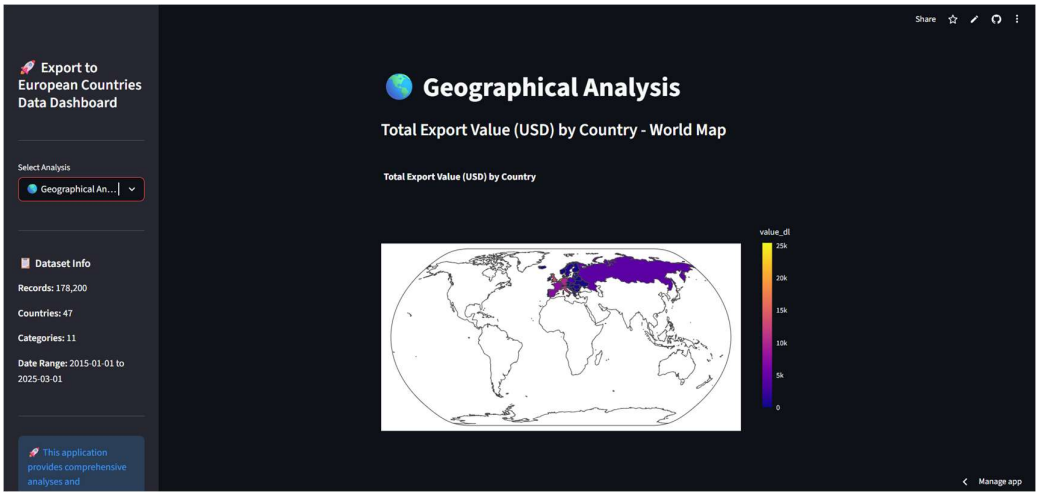
The analysis produced several meaningful insights through both EDA and machine learning. From the exploratory analysis, it was observed that India's top exported commodities to Europe include fuels, textiles, machinery, metals, chemicals, and a variety of agricultural products, with notable differences between commodity frequency and actual export value—some items appeared infrequently yet generated exceptionally high USD revenue. Sub-region analysis revealed Western and Southern Europe as the strongest markets, with significant unit price variability across regions, especially in mineral and textile products. Price dynamics further showed extremely high variance in unit prices, requiring log-scale visualization due to skewed distributions. Time-series trends demonstrated steady export growth over the years with seasonal fluctuations across months and quarters. Correlation analysis showed a very strong relationship between INR and USD values (0.99) and moderate correlation between quantity and USD value, indicating stable monetary patterns but variable quantities depending on commodity type.

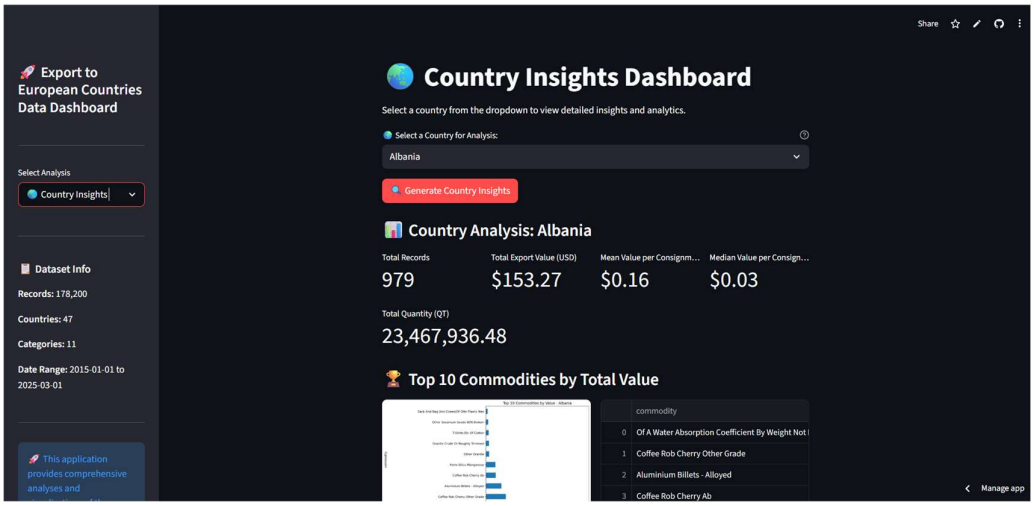
Machine learning results supported these findings: Linear Regression achieved an R^2 score of 0.99, MAE of 0.03, and MSE of 0.11, proving that USD value is highly dependent on INR value. The Random Forest Regressor performed even better, achieving an R^2 of 0.99, MAE of 0.02, and MSE of 0.10, effectively handling non-linear interactions and providing robust predictions for real-world market fluctuations. Logistic Regression successfully classified transactions into high-value and low-value categories with stable accuracy, making it helpful for prioritising logistics and shipment planning. K-Means clustering grouped countries and commodities into meaningful segments, identifying high-value low-frequency markets and low-value high-frequency markets, thereby offering insights into diversification needs and pricing strategies.

Overall, the combined EDA and ML findings highlight Europe as a high-value export destination for India, particularly in Western and Southern regions. Revenue is largely driven by mineral products, metals, and chemicals, while time-series analysis confirms growing export opportunities. The ML models proved highly reliable for value prediction and shipment classification, and clustering revealed strategic market segmentation opportunities. These insights have strong real-world implications: exporters can use the dashboard for country-specific strategies, policymakers can study trade imbalances and niche markets, and commodity-level weaknesses or opportunities can be identified for more informed decision-making.

6.2Graphs, tables, and visualizations







7. Conclusion and Future Work

7.1 Conclusion

This project successfully analyzed India's export patterns to European countries by combining exploratory data analysis (EDA), feature engineering, machine learning models, and an interactive Streamlit-based dashboard. The study began with a large, complex dataset of 2.6 million records, which was cleaned and refined to 330,610 high-quality records suitable for analysis. Through univariate, bivariate, and multivariate analysis, major insights were uncovered regarding the distribution of export values, commodity behaviour, destination country patterns, and sub-regional trade concentration.

The analysis confirmed that Europe remains one of the most valuable export destinations, with countries like the United Kingdom, Germany, and Albania showing high economic engagement. Commodity-wise, Mineral Products, Metals, Textiles, and Chemicals dominate both in frequency and monetary value.

Machine Learning models built as part of the study were highly effective:

- Linear Regression and Random Forest Regression achieved an R^2 score of 0.99, demonstrating almost perfect predictive capability for export USD value.

- Logistic Regression successfully classified exports into high-value vs. low-value categories.
- K-Means clustering revealed distinct groups of countries and commodities, offering insights for strategic export segmentation.

The integrated Streamlit dashboard enhanced interpretability and provided interactive access to insights, making the system suitable for stakeholders, analysts, policymakers, and exporters. Overall, the project met its objectives by delivering both descriptive insights and predictive capabilities, contributing valuable understanding to the domain of international trade analytics.

7.2 Future Work

While the project achieved strong analytical and predictive outcomes, several areas can be expanded in future iterations:

1. Anomaly Detection for Fraud or Irregular Trade Patterns

Future work can include building an anomaly detection module to identify suspicious or irregular trade activities. Using unsupervised learning models such as Isolation Forest, Autoencoders, or Local Outlier Factor (LOF), the system can flag abnormal transactions, sudden spikes in country-level exports, or unusual high-value or low-value shipments. This would help detect fraud, errors, or data inconsistencies, making the system more secure and dependable for stakeholders.

2. Building a Recommendation System for Exporters

A recommendation system can be added to assist exporters by providing actionable insights tailored to their products and markets. Such a system can suggest high-potential countries, identify fast-growing commodity markets, determine the best trade seasons, and highlight competitive pricing strategies. This transforms the dashboard into a powerful decision-support tool, helping exporters optimize their planning and increase market opportunities.

3. Real-Time Data Pipeline

Future improvements can involve creating a real-time data ingestion and processing pipeline. Tools like Apache can be used for streaming data, while Airflow or Prefect can manage automated scheduling and workflow orchestration. The system could continuously clean the data, retrain models, and update dashboards dynamically. This would evolve the current static platform into a real-time export analytics system with instant insights.

4. Enhancing Visual Analytics

The visual analytics component of the dashboard can be further enriched by adding more interactive and intuitive visualizations. Features like prediction sliders, commodity comparison simulators, and advanced 3D geographical charts can improve user engagement and analytical depth. These enhancements will make the interface more powerful, visually appealing, and easier for policymakers and exporters to interpret complex trends.

The study establishes a strong foundation for export analytics and predictive modelling in international trade. With the proposed future enhancements—especially forecasting, deep learning, and cloud deployment—the system can evolve into a powerful, industry-ready platform for strategic trade decision-making.