**Project Milestone 2:** **Data Selection and Project Proposal**

**Summary**

This project aims to leverage predictive analytics to optimize strategies in the sheep and goat marketing sector. By analyzing the FAOSTAT historical dataset, which covers global food and agriculture statistics from 1961 to 2013, we will develop models to forecast export and import quantities and values. The key objectives include identifying trends, comparing country performances, and understanding the correlation between export quantities and values. This analysis will help in making informed decisions to enhance market strategies for sheep and goats.

**Introduction**

The global trade of sheep and goats is a critical component of the agricultural sector, influencing economies and food security. Understanding the dynamics of this trade is essential for developing strategies to optimize market performance. This project seeks to use historical data to identify patterns, forecast future trends, and provide insights into the factors influencing sheep and goat exports and imports. By doing so, stakeholders can make data-driven decisions to improve market outcomes (Cambridge Core, 2020; MDPI, 2023; Springer, 2021).

**Statement of the Problem**

The sheep and goat market faces several challenges, including fluctuating prices, varying demand, and changes in trade policies. Predicting future trends and understanding the factors driving these changes is crucial for developing effective market strategies. This project addresses the need for robust predictive models to forecast export and import quantities and values, providing a foundation for optimizing market strategies and improving trade performance.

**Preliminary Requirements**

To successfully execute this project, the following preliminary requirements are necessary:

Data Access: Access to the FAOSTAT historical dataset from Kaggle.

Computing Resources: Adequate computational power to handle large datasets and perform complex analyses.

Software: Tools such as Python, Jupyter Notebook, and relevant libraries (pandas, scikit-learn, matplotlib, seaborn) for data analysis and visualization.

Domain Knowledge: Understanding of the agricultural trade, specifically the sheep and goat market.

**Summary of the dataset**:

The FAOSTAT historical dataset, sourced from Kaggle, provides comprehensive global food and agriculture statistics from 1961 to 2013, covering over 200 countries. This dataset includes data on more than 25 primary products and inputs. For this analysis, I focused on the "Live Pigs" dataset, specifically examining export quantities (heads of live pigs) and export values (US dollars). Key variables in the dataset include the country (Area), item (Agricultural Products), element (Export Quantity and Export Value), year, unit, and value.

The dataset tracks the export performance of live pigs, offering insights into how quantities and values have evolved over the decades. It reveals trends in global trade, highlights the major exporting countries, and examines the correlation between the number of pigs exported and the financial return from these exports. This rich historical data is invaluable for understanding past market dynamics, informing predictive models, and shaping future agricultural policies and trade strategies. The dataset's granularity allows for detailed trend analysis, country comparisons, and the exploration of economic impacts and external influences on pigs exports. Accessing this data is straightforward using the Kaggle API, making it a valuable resource for researchers and analysts in the agricultural sector.

**Key variables of the dataset are:**

Country (Area): The geographical area represented.

Item (Agricultural Products): The specific agricultural product, focusing on sheep and goats.

Element (Export Quantity and Export Value): Metrics representing the quantity and value of exports and imports.

Year: The year of the record.

Unit: Measurement units for the quantity and value.

Value: The numerical value representing quantities and values of exports and imports.

**Research questions:**

1. Trend Analysis:

1.1. How do the export quantities of live sheep changed from 1961 to 2013 globally and regionally? Are there any noticeable patterns or trends over specific decades?

1.2. What trends can be observed in the export values of live sheep over the same period? Are there any significant fluctuations corresponding to global economic events or changes in agricultural policies?

2. Country Comparison:

2.1. Which countries consistently rank in the top 10 for the highest export quantities of live sheep? How do these rankings change over time?

2.2. Which countries have seen the most significant increases or decreases in export values of live sheep? What economic or political factors might explain these changes?

3. Correlation Analysis:

3.1. What is the correlation between the export quantities and export values of live sheep? Does this relationship vary significantly between different countries or regions?

3.2. How does the correlation between export quantities and values change over different time periods? Are there periods where the correlation is stronger or weaker?

4. Predictive Modeling:

4.1. Can we develop a predictive model to forecast future export quantities of live sheep based on historical data? What are the key predictors that influence these quantities?

4.2. Can a similar predictive model be developed for forecasting export values of live sheep? How accurately can the model predict future values, and what are the significant factors influencing these predictions?

**Methodology**

The methodology follows the data science process, encompassing data importation, inspection, wrangling, preprocessing, actual analysis, visualization, model building and evaluation, and interpretation of insights.

Data Importation and Inspection:

Import Data: Load the dataset using pandas.

Initial Inspection: Examine the dataset structure, missing values, and summary statistics.

Data Wrangling and Preprocessing

Data Cleaning: Handle missing values, outliers, and inconsistencies.

Feature Engineering: Create new features or modify existing ones to enhance model performance.

Data Transformation: Convert categorical variables into numerical values using techniques like one-hot encoding.

Exploratory Data Analysis (EDA):

Trend Analysis: Plot time series data to identify trends and patterns in export and import quantities and values.

Country Comparison: Analyze the performance of different countries over time.

Correlation Analysis: Examine the relationships between different variables.

Predictive Modeling:

Model Selection: Choose appropriate models (Linear Regression, Random Forest, KNN, SVR) based on the problem context.

Model Training: Split the data into training and testing sets and train the models.

Hyperparameter Tuning: Optimize model parameters using techniques like GridSearchCV.

Model Evaluation: Evaluate models using metrics such as MSE, MAE, R², EVS, and MAPE.

**Expected Results**

The expected outcomes of this project include:

Trends and Patterns: Identification of significant trends and patterns in the export and import of sheep and goats.

Predictive Models: Development of accurate predictive models for forecasting export and import quantities and values.

Key Predictors: Insights into the key factors influencing sheep and goat trade.

Optimization Strategies: Recommendations for optimizing market strategies based on predictive insights.

**Execution and Management of the Project**

The project will be executed in phases, each focusing on different aspects of the data science process. Regular meetings and updates will ensure that the project stays on track and any issues are promptly addressed. The team will consist of data scientists, domain experts, and stakeholders from the sheep and goat marketing sector.

**Models and Evaluation Plan**

Models

Linear Regression: Baseline model for comparison.

Random Forest: Handles non-linear relationships and interactions well.

K-Nearest Neighbors (KNN): Simple and effective for certain types of data.

Support Vector Machine (SVR): Robust for high-dimensional spaces.

**Evaluation Metrics**

Mean Squared Error (MSE)

Mean Absolute Error (MAE)

R-squared (R²)

Explained Variance Score (EVS)

Mean Absolute Percentage Error (MAPE)

**Suggested Visualization**

Visualizations will include:

Time Series Plots: To show trends over time.

Bar Charts: To compare country performances.

Scatter Plots: To examine correlations between variables.

Feature Importance Plots: To highlight key predictors in the models.

**Assumptions**

The historical data is representative of future trends.

The models selected are appropriate for the data characteristics.

There is sufficient computational power to handle the dataset and perform analyses.

**Potential Risks**

Data Quality Issues: Incomplete or inaccurate data can lead to incorrect predictions.

Model Overfitting: Models may perform well on training data but poorly on unseen data.

Bias in Data: Historical biases can affect model predictions and perpetuate inequities.

**Ethical Concerns**

Data Privacy: Ensuring that data usage complies with privacy regulations.

Bias Mitigation: Identifying and addressing biases in the dataset to avoid unfair outcomes.

Transparency: Maintaining transparency in model development and interpretation.

**Contingency Plan**

If the original project plan encounters issues:

Alternative Models: Explore other models like Gradient Boosting Machines (GBM) or Neural Networks.

Feature Engineering: Improve model performance by creating new features or using dimensionality reduction techniques.

Data Augmentation: Use techniques to augment the data or obtain additional data smyces to improve model robustness.

By following this comprehensive approach, the project aims to provide valuable insights and optimization strategies for the sheep and goat marketing sector, enhancing decision-making and market performance.

**Addressing Questions of Milestone 2:**

1. What types of model or models do you plan to use and why?

I plan to use the following models for predictive analysis:

Linear Regression: A simple and interpretable model that provides a baseline performance. It is computationally efficient and easy to understand, making it useful for initial analysis and comparison.

Random Forest: An ensemble learning method that builds multiple decision trees and merges them to get a more accurate and stable prediction. It handles non-linear relationships and interactions well.

K-Nearest Neighbors (KNN): A non-parametric method that makes predictions based on the closest training examples in the feature space. It is simple to implement and understand.

Support Vector Machine (SVR): A robust model for regression tasks that tries to find a hyperplane that best fits the data. It is effective in high-dimensional spaces and when the number of dimensions is greater than the number of samples.

2. How do you plan to evaluate your results?

I will evaluate the models using the following metrics:

Mean Squared Error (MSE): Measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated.

Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction.

R-squared (R²): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

Explained Variance Score (EVS): Measures the proportion to which a model accounts for the variance of the outcome.

Mean Absolute Percentage Error (MAPE): Measures the accuracy of a forecast system. It is expressed as a percentage.

3. What do you hope to learn?

I aim to achieve the following:

Identify Key Predictors: Understand which factors significantly impact the export and import quantities and values of sheep and goats.

Model Performance: Compare the performance of different predictive models to determine the most accurate and reliable model.

Data Insights: Gain insights into the trends and patterns in the trade of sheep and goats over the past 15 years.

4. Assess any risks or ethical concerns with your proposal.

Data Quality: Incomplete or inaccurate data can lead to incorrect predictions. Ensuring data quality through thorough cleaning and preprocessing is critical.

Bias in Data: Historical data may contain biases that can affect model predictions. It is important to identify and mitigate these biases to avoid perpetuating them.

Ethical Use of Data: Ensuring that the data used complies with legal and ethical standards, especially if the data contains sensitive information.

5. Identify a contingency plan if your original project plan does not work out.

If the original project plan encounters issues:

Alternative Models: Explore other machine learning models such as Gradient Boosting Machines (GBM) or Neural Networks.

Feature Engineering: Improve the model performance by creating new features or using dimensionality reduction techniques such as PCA.

Data Augmentation: Use techniques to augment the data or obtain additional data smyces to improve model robustness.

6. Include anything else you believe is important.

Model Interpretability: While advanced models may provide better performance, they may also be less interpretable. Ensuring a balance between performance and interpretability is important, especially for stakeholders who need to understand the model's decisions.

Scalability: Ensure that the chosen models can scale with the data size, especially if the dataset grows in the future.

Continuous Monitoring: Implement a system to continuously monitor the model's performance and update it with new data to maintain accuracy over time.

**References**

Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A., & Saker, L. (2023). Crop Prediction Model Using Machine Learning Algorithms. Applied Sciences, 13(16), 9288. https://www.mdpi.com/2076-3417/13/16/9288iCROPM 2020: Crop Modeling for the Future | The Jmynal of Agricultural Science |

Cambridge Core. (2020). https://www.cambridge.org/core/jmynals/jmynal-of-agricultural-science/article/icropm-2020-crop-modeling-for-the-future/94520DEBD9EA9785EB3F2B75BC3AC5F4 Smart Farming Prediction Models for Precision Agriculture: A Review. (2021).

FAOSTAT Data. (n.d.). FAOSTAT Data. FAO. http://www.fao.org/faostat/en/#data

Hindawi. (2021). Broadening the Research Pathways in Smart Agriculture: Predictive Analysis Using Semiautomatic Information Modeling. Hindawi. https://www.hindawi.com/jmynals/complexity/2021/5391823/

Kaggle. (n.d.). FAOSTAT: Food and Agriculture Data. Kaggle. https://www.kaggle.com/datasets/faoorg/faostat-food-and-agriculture-data

MDPI. (2023). Integrated Predictive Modeling and Policy Recommendations for Agriculture. MDPI. https://www.mdpi.com/2076-3417/13/16/9288

Springer. (2021). Smart Farming Prediction Models for Precision Agriculture: A Review. Springer. https://link.springer.com/article/10.1007/s11356-021-12577-3

Routledge. (2021). Predictive Analytics in Smart Agriculture. Routledge. https://www.routledge.com/Predictive-Analytics-in-Smart-Agriculture/Hsu/p/book/9780367424218

ScienceDirect. (2021). Crop Yield Prediction Using Machine Learning: A Review. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0168169921003423

ScienceDirect. (2022). Precision Agriculture Using IoT Data Analytics and Machine Learning. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S153751102200014X

Taylor & Francis. (2022). Decision Support System for Smart Agriculture in Predictive Analysis. Taylor & Francis. https://www.taylorfrancis.com/books/mono/10.1201/9781003241523/decision-support-systems-smart-agriculture-predictive-analysis-syed-akhtar

Project Activities & Weekly Implementation Schedule:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Week** | **Milestone** | **Task** | **MON** | **TUE** | **WED** | **THU** | **FRI** | **SAT** | **SUN** |
| 1 | 1 | Craft rough project plan | x | x | x |  |  |  |  |
| 1 | 1 | Identify a peer & Create Teams folder |  |  |  | x | x | x | x |
| 2 | 2 | Select data | x | x | x |  |  |  |  |
| 2 | 2 | Build project proposal |  |  |  | x | x | x | x |
| 2 | 2 | peer for review & submit milestone 2 | x | x | x | x |  |  |  |
| 3 | 3 | Start preliminary data analysis and visualization | x | x | x | x |  |  |  |
| 3 | 3 | Continue data exploration and begin preprocessing |  |  |  |  | x | x | x |
| 4 | 3 | Complete data cleaning and preprocessing | x | x | x | x |  |  |  |
| 4 | 3 | Select and justify predictive models |  |  |  |  | x | x | x |
| 5 | 3 | Conduct preliminary analysis using selected models | x | x | x | x |  |  |  |
| 5 | 3 | Evaluate initial results and make necessary adjustments |  |  |  |  | x | x | x |
| 6 | 3 | Finalize preliminary analysis | x | x | x |  |  |  |  |
| 6 | 3 | Submit Milestone 3 for peer review and feedback |  |  |  | x | x | x | x |
| 7 | 4 | Complete data preprocessing and finalize models | x | x | x | x |  |  |  |
| 7 | 4 | Evaluate model performance and interpret results |  |  |  |  | x | x | x |
| 8 | 4 | Formulate conclusions and recommendations | x | x | x | x |  |  |  |
| 8 | 4 | Analyze ethical implications and submit Milestone 4 |  |  |  |  | x | x | x |
| 9 | 5 | Project presentation preparation and status review | x | x | x | x | x |  |  |
| 9 | 5 | Begin drafting the final project paper |  |  |  |  |  | x | x |
| 10 | 5 | Continue drafting the final project paper | x | x | x | x |  |  |  |
| 10 | 5 | Start working on the presentation slides |  |  |  |  | x | x | x |
| 11 | 5 | Finalize the project paper | x | x | x | x |  |  |  |
| 11 | 5 | Record the audio/video presentation |  |  |  |  | x | x | x |
| 12 | 5 | Finalize and review the presentation | x | x | x | x |  |  |  |
| 12 | 5 | Submit the final project paper and presentation |  |  |  |  | x | x | x |