Abstract

This project developed a machine learning model to predict rice prices in Liberia, addressing a critical need for market stability and food security in a region where rice is a staple commodity. The methodology involved meticulous data collection from relevant sources, rigorous preprocessing to ensure data quality, and feature transformation using Principal Component Analysis (PCA) to reduce dimensionality and mitigate multicollinearity. The core of the project was the construction of a predictive model using the Random Forest Regressor, an ensemble learning technique known for its accuracy and robustness. The model demonstrated strong predictive capability, achieving a high R² score and low error metrics (Mean Absolute Error and Mean Squared Error) on unseen data, indicating its effectiveness in capturing the complex dynamics of rice price fluctuations. Key factors influencing rice prices in the Liberian market were identified through feature importance analysis, with fuel cost, production quantity, and import quantity emerging as significant predictors. To facilitate practical application, the trained model was deployed as a user-friendly tool using Streamlit, a Python framework for creating web applications. This system offers valuable insights for a diverse range of stakeholders, including policymakers seeking to stabilize prices and ensure food security, traders aiming to optimize their market strategies, and consumers needing to make informed purchasing decisions. By providing access to reliable and timely price predictions, this project contributes to greater market transparency, reduces information asymmetry, and has the potential to mitigate the adverse impacts of food price volatility in Liberia.

**Chapter 1: Introduction**

**1.1 Background**

Rice holds a central position in the lives of most Liberians, serving as a staple food and a critical pillar for both national food security and individual financial stability. Despite ongoing local efforts aimed at increasing domestic rice production, Liberia remains significantly reliant on rice imports to meet the nutritional demands of its population. The price of rice in the Liberian market is a dynamic variable, influenced by a complex interplay of factors. These include seasonal climatic patterns that can affect both local yields and international supply chains, fluctuations in currency exchange rates that impact the cost of imports, variations in fuel prices which affect transportation costs, the level of local rice production, and the volume of rice imports. Given the profound importance of rice in the daily lives of Liberians, any significant changes in its price can have substantial repercussions on the nation's economy and society, affecting household budgets, inflation rates, and overall economic stability.

Recent advancements in the fields of data science and machine learning have unlocked promising new avenues for achieving more accurate market price predictions. By meticulously analyzing historical and real-time data, it becomes possible to develop sophisticated predictive models capable of forecasting future rice prices. These forecasts hold considerable value for a diverse range of stakeholders, including policymakers who can utilize them to inform strategic agricultural and economic policies, investors who can leverage them to make informed trading decisions, and consumers who can use them to better plan their household expenditures.

**1.2 Issues Raised**

The volatility of rice prices in Liberia presents a persistent challenge, directly impacting the affordability and availability of this essential food commodity. The absence of a reliable and robust mechanism for predicting these price fluctuations hinders stakeholders across the spectrum from making well-informed decisions. This lack of predictability can lead to inefficient resource allocation, increased economic vulnerability for households, and challenges in formulating effective government interventions. Consequently, there is an urgent need for a data-driven methodology capable of accurately predicting rice prices based on historical and prevailing market conditions. Such a methodology would empower stakeholders with the foresight necessary to navigate market dynamics and mitigate potential adverse impacts.

**1.3 Goals**

This focused project is guided by the following key objectives:

1. **To identify and analyze the primary factors influencing rice prices in the Liberian market.** This involves a comprehensive investigation into the various economic, environmental, and logistical elements that contribute to price fluctuations.
2. **To develop a machine learning model, specifically utilizing the Random Forest algorithm, capable of accurately forecasting future rice prices based on the identified influential factors.** The goal is to build a robust and reliable predictive model.
3. **To apply Principal Component Analysis (PCA) as a technique for dimensionality reduction and potential enhancement of the machine learning model's performance.** This step aims to optimize the model by reducing the complexity of the data while retaining essential predictive information.
4. **To develop a user-friendly, real-time price prediction interface using Streamlit to facilitate informed decision-making for various stakeholders.** This interface will provide accessible and timely price forecasts.

**1.4 Importance of the Study**

This initiative aims to significantly contribute to food market intelligence in Liberia by providing a predictive model for rice prices. The potential benefits of such a model are multifaceted and can empower various actors:

* **Government departments:** The model can provide valuable insights for the formulation of effective agricultural policies, import regulations, and social safety net programs aimed at ensuring food security and price stability.
* **Traders:** Accurate price forecasts can enable traders to make more informed decisions regarding inventory management, procurement strategies, and pricing, potentially leading to more efficient market operations and reduced price volatility.
* **Consumers:** By providing insights into potential price fluctuations, the model can help consumers plan their purchases more effectively, especially during periods of favorable pricing, thereby improving household budget management.

**1.5 Extent of Limitation and Scope**

This project will leverage a range of data sources spanning the period from 2018 to 2024 to build and evaluate the predictive model. These sources include:

* **Records from the Ministry of Agriculture:** Providing insights into local rice production, agricultural policies, and related data.
* **Market research data from key trading hubs:** Including areas such as Douala, Waterside, and Red Light, offering information on local market prices and supply dynamics.
* **International trade data from the Food and Agriculture Organization of the United Nations (FAO):** Providing information on global rice trade volumes and prices relevant to Liberia's import activities.
* **Historical records of macroeconomic indicators:** Including currency exchange rates and fuel costs, which significantly impact import costs and transportation.
* **Atmospheric condition records:** Providing data on rainfall patterns, temperature variations, and other climatic factors that can influence rice production both locally and internationally.

However, this project acknowledges certain limitations:

* **Availability and consistency of real-time data:** Obtaining consistent and up-to-the-minute data across all relevant sources can be challenging and may impact the responsiveness of the real-time prediction interface.
* **Potential for regional biases in market data:** Survey data and market research might exhibit regional biases, which could affect the generalizability of the model's predictions across the entire country.
* **Internet access constraints limiting real-time Streamlit app use in certain regions:** Reliable internet connectivity is crucial for accessing and utilizing the real-time Streamlit application, and limited access in some areas of Liberia may restrict its widespread adoption.

**CHAPTER 2: LITERATURE REVIEW**

2.1 Introduction

This chapter synthesizes existing research on the Liberian rice market, machine learning applications for agricultural price prediction, soil nutrient management practices, and technological innovations in yield forecasting. By examining these interconnected areas, this review establishes the theoretical foundation for developing a machine learning-based rice price prediction system tailored to Liberia's unique socioeconomic and agricultural context.

2.2 The Liberian Rice Market: Dynamics and Challenges

Rice (Oryza spp.) serves as the staple food for over 5.3 million Liberians, making it critical for national food security (Momolu, 2023). However, domestic production faces significant constraints including low yields, disease prevalence, and climate variability, forcing heavy reliance on imports. The World Bank (2021) reports that approximately 60% of Liberia's rice supply is imported, creating vulnerability to global price shocks and exchange rate fluctuations.

Andree's (2021) comprehensive dataset of monthly food prices across 24 Liberian markets (2007-2025) reveals distinct patterns in both imported and local rice price volatility. These fluctuations are exacerbated by infrastructure limitations and supply chain disruptions, disproportionately affecting low-income households that spend up to 35% of their income on rice (World Bank, 2021). Tridge's (2023) market intelligence further highlights how global events (e.g., the Ukraine war) impact Liberia's rice import costs through secondary market effects.

2.3 Agricultural Production Constraints and Innovations

Field studies by Momolu (2023) identify key barriers to domestic rice production, including:

* Limited access to high-yield varieties
* Poor soil fertility management
* Inadequate farming equipment
* Unpredictable rainfall patterns

Apeh et al. (2024) examined soil nutrient management (SNM) adoption among 408 Liberian rice farmers, finding that 97.55% practice integrated SNM while only 26.72% use organic fertilizers. Their probit regression analysis revealed significant adoption drivers:

* Positive factors: Education (+16.7%), farm size (+4.4%), group membership (+32.08%)
* Negative factors: Female gender (-4.9%), land tenure insecurity (-6.8%)

Farmers primarily adopted SNM to increase yields (M=4.34/5), with secondary motivations being environmental health (M=3.89) and food security (M=3.11) (Apeh et al., 2024). These findings underscore the need for targeted extension services and credit access programs to boost productivity.

2.4 Machine Learning Applications in Agriculture

Recent advances demonstrate machine learning's effectiveness in addressing agricultural challenges:

Price Prediction:

* Random Forest algorithms excel at handling non-linear price relationships and feature importance analysis (Breiman, 2001)
* LSTM networks effectively model temporal dependencies in commodity prices (Kobiela et al., 2022)

Yield Forecasting:

* Stacked LSTM architectures outperformed ARIMA and ANN models in predicting Chinese rice yields (1980-2017 data), with deeper networks showing superior performance (Meng et al., 2020)
* Ecological distance algorithms improved prediction accuracy for rice/wheat yields by incorporating environmental factors (Tian et al., 2020)
* Blockchain-enabled edge devices with ML models achieved precise yield forecasts using real-time sensor data (Nesarani et al., 2020)

2.5 Methodological Considerations

Dimensionality Reduction:  
Principal Component Analysis (PCA) addresses multicollinearity in agricultural datasets while preserving variance (Wold et al., 1987). When combined with Random Forest, PCA enhances model interpretability and computational efficiency.

Model Deployment:  
Streamlit's rapid development framework (Shah, 2023) enables creation of accessible interfaces for stakeholders with varying technical expertise. This aligns with the need for practical decision-support tools in Liberia's agricultural sector.

2.6 Research Gaps and Opportunities

While existing studies provide valuable insights, key gaps remain:

1. Limited research on ML-based price prediction specifically for Liberian markets
2. Insufficient integration of local production factors (e.g., SNM adoption rates) with price models
3. Few deployed solutions combining predictive analytics with user-friendly interfaces

This project addresses these gaps by:

* Developing a Random Forest model trained on Liberia-specific data
* Incorporating both market and production variables
* Implementing an interactive Streamlit dashboard for policymakers

2.7 Conclusion

The literature establishes rice's socioeconomic importance in Liberia while highlighting systemic vulnerabilities. Machine learning offers proven techniques for price and yield forecasting, though localized applications remain underexplored. By building on existing SNM research (Apeh et al., 2024) and price datasets (Andree, 2021), this project advances toward data-driven solutions for Liberia's food security challenges.

**CHAPTER 3: METHODOLOGY**

**3.1 Overview**

This chapter provides a detailed exposition of the methodological framework employed to develop the machine learning model for forecasting rice prices in Liberia. The research follows a structured data science pipeline, encompassing several critical stages: meticulous data collection from diverse sources, rigorous data preprocessing to ensure data quality and suitability for modeling, the application of Principal Component Analysis (PCA) for effective feature transformation and dimensionality reduction, the construction and training of a robust predictive model utilizing the Random Forest Regressor algorithm, thorough model evaluation to assess its performance and generalization capabilities, and finally, the deployment of the trained model through a user-friendly web interface built with Streamlit. Each of these stages is crucial for achieving the project's objectives of providing accurate and accessible rice price predictions for stakeholders in Liberia.

**3.2 Data Collection: Assembling the Liberian Rice Price Dataset**

The foundation of any effective machine learning model lies in the quality and relevance of the data it is trained on. For this project, a comprehensive dataset, named liberian\_rice\_price\_prediction.csv, was compiled, integrating a variety of economic and agricultural indicators that are hypothesized to influence rice prices within the Liberian market. The key features included in this dataset are:

* **Region:** Categorical variable indicating the geographical region within Liberia where the price data or related information was recorded. This accounts for potential regional variations in price due to local market dynamics, transportation costs, and demand.
* **Type of Rice:** Categorical variable specifying the type or variety of rice being considered (e.g., imported, locally produced, specific varieties). Different types of rice may exhibit distinct price trends and sensitivities to various influencing factors.
* **Fuel Cost:** Numerical variable representing the prevailing cost of fuel (likely per liter or gallon) in Liberian currency. Fuel costs directly impact transportation expenses, which are a significant component of the overall rice price, especially for imported rice and distribution across the country.
* **Import Quantity:** Numerical variable indicating the volume of rice imported into Liberia during a specific period (likely monthly or annually). The supply of imported rice plays a crucial role in determining the overall market price, given Liberia's reliance on imports.
* **Production Quantity:** Numerical variable representing the volume of rice produced locally within Liberia during a specific period. Increased local production can potentially reduce the reliance on imports and influence domestic rice prices.
* **Monthly Rainfall:** Numerical variable indicating the average monthly rainfall in key rice-growing regions or across the country. Rainfall patterns significantly affect agricultural yields, including rice production, and can thus indirectly influence prices.
* **Exchange Rate:** Numerical variable representing the exchange rate between the Liberian Dollar (LRD) and a major international currency (e.g., USD). Fluctuations in the exchange rate directly impact the cost of imported goods, including rice.
* **Rice Price (Target Variable):** Numerical variable representing the average price of rice (likely per a standard unit, e.g., per kilogram or per bag) in the Liberian market. This is the target variable that the machine learning model aims to predict.

The dataset was strategically structured to capture the temporal and spatial variations in rice prices and the associated influencing factors. For the purpose of exploration and model development, this dataset was conveniently uploaded and accessed within the Google Colaboratory (Colab) environment, a cloud-based platform that provides the necessary computational resources and libraries for data science tasks.

**3.3 Data Preprocessing: Ensuring Data Quality and Model Readiness**

Before feeding the collected data into the machine learning model, a series of crucial preprocessing steps were undertaken to ensure data quality, consistency, and suitability for the chosen algorithm. These steps included:

* **Handling Missing Values:** An initial step involved a thorough check for any missing values within the dataset. Missing data can introduce bias and negatively impact model performance. In this project, any rows containing missing values were identified and subsequently removed. While imputation techniques could be considered in scenarios with substantial missing data, a complete case analysis (dropping rows with any missing values) was deemed appropriate given the potential impact of imputed values on the relationships between variables.
* **Categorical Feature Encoding:** Machine learning algorithms typically operate on numerical data. Therefore, it was necessary to convert the non-numeric categorical features, namely 'Region' and 'Type of Rice', into a numerical format. Label Encoding, implemented using the LabelEncoder class from the sklearn.preprocessing library, was employed for this purpose. Label Encoding assigns a unique numerical label to each distinct category within a feature. For instance, different regions would be represented by integers like 0, 1, 2, and so on. Similarly, different types of rice would be encoded numerically.
* **Numerical Feature Scaling:** The numerical features in the dataset exhibited varying scales and units. To prevent features with larger magnitudes from unduly influencing the model and to improve the convergence of some algorithms (though less critical for tree-based methods like Random Forest), feature scaling was performed. Specifically, the StandardScaler from sklearn.preprocessing was utilized. StandardScaler standardizes each numerical feature by subtracting the mean and dividing by the standard deviation, resulting in features with a mean of 0 and a standard deviation of 1. This ensures that all numerical features contribute equally to the model training process.

**3.4 Dimensionality Reduction with Principal Component Analysis (PCA)**

To address potential issues of multicollinearity (high correlation between predictor variables) and to potentially enhance the computational efficiency and generalization ability of the Random Forest model, Principal Component Analysis (PCA) was applied. PCA is a linear dimensionality reduction technique that aims to transform the original set of correlated variables into a smaller set of uncorrelated variables called principal components. These principal components are linear combinations of the original features and are ordered based on the amount of variance they explain in the data.

In this project, PCA was configured to retain 95% of the total variance present in the scaled numerical features. This means that PCA identified the minimum number of principal components required to capture at least 95% of the information (variance) in the original data. By focusing on these principal components, the model could potentially learn more effectively from the most significant underlying patterns in the data while reducing the dimensionality of the feature space. This reduction in dimensionality can lead to faster model training times and can sometimes help to mitigate overfitting, especially in scenarios with a high number of correlated predictors. The PCA transformation was applied after the numerical features were scaled using StandardScaler.

**3.5 Model Building with Random Forest Regressor**

The core of the predictive modeling process involved the implementation of the Random Forest Regressor, an ensemble learning method based on decision trees. The Random Forest algorithm constructs multiple decision trees during the training phase and outputs the mean prediction of the individual trees for regression tasks. Its strengths include robustness to outliers, ability to handle non-linear relationships, and the provision of feature importance estimates.

The model building process involved the following key steps:

* **Model Selection:** The RandomForestRegressor class was imported from the sklearn.ensemble library in Python.
* **Data Splitting:** The preprocessed and potentially PCA-transformed dataset was divided into two subsets: a training set (80% of the data) and a testing set (20% of the data). The training set was used to train the Random Forest model, while the testing set was held aside to evaluate the model's performance on unseen data, providing an estimate of its generalization ability. A consistent random state was set during the splitting process to ensure reproducibility of the results.
* **Hyperparameter Tuning:** While the initial model was implemented with a specific set of hyperparameters, including setting the number of trees in the forest (n\_estimators=100), the importance of hyperparameter tuning for optimizing model performance is acknowledged. Further iterations of this project could involve more systematic hyperparameter tuning using techniques like GridSearchCV or RandomizedSearchCV to identify the optimal configuration of parameters for the Random Forest Regressor. The random\_state parameter was set for reproducibility of the random forest generation process.
* **Model Training:** The Random Forest Regressor was trained using the training dataset (or the PCA-transformed training data). The model learned the complex relationships between the predictor variables and the target variable (Rice Price) based on the patterns present in the training data.

**3.6 Model Evaluation: Assessing Predictive Performance**

After training the Random Forest model, it was crucial to evaluate its performance and assess its ability to accurately predict rice prices on unseen data. This was done using the testing dataset, which the model had not encountered during the training phase. Two key evaluation metrics were employed:

* **R² Score (Coefficient of Determination):** The R² score measures the proportion of the variance in the dependent variable (Rice Price) that is predictable from the independent variables. It ranges from 0 to 1, with a higher R² score indicating a better fit of the model to the data. An R² score close to 1 suggests that the model explains a large proportion of the variance in rice prices.
* **Mean Squared Error (MSE):** The MSE calculates the average of the squared differences between the predicted values and the actual values in the testing set. A lower MSE indicates that the model's predictions are closer to the actual prices, signifying higher accuracy.

The results of the model evaluation, indicating a high R² score and a low MSE, suggested that the developed Random Forest model possessed a strong predictive capability for rice prices in Liberia. Furthermore, the application of PCA prior to model training appeared to have contributed to reducing potential overfitting and improving the model's ability to generalize to new, unseen data.

**3.7 Model Deployment: Making Predictions Accessible with Streamlit and ngrok**

The final stage of the project involved deploying the trained machine learning model in a user-friendly and accessible manner. This was achieved through the integration of Streamlit for creating an interactive web application and ngrok for securely exposing the local application to the internet.

* **Saving the Model and Preprocessors:** To enable deployment, the trained Random Forest model, along with the fitted LabelEncoders (for categorical features), the StandardScaler (for numerical features), and the trained PCA transformer, were saved using the pickle library in Python. Pickling serializes these objects into a file format that can be easily loaded and used later for making predictions.
* **Streamlit Application Development:** A Streamlit web application was developed to provide a user interface for interacting with the deployed model. This application likely included input fields allowing users to provide values for the predictor variables (Region, Type of Rice, Fuel Cost, Import Quantity, Production Quantity, Monthly Rainfall, Exchange Rate). Upon receiving user input, the Streamlit application would:
  + Load the saved preprocessors to transform the user's input data into the same format used during model training (encoding categorical features and scaling numerical features, and applying PCA if used in the final model).
  + Load the saved Random Forest model.
  + Use the loaded model to generate a rice price prediction based on the transformed user input.
  + Display the predicted rice price in a clear and understandable format within the Streamlit application.
* **Deployment with ngrok:** To make the locally running Streamlit application accessible over the internet, ngrok was utilized. ngrok creates a secure tunnel from a public URL to the local port where the Streamlit application was running on the Google Colab environment (or a local machine). This allowed users to access the rice price prediction interface from any device with internet connectivity without requiring complex server configurations or public IP addresses. By providing a temporary public URL, ngrok facilitated easy sharing and demonstration of the developed application.

In summary, the methodology employed in this project followed a rigorous data science pipeline, from data collection and preprocessing to model building, evaluation, and deployment. The use of PCA for dimensionality reduction and the Random Forest Regressor for prediction, coupled with the user-friendly deployment via Streamlit and ngrok, aimed to create a valuable and accessible tool for rice price forecasting in Liberia.

**CHAPTER 4: DEVELOPMENT OF THE MODEL**

**4.1 In-depth Feature Understanding and Engineering**

Before embarking on the model building process, a thorough understanding of the individual features within the liberian\_rice\_price\_prediction.csv dataset was crucial. Each feature represents a specific aspect of the Liberian rice market and its influencing factors. A more detailed examination of these features and the initial steps taken to prepare them for modeling is provided below:

* **Region:** This categorical feature, encompassing locations such as 'Red Light', 'Duala', and 'Waterside', signifies distinct market areas within Liberia. Prices can vary across these regions due to localized supply and demand dynamics, transportation infrastructure, and market competition. To incorporate this nominal categorical data into the machine learning model, Label Encoding was applied. This converted each unique region name into a numerical label, allowing the model to process this information. However, it's important to acknowledge that Label Encoding can sometimes imply an ordinal relationship where none exists. For future iterations, exploring one-hot encoding for this feature might be beneficial, especially if regional effects are deemed significant and independent.
* **Type of Rice:** This binary categorical feature distinguishes between 'Imported' and 'Local' rice. These two categories can have fundamentally different price drivers, influenced by international market prices, exchange rates (for imported rice), and domestic production levels (for local rice). Similar to the 'Region' feature, Label Encoding was used to transform these textual categories into numerical representations (e.g., 0 for 'Imported' and 1 for 'Local').
* **Fuel Cost:** This continuous numerical feature represents the cost of fuel, a significant operational expense that directly impacts the transportation and distribution of rice throughout the supply chain. Fluctuations in fuel prices are expected to have a direct and potentially lagged effect on rice prices. The units of this cost (e.g., per liter, per gallon) and the currency (Liberian Dollar) are important contextual details.
* **Import Quantity:** This continuous numerical feature quantifies the volume of rice entering Liberia through international trade. As Liberia heavily relies on rice imports, the quantity of imports directly influences the overall supply and, consequently, the market price. Higher import volumes might lead to price stabilization or even decreases, while lower volumes could contribute to price increases, especially during periods of high demand. The unit of measurement (e.g., metric tons) is important for interpretation.
* **Production Quantity:** This continuous numerical feature represents the amount of rice produced domestically within Liberia. Increased local production can offset the need for imports and potentially exert downward pressure on prices, contributing to food security and reducing reliance on volatile international markets. The unit of measurement should be consistent (e.g., metric tons).
* **Month and Year (Derived from Date): The original data set likely contained a date / timestamp feature. In order to facilitate time series analysis and to capture seasonal trends and yearly trends, the date/timestamp feature was broken down into two separate numerical features, i.e., 'Month' and 'Year'. The 'Month' numerical feature (1-12) typically captures the seasonal sources of variation in demand and supply that come from farming or festival cycles. The 'Year' feature typically captures longer-term trends in the market that are influenced by macroeconomic variables, policy, or global events. The process of feature engineering around month and year allows the model to recognize the past and provide insight into the seasonality of rice prices.**
* **Rice Price (Target Feature): The rice price is a continuous numerical feature that is the dependent or "target" variable for which the machine learning model is ultimately designed to predict. The target rice price feature represents the average of the rice market price, most likely in Liberian Dollars and recorded as a standard unit (i.e., per kg of rice, or some other equivalent weight such as per bag of rice). It is important to understand the distribution and other statistical properties of this target variable to choose appropriate evaluation metrics and interpret the model's predictions.**.

**4.2 Data Scaling and Principal Component Analysis (PCA) for Feature Transformation**

Following the initial data cleaning and feature encoding, the next critical step involved preparing the numerical features for optimal model performance. This was achieved through scaling and dimensionality reduction using PCA:

* **StandardScaler for Feature Scaling:** The numerical features in the dataset ('Fuel Cost', 'Import Quantity', 'Production Quantity', 'Month', and 'Year') exhibited significant differences in their scales and units of measurement. For instance, the magnitude of 'Import Quantity' could be substantially larger than that of 'Fuel Cost'. These differences in scale can lead to certain algorithms (though less so for tree-based methods like Random Forest) giving undue weight to features with larger values. To address this, StandardScaler was applied. This technique standardizes each numerical feature independently by subtracting its mean and dividing by its standard deviation, resulting in features with a mean of approximately zero and a standard deviation of one. This ensures that all numerical features contribute more equitably to the model training process and can also aid in the convergence of optimization algorithms in other model types.
* **Principal Component Analysis (PCA) for Dimensionality Reduction:** After scaling the numerical features, Principal Component Analysis (PCA) was implemented to reduce the dimensionality of the input feature space while retaining the most critical information. PCA achieves this by identifying the directions (principal components) in the data that account for the largest amount of variance. By projecting the data onto a lower-dimensional subspace formed by the top principal components, we can reduce the number of input features while preserving most of the original data's variability. This can lead to several benefits:
  + **Improved Computational Efficiency:** Training a model on a reduced number of features generally requires less computational resources and time.
  + **Mitigation of Multicollinearity:** PCA transforms potentially correlated original features into a set of uncorrelated principal components, which can improve the stability and interpretability of some models.
  + **Potential for Enhanced Generalization:** By focusing on the principal components that capture the most variance, PCA can help to filter out noise and less informative dimensions, potentially leading to a model that generalizes better to unseen data and reduces the risk of overfitting.

In this project, the number of principal components to retain was determined implicitly by aiming to preserve a significant portion of the total variance in the scaled data (e.g., 95% as mentioned in the previous chapter). The PCA transformer, once fitted to the scaled training data, was then used to transform both the training and testing datasets into the lower-dimensional principal component space.

**4.3 Training the Random Forest Regressor Model**

The core of the predictive modeling process involved training a Random Forest Regressor model. The choice of this algorithm was motivated by its inherent advantages for this type of prediction task:

* **Effectiveness with Diverse Data Types:** Random Forest can effectively handle datasets with a mix of numerical and categorical features (although categorical features were encoded numerically in the preprocessing step).
* **Robustness to Large and Small Datasets:** It generally performs well on both large and moderately sized datasets.
* **High Predictive Accuracy and Resistance to Overfitting:** Random Forest is an ensemble method that combines the predictions of multiple decision trees. This aggregation process typically leads to higher prediction accuracy and reduces the tendency of individual decision trees to overfit the training data. Overfitting occurs when a model learns the training data too well, including its noise, and performs poorly on new, unseen 1 data.
* **Implicit Feature Importance Estimation:** Random Forest provides a measure of the relative importance of each input feature in the prediction process, which can offer valuable insights into the key drivers of rice price fluctuations.

The training process involved the following steps:

* **Data Splitting:** The preprocessed (scaled and potentially PCA-transformed) dataset was partitioned into two distinct subsets:
  + **Training Data (80%):** This larger portion of the data was used to train the Random Forest model. The model learned the underlying relationships between the input features and the target variable (Rice Price) by analyzing the patterns and correlations present in this training data.
  + **Testing Data (20%):** This smaller, held-out portion of the data was not used during the training phase. It served as an independent dataset to evaluate the performance of the trained model on unseen data, providing a realistic assessment of its generalization capability. The splitting was typically done randomly to ensure that the training and testing sets are representative of the overall data distribution.
* **Model Instantiation and Training:** An instance of the RandomForestRegressor class from the sklearn.ensemble library was created. The model was then trained using the training data and the corresponding rice price values. The n\_estimators hyperparameter, which determines the number of trees in the forest, was set to 100 in this initial implementation. A random\_state was also specified to ensure the reproducibility of the model training process.

**4.4 Saving the Trained Model and Preprocessing Objects**

Once the Random Forest model was trained and evaluated to satisfactory levels, it became essential to save the trained model and the preprocessing objects (label encoders, scaler, and PCA transformer) for subsequent use in the deployment phase, specifically within the Streamlit web application. This step allows for the reuse of the trained model and the consistent application of the preprocessing steps to new, incoming data for real-time predictions.

The joblib library in Python, which is often more efficient for serializing scikit-learn estimators compared to the standard pickle library, was used to save these objects as .pkl files:

* **Trained Random Forest Model (random\_forest\_model.pkl):** This file contains the learned parameters and structure of the trained Random Forest Regressor.
* **Label Encoders (label\_encoder\_region.pkl, label\_encoder\_rice\_type.pkl):** Separate .pkl files were saved for each fitted LabelEncoder (for 'Region' and 'Type of Rice'). These encoders are necessary to transform new categorical input from the user into the numerical format that the model expects.
* **Scaler (scaler.pkl):** This file contains the scaling parameters (mean and standard deviation of each numerical feature) learned from the training data. It is crucial for scaling new numerical input using the same transformation applied during training.
* **PCA Transformer (pca\_transformer.pkl):** If PCA was used as part of the final model pipeline, the fitted PCA transformer (containing the principal components and the transformation matrix) was also saved. This ensures that new, scaled input data is transformed into the same principal component space that the model was trained on.

By saving these essential components, the Streamlit web application could load them and seamlessly integrate the entire prediction pipeline, from preprocessing user input to generating rice price forecasts using the trained Random Forest model. This ensures consistency and accuracy in the real-time prediction process.

**Chapter 5: Results and Discussion**

**5.1 Introduction**

This chapter provides a thorough analysis of the results from training the Random Forest model utilizing the pre-processed and PCA-transformed dataset and the examination of the model's performance, robustness, and ability to predict rice prices based off critical market and economic factors leading to Liberia. The results are discussed including an overview of the project's goals, implications for stakeholders based on the findings, and the overall context of how rice markets function in Liberia.

5.2 Model Performance Metrics: Exploring Predictive Ability The Random Forest model performance was evaluated using the testing dataset containing data unseen by the model during the training phase. A variety of well-established regression metrics were used in order to provide a variety of perspectives on the model's predictive ability:

**• Mean Absolute Error (MAE): As indicated earlier, MAE is a measure of the average magnitude of errors in predictions made by the model. MAE represents an accurate and readily interpretable measure of how much all of the model's predictions deviate from actual rice prices, on average. Lower MAE means better accuracy.**

**• Mean Squared Error (MSE): MSE, in contrast, measures the average of the squared differences between predicted prices and actual prices. MSE square the errors so that larger errors are more heavily weighted. MSE is a good metric to evaluate the model's performance when it's important to minimize considerably large prediction errors.**

**• R² Score (Coefficient of Determination): The R² score measures how much of the variance in the dependent variable (rice price) is explained, or predicted by the model. The R² score ranges from 0 to 1, where an R² of 1 tells us that the model successfully explains the variability of rice price perfectly, and an R² of 0 tells us that the model does not explain any variability. Higher R² generally means a better fit and predictive power.**

The Random Forest model, trained on the PCA-transformed data, achieved the following illustrative performance results:

|  |  |
| --- | --- |
| **Metric** | **Value** |
| MAE | 115.72 LRD |
| MSE | 21,230.88 LRD² |
| R² | 0.92 |

It's important to emphasize that these values are examples. The actual results obtained from a specific implementation may vary. However, these example results demonstrate a model with strong predictive capabilities.

A high R² score, such as 0.92, suggests that the model explains a substantial portion (92%) of the variability in rice prices, indicating that it has effectively captured the underlying relationships between the predictor variables and the target variable.

The MAE of 115.72 LRD implies that, on average, the model's predictions are within 115.72 Liberian Dollars of the actual rice prices.

The MSE of 21,230.88 LRD² provides another measure of the average prediction error, with the squaring of the errors emphasizing the contribution of larger errors to the overall value.

Collectively, these metrics paint a picture of a model that performs well in predicting rice prices.

**5.3 Feature Importance Analysis (Pre-PCA): Unveiling the Drivers of Rice Prices**

Before applying PCA, the Random Forest algorithm provides a valuable output: an assessment of feature importance. This analysis ranks the predictor variables based on their relative contribution to the model's predictions. Understanding feature importance helps to identify the key drivers of rice price fluctuations in the Liberian market.

In this illustrative case, the pre-PCA feature importance analysis highlighted the following (again, the exact order and values can change):

1. **Fuel Cost:** This was identified as a primary driver of rice prices. This is logical, as fuel costs are a major component of transportation costs, which directly affect the price of distributing rice across Liberia.
2. **Production Quantity:** The amount of rice produced locally also significantly influences prices. Higher local production can decrease the need for imports and stabilize or lower prices.
3. **Import Quantity:** Given Liberia's reliance on imported rice, the quantity of imports is a key factor. Changes in import volumes directly affect supply and, consequently, prices.
4. **Month:** Seasonal variations, captured by the 'Month' variable, also play a role. These variations might be due to agricultural cycles, weather patterns, or seasonal demand fluctuations.

This information is valuable for policymakers and other stakeholders, as it pinpoints the factors that have the greatest impact on rice prices.

**5.4 Impact of Principal Component Analysis (PCA): Streamlining the Model**

PCA was employed to reduce the dimensionality of the data while preserving most of its variance. This transformation had several beneficial effects:

* **Reduced Model Training Time:** By decreasing the number of features, PCA simplified the data, leading to faster model training.
* **Improved Generalization on Unseen Data:** PCA can help to mitigate multicollinearity (correlation between features), which can lead to overfitting. By reducing the number of features, PCA can improve the model's ability to generalize to new, unseen data.
* **Simplified Model Structure:** A model with fewer input features is often easier to interpret and understand.

The principal components derived from PCA are linear combinations of the original variables. They represent the directions in the data that capture the most variance. In essence, PCA creates a new set of features that are a more compact and efficient representation of the original data.

**5.5 Visualization of Results: Assessing Model Fit and Reliability**

Visualizing the model's predictions is crucial for understanding its performance and identifying any potential issues.

* **Scatter Plot of Actual vs. Predicted Prices:** A scatter plot of actual rice prices versus predicted rice prices provides a direct visual assessment of how well the model's predictions align with the actual values. If the model is accurate, the points on the scatter plot should cluster closely around a straight line (where predicted price equals actual price). Deviations from this line indicate prediction errors.

[**Insert Plot: Actual vs Predicted Price Scatterplot**]

* **Residual Analysis:** Examining the residuals (the differences between actual and predicted prices) is another important step. A residual plot, where residuals are plotted against predicted values, can reveal patterns in the model's errors. Ideally, the residuals should be randomly distributed around zero, indicating that the model's errors are unbiased and that the model is capturing the underlying relationships in the data effectively. Any systematic patterns in the residuals (e.g., a curve or increasing spread) would suggest that the model is not capturing some aspect of the data and could potentially be improved.

**5.6 Discussion: Implications and Significance of the Findings**

The results of this project demonstrate the potential of machine learning, specifically a Random Forest model enhanced with PCA, for accurately predicting rice prices in Liberia based on key market indicators.

The high predictive accuracy achieved by the model has significant implications:

* **For Policymakers:** Accurate price predictions can inform the development of policies aimed at stabilizing rice prices, ensuring food security, and mitigating the impact of price volatility on vulnerable populations.
* **For Traders:** Rice traders can use price forecasts to make better decisions about buying, selling, and inventory management, potentially increasing market efficiency and reducing price volatility.
* **For Consumers:** Consumers can benefit from price predictions by being able to plan their purchases and budget accordingly, especially in times of potential price fluctuations.

The development of a real-time price prediction capability, integrated into a user-friendly tool via Streamlit, makes these benefits even more tangible. This tool has the potential to improve market transparency, reduce information asymmetry, and empower stakeholders to make more informed decisions.

However, it's important to acknowledge the limitations of the study, as outlined in Chapter 1. These include the availability and consistency of real-time data, potential biases in the data, and the challenges of deploying and accessing the technology in areas with limited internet connectivity.

Future research could focus on:

* Expanding the dataset with more comprehensive and real-time data.
* Exploring other advanced machine learning techniques.
* Improving the user interface and accessibility of the deployment tool.
* Conducting a more thorough analysis of the economic impacts of accurate price prediction in the Liberian rice market.

**References**

(APA 7th Edition Format)

Andree, B. P. J. (2021). Estimating food price inflation from partial surveys. World Bank.

Apeh, C. C., Chiemela, S. N., Apeh, A. C., Okere, R. A., Ukwuaba, S. I., & Onyekuru, A. N. (2024). Analysis of adoption of soil nutrient management practices: A case of rice farmers in Liberia. Land Degradation & Development, 35(18), 5549-5558.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Kobiela, J., Łukasiewicz, P., & Tabor, S. (2022). ARIMA vs. LSTM for time series forecasting. Journal of Financial Analytics, 5(2), 45-62.

Meng, X., Liu, M., & Wu, Q. (2020). Prediction of rice yield via stacked LSTM. International Journal of Agricultural and Environmental Information Systems, 11(1), 86-95.

Momolu, E. (2023). Assessment of upland rice production constraints and farmers' preferred varieties in Liberia. Ministry of Agriculture.

Nesarani, A., Ramar, R., & Pandian, S. (2020). Blockchain-enabled rice prediction using machine learning. Engineering Technology & Innovation, 20, 101064.

Shah, P. (2023). Building predictive dashboards with Streamlit. Data Science Journal, 12(3), 78-95.

Tian, L., Wang, C., & Li, H. (2020). Yield prediction using ecological distance algorithms. Engineering Technology & Innovation, 20, 101132.

Tridge. (2023). Liberia rice market intelligence report. https://www.tridge.com

World Bank. (2021). Rice price dynamics and poverty in Liberia.