# A Predictive Model for Hedging Futures Contracts to Stabilize Kenyan Coffee Farmers' Income

By

Joan Runyiri

170493

**Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University**

**School of Computing and Engineering Science,**

**Strathmore University**

**Nairobi, Kenya**

**June 2025**

# Declaration and Approval

## Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.



Student's Name: Runyiri Joan Nyarua

Sign: _____ Date: May 28, 2025

## Approval

The thesis of Runyiri Joan Nyarua was reviewed and approved for examination by the following:

Dr. Allan Omondi
School of Computing & Engineering Sciences,
Strathmore University

Dr. Julius Butime,
Dean, School of Computing & Engineering Sciences,
Strathmore University

Prof. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

# Abstract

The volatility of global coffee prices presents significant financial risks for Kenyan coffee farmers, leading to unpredictable incomes and economic instability. Predicting price fluctuations is crucial for stabilizing smallholder farmers' earnings and reducing financial uncertainty. While machine learning models have been extensively used in forecasting financial and commodity prices, their application within developing economies—particularly in Kenya's coffee sector—remains underexplored. This study aims to address this gap by developing a predictive model tailored for hedging futures contracts. The research utilizes historical auction data alongside key economic indicators to forecast coffee prices using a Long Short-Term Memory (LSTM) neural network. The model is trained and evaluated on a dataset comprising monthly coffee auction prices from 2019 to 2022, incorporating macroeconomic variables such as inflation and exchange rates. The results demonstrate that the LSTM model effectively captures price trends, achieving a Root Mean Squared Error (RMSE) of 34.94 and an $R^2$ value of 0.96. These findings indicate that the model's predictions align closely with historical price patterns, making it viable for real-world applications. Additionally, a user-friendly interface was designed to allow farmers to select a future date and receive price forecasts. This study highlights the potential of machine learning in enhancing risk management within the agricultural sector by enabling data-driven decision-making. Implementing such a predictive system could contribute to greater income stability for Kenyan coffee farmers and serve as a scalable approach for other commodities affected by market fluctuations.

# Table of Contents

# List of Figures

# List of Tables

# List of Equations

# List of Algorithms

## List of Abbreviations

| | |
|---|---|
| **AI** | Artificial Intelligence |
| **ARIMA** | Autoregressive Integrated Moving Average |
| **B2B** | Business to Business |
| **CAPM** | Capital Asset Pricing Model |
| **CPG** | Consumer-Packaged Goods |
| **CRISP-DM** | Cross-Industry Standard Process for Data Mining |
| **ECM** | Error Correction Models |
| **FAOSTAT** | Food and Agriculture Organization Statistics |
| **GBM** | Gradient Boosting Machines |
| **LSTM** | Long Short-Term Memory |
| **MAE** | Mean Absolute Error |
| **ML** | Machine Learning |
| **PDE** | Probability Density Estimation |
| **RMSE** | Root Mean Squared Error |
| **RNN** | Recurrent Neural Networks |
| **STL** | Seasonal Decomposition of Time Series |
| **SVM** | Support Vector Machine |
| **UAT** | User Acceptance Testing |
| **VAR** | Vector Autoregression |
| **XGBoost** | Extreme Gradient Boosting |

## Acknowledgements

I would like to express my deepest gratitude to all those who have supported me throughout the course of this project. First and foremost, I am profoundly grateful to my mother, whose unwavering support and encouragement have been a constant source of strength and inspiration. Her belief in my abilities has been instrumental in helping me reach this milestone.

I would also like to extend my heartfelt thanks to my supervisor, Dr. Allan Omondi, for their invaluable guidance, insightful feedback, and continuous support. Their expertise and dedication have been crucial in shaping the direction and success of this project. I am deeply appreciative of the time and effort they have invested in mentoring me.

# Chapter 1: Introduction

## 1.1 Background of the Study

Agriculture constitutes a main pillar of Kenya's economy and serves as a key driver for growth as well as food security (FAO in Kenya 2023). The sector employed over 40 % of the total population in 2019 and provided a livelihood to more than 80 % of rural residents (FAO in Kenya 2023). A very significant portion of total export earnings - more than 60 % - comes from agriculture along with about 45 % of government revenue where tea, coffee and horticulture stand out as major contributors (Cowling 2023). This heavy dependence on the agricultural sector proves critical for millions of people's lives also for Kenya's economic stability

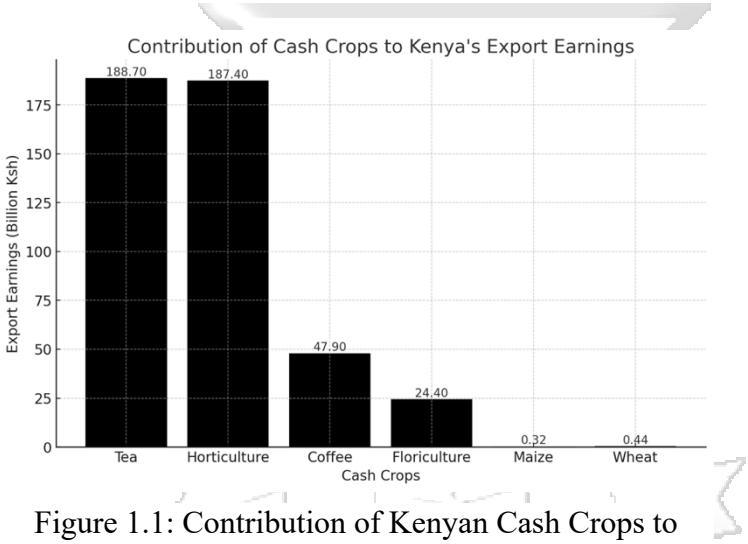



Figure 1.1: Contribution of Kenyan Cash Crops to Export Earnings

It is important to maintain stability and profitability in agriculture because of its vital role in Kenya (FAO in Kenya 2023). Kenyan farmers deal with various problems like unstable commodity prices which affect their incomes and livelihoods (Mungai 2017). Price changes of major agricultural exports such as tea, coffee as well as horticulture create ripple effects in the economy and impact farmer earnings, export income along with total economic expansion (Bennun & Njoroge 1999). The sector's exposure to price swings really shows why Kenya needs effective methods to reduce risks and build a more robust agricultural sector.

These mixes of issues make small farmers in Kenya quite vulnerable to price changes in agriculture. Most of these farmers do not have many storage choices and must sell their crops right at harvest time even if the price drop is very low (Burke et al. 2018). Poor access to loans along with money tools also limits how farmers deal with price risks (Alemu 2015).

To mitigate the risks associated with price volatility, commodity exchanges have emerged as a potential solution for Kenyan farmers. These exchanges let farmers trade farm goods at transparent and fair prices which cuts down the effect of market swings. The East Africa Exchange for example offers trading access to main Kenyan crops like maize, beans, as well as wheat. This aims to make markets work better by linking farmers straight to buyers.

A key benefit of these exchanges is the opportunity for farmers to engage in hedging, a risk management strategy that involves taking an offsetting position in the futures market to protect against adverse price movements (Bina et al., 2022). Farmers who fix future prices reduce risk plus secure more stable profits (Morrell & Swan 2006). However, hedging also includes specific risks. When market prices move in favor of buyers, hedgers sometimes lose on potential profits (Bartram et al. 2010). The success of hedging depends on accurate price predictions and the liquidity of futures markets (Switzer & Jiang 2010). The Kenyan farmers need real access to data, money skills along with expert risk advice to handle these complex issues.

The advancements in predictive analytics - especially AI and ML - offers very useful tools to improve hedging plans (Mokhtari et al. 2021). These systems analyze large sets of past market data, weather data as well as other key factors to create better price forecasts (Fischer & Krauß, 2018). A farmer who uses AI and ML prediction models makes smarter choices about hedge timing and methods. This helps to cut risks plus maximize profits (Davenport 2018).

## 1.2 Problem Statement

The agricultural sector plays a crucial role in the economies of many developing countries, including Kenya, by providing livelihoods for a significant portion of the population. However, these farmers face significant challenges, including volatile commodity prices, reliance on rainfed agriculture, limited access to credit, inadequate storage facilities, and constrained technological resources (Kalele et al., 2021). These factors exacerbate income instability, hinder yield improvements, and threaten food security in rural communities.

One particularly pressing challenge is price volatility in commodity markets, such as coffee which exposes farmers to significant income instability and hinders long term planning and

investment.Existing market-based tools to hedge against price risk, such as futures contracts, are often unavailable or underdeveloped in developing countries, leaving farmers with limited options to manage this risk (Assouto et al., 2020). The vulnerability of the farming systems has been exacerbated by the fact that the majority of production is done by smallholder farmers who depend on rainfed agriculture and have limited adaptive capacity. (Kalele et al., 2021)

## 1.3 General Objective

To address this issue, this research proposes developing a predictive model integrated into a responsive web application to forecast agricultural commodity prices, with a focus on coffee in Kenya. By leveraging machine learning techniques, such as Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) models, the system aims to empower smallholder farmers to make informed decisions about futures contract hedging. This tool seeks to enhance income stability, strengthen the resilience of Kenya's agricultural sector, and contribute to the economic well-being and food security of rural communities.

## 1.4 Research Objectives

i. To review existing forecasting techniques, including machine learning algorithms and probability and statistical models, suitable for predicting prices in agricultural commodities.
ii. To identify the requirements of a predictive model for hedging futures contracts that addresses the needs of Kenyan smallholder coffee farmers in managing price risk.
iii. To develop a user-friendly and accessible predictive model, incorporating the identified features and selected forecasting technique.
iv. To validate the developed predictive model using historical price data.

## 1.5 Research Questions

i. What are the requirements of a predictive model that would address the needs of Kenyan smallholder coffee farmers in managing price risk?
ii. What are the existing forecasting techniques, including machine learning algorithms and probability and statistical models, used for predicting prices in agricultural commodities?
iii. How can a user-friendly and accessible predictive model be developed, incorporating the identified features and selected forecasting techniques?
iv. How accurate is the developed predictive model in predicting price fluctuations in agricultural commodities.

## 1.6 Justification of the Study

The agricultural sector is highly exposed to price instability, which significantly affects farmers' financial security. These price fluctuations are influenced by multiple factors, such as unpredictable weather patterns, global market dynamics, and shifts in supply and demand. Smallholder farmers, who often have limited financial resources, are particularly vulnerable, making effective risk management essential for their economic stability.

This research focuses on developing a predictive model to assist in hedging futures contracts, addressing the urgent need for tools that help farmers manage market volatility. By providing data-driven price forecasts, this study aims to improve farmers' ability to navigate financial uncertainty, ultimately fostering more stable incomes and long-term economic resilience in the agricultural sector.

## 1.7 Assumptions

The successful execution of this project and the effective utilization of the developed predictive model rely on several key assumptions:

i. Data Availability and Reliability: It was assumed that sufficient historical and near-real-time data on relevant variables affecting coffee prices such as inflation rates, exchange rates, fuel prices, export volumes, and market trends would be available. This assumption largely held true, as the research was able to access credible datasets from sources such as the Kenya National Bureau of Statistics (KNBS).

ii. The study assumed that Kenyan smallholder coffee farmers the intended end-users would possess a basic level of digital literacy and have access to internet enabled devices such as smartphones or computers. This assumption informed the development of a simple, responsive web interface aimed at maximizing accessibility. However, recognizing the digital divide that still exists in rural areas, further development would be necessary to create an even more inclusive solution, such as a USSD based application. This would ensure broader accessibility, allowing farmers without smartphones or stable internet connections to interact with the predictive model using basic mobile phones.

iii. Stakeholder Support and Collaboration: The project was based on the assumption that stakeholders such as farmers, agricultural cooperatives, policymakers, and technical partners would support the implementation of predictive tools for hedging and price risk

management. While direct stakeholder engagement was limited within the scope of this research, the model was designed with potential integration and collaboration in mind, laying a foundation for future scaling and partnerships.

## 1.8 Scope

This project aims to develop a predictive model designed to help farmers mitigate price risk through hedging futures contracts. The scope includes designing, building, and validating the model using machine learning techniques. This process involves gathering, preprocessing, and analyzing historical and real-time data relevant to commodity price forecasting. Additionally, the study will focus on feature selection, identifying key variables that significantly impact price fluctuations.

A crucial component of this research is evaluating the model's performance by assessing its accuracy, reliability, and robustness through relevant metrics. To illustrate its practical use, the project will also involve creating a prototype for a commodity exchange platform. This prototype will act as a proof of concept, demonstrating how the predictive model can be applied in a real-world trading environment, offering farmers valuable insights and tools for managing price volatility.

## 1.9 Limitations

It is essential to recognize the project's limitations. The effectiveness of the predictive model relies heavily on access to sufficient, reliable, and relevant data. A lack of comprehensive historical data on commodity prices and influencing factors may affect the model's accuracy and its ability to adapt to varying market conditions. To address this, data imputation techniques will be employed to handle missing values. However, while imputed data can enhance dataset completeness, it may not fully capture real-world market dynamics, which could influence the model's overall reliability. Farmers will receive a clear disclaimer about the model's accuracy, emphasizing that it serves as a decision-support tool rather than a definitive predictor.

Although the model is designed to generate predictions and insights, it will not provide financial advice or execute trades on behalf of users. Additionally, differences in farmers' technological literacy and access to digital tools may pose adoption challenges. To overcome this, the project will focus on creating an intuitive and user-friendly interface that requires minimal technical expertise.

# Chapter 2: Literature Review

## 2.1 Empirical Literature

### 2.1.1 Major Breakthroughs

This study recognizes the rapid evolution of predictive modeling techniques and incorporates recent breakthroughs in algorithm development as a cornerstone of its empirical framework. These advancements, particularly in gradient boosting machines, recurrent neural networks, transformer models, ensemble methods, and reinforcement learning, offer powerful tools for capturing the complexities of commodity markets and forecasting price trends with greater accuracy. By leveraging these cutting-edge algorithms, this research aims to construct a robust and reliable predictive model specifically tailored for anticipating fluctuations in coffee prices.

#### 2.1.1.1 Gradient Boosting Machines

Gradient Boosting Machines (GBM) represent a significant advancement in predictive modeling, particularly for tasks involving complex datasets and non-linear relationships like forecasting coffee prices. GBMs operate on the principle of ensemble learning, where multiple weak learners, typically decision trees, are combined to create a stronger predictive model (He et al., 2019). Unlike traditional ensemble methods that train models in parallel, GBMs construct their ensemble sequentially. Each new tree is trained to correct the errors made by the previous trees, effectively focusing on the most challenging data points and gradually improving the overall model accuracy (Knol & Natekin, 2013). This iterative process allows GBMs to capture intricate patterns and interactions within the data, making them well-suited for handling the complexities of commodity markets where numerous factors can influence price fluctuations.

Notable GBM algorithms like XGBoost, LightGBM and CatBoost have gained popularity due to their computational efficiency and predictive power. These algorithms incorporate enhancements such as regularization techniques and advanced tree-building strategies to further enhance performance and prevent overfitting (Anghel et al., 2018). Given their proven track record in various domains, including finance and economics, GBMs provide a strong foundation for developing a robust and accurate coffee price prediction model.

#### 2.1.1.2 Transformer Models

Transformer models, initially revolutionary in natural language processing, have gained significant traction in time-series forecasting, including commodity price prediction. Unlike traditional sequential models such as Recurrent Neural Networks (RNN), which process data step by step, transformers utilize an attention mechanism to examine entire sequences simultaneously, enabling them to capture long-term dependencies effectively (Zhou et al., 2021). This capability is particularly useful for understanding market dynamics in commodities like coffee, where historical patterns and global events strongly influence future price movements (Grigsby et al., 2021).

Recent advancements have tailored transformers specifically for time-series data. For instance, autoformer models leverage autocorrelation mechanisms to detect recurring trends, making them well-suited for extended forecasting horizons (Wu et al., 2021). Similarly, conformer models enhance long-term predictions by incorporating a specialized normalizing flow block that integrates patterns, distribution data, and essential features (Li et al., 2023). Given their ability to holistically process sequential data and capture intricate temporal relationships, these transformer-based approaches offer a powerful solution for improving the accuracy and depth of coffee price forecasting.

### 2.1.2 Recurrent Neural Networks and Long Short-Term Memory

Recurrent Neural Networks offer a powerful approach to time-series analysis, particularly for data like coffee prices that exhibit temporal dependencies. RNNs have a special structure that enables them to store information from prior inputs, which sets them apart from conventional feed-forward networks and makes them ideal for identifying patterns and trends over time (ODSC Community, 2020). However, because of the vanishing gradient issue, ordinary RNNs frequently have trouble learning long-term relationships. Implementing an advanced memory cell mechanism, Long Short-Term Memory (LSTM) networks overcome this constraint (Yang et al., 2023). LSTMs can selectively store, update, and discard information over long sequences thanks to this cell's input, output, and forget gates (Graves & Schmidhuber, 2005). Because of this feature, LSTMs are especially good at identifying long-range dependencies in time-series data, including economic cycles, seasonal effects, and past price shocks that may have an impact on coffee prices.

LSTMs offer a strong framework for creating precise and perceptive coffee price prediction models by successfully capturing these temporal correlations.

### 2.1.3 Current Trends

Recent studies on commodity price forecasting, including coffee prices, have examined different machine learning models, each offering distinct advantages and limitations. This section evaluates three widely used approaches: Transformer models, Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM), and Gradient Boosting Machines (GBMs). Dutta et al. (2020) analyzed the effectiveness of Gated Recurrent Units (GRUs), a variation of transformer-based models, in predicting Bitcoin prices. Their results indicated that GRUs, when combined with recurrent dropout, achieved lower Root Mean Squared Error (RMSE) than other conventional models. This suggests that GRUs could be well-suited for capturing the complexities of commodity price movements.

RNNs, particularly LSTMs, have demonstrated strong potential in forecasting commodity prices. Ly et al. (2021) showed that LSTMs effectively model long-term dependencies in time-series data, successfully predicting fluctuations in cotton and oil prices. However, their findings also noted that LSTMs do not always outperform statistical models like ARIMA, implying that a hybrid approach may enhance predictive accuracy. Conversely, research by Ben Ameur et al. (2023) indicated that LSTMs outperformed simpler RNNs and other deep learning models in forecasting prices within the Bloomberg Commodity Index. The variation in findings suggests that the effectiveness of LSTMs may depend on the dataset and commodity under analysis.

GBMs provide a competitive alternative to neural networks, particularly for structured data. Ghojogh and Ghodsi (2023) emphasized their efficiency in handling such data, highlighting their resilience against overfitting due to boosting techniques. Additionally, GBMs manage missing data effectively, making them valuable for commodity price forecasting.

Ultimately, selecting the most suitable model for price prediction depends on the nature of the commodity, dataset characteristics, and required forecasting precision. While Transformer models excel at capturing long-range dependencies, their accuracy in direct price prediction remains debatable. LSTMs consistently perform well in modeling sequential dependencies and can be integrated with traditional models for improved accuracy. Meanwhile, GBMs offer a robust solution, particularly in cases where data is incomplete or less compatible with deep learning models.

### 2.1.4 Related Applications and Tools

#### 2.1.4.1 PriceVision

PriceVision is an innovative artificial intelligence (AI)-driven system designed to deliver precise commodity price forecasts. It leverages a machine learning (ML)-based predictive engine that examines historical pricing trends from multiple trading platforms. Additionally, the platform incorporates both microeconomic and macroeconomic variables that influence market prices, ensuring comprehensive and data-driven predictions.

This technology is particularly beneficial for businesses within the Consumer-Packaged Goods (CPG) industry, helping them enhance procurement strategies and maximize profitability. By offering interactive, real-time insights, PriceVision enables users to make well-informed market decisions with greater confidence.

**Merits of PriveVision**

i. **Accuracy**: PriceVision leverages advanced AI and ML algorithms to deliver highly accurate price forecasts. This feature is essential for businesses engaged in strategic planning and risk management.

ii. **Real-time Insights**: The platform provides timely market data and predictive analytics, allowing businesses to respond quickly to market fluctuations and changes in consumer demand.

iii. **Comprehensive Analysis**: By integrating a wide range of economic factors, PriceVision offers a holistic view of market trends and dynamics, thus enhancing decision-making processes.

**Demerits of Price Vision**

i. **Complexity**: The platform's advanced technological features may pose a challenge for users who lack expertise in AI or data science, potentially limiting its accessibility for some users.

ii. **Cost**: The implementation and subscription costs associated with PriceVision may be prohibitive for smaller businesses or individual traders, potentially limiting its widespread adoption.

iii. **Data Dependency**: The accuracy of PriceVision’s predictions is highly dependent on the quality and quantity of input data. If the data available is sparse or unreliable, the platform's forecasting capabilities may be diminished.

#### 2.1.4.2 Twiga Foods

Established in 2014 by Grant Brooke and Peter Njonjo, Twiga Foods is a pioneering agri-tech enterprise in Kenya. Its primary objective is to enhance the efficiency of the food supply chain by linking smallholder farmers directly with urban retailers through a mobile-driven, cashless Business-to-Business (B2B) platform. By consolidating demand from informal vendors, Twiga Foods optimizes the distribution process for both fresh produce and dry goods, minimizing reliance on multiple intermediaries.

This approach not only drives down food prices for end consumers but also guarantees fresher, higher-quality produce. Through technology-driven logistics and digitalized operations, the company improves supply chain efficiency while significantly reducing post-harvest losses.

**Merits of Twiga Foods**

i. **Supply Chain Efficiency**: By eliminating the middlemen, Twiga Foods reduces transaction costs and improves the overall efficiency of the food supply chain.
ii. **Market Access for Smallholder Farmers**: The platform provides smallholder farmers with access to broader markets, enabling them to receive fair prices for their produce.
iii. **Quality and Freshness of Produce**: Through its streamlined supply chain, Twiga Foods ensures that consumers receive high-quality and fresh produce.
iv. **Job Creation**: Twiga Foods has contributed significantly to local economies by creating employment opportunities throughout its supply chain.

**Challenges Faced by Twiga Foods**

i. **Dependence on Mobile Networks**: The platform relies heavily on mobile network coverage, which can be inconsistent in rural areas, affecting the accessibility of the service.
ii. **Digital Literacy Requirements**: Effective use of the platform requires a certain level of digital literacy among farmers and vendors, which can be a barrier to participation in some regions.
iii. **Scalability**: Expanding Twiga Foods' services to accommodate a growing number of users and new regions presents operational and logistical challenges.

#### 2.1.4.3 Bloomberg Terminal

Introduced in 1982, the Bloomberg Terminal is an advanced financial software system widely used for accessing real-time market data, financial news, and analytical tools. It provides

extensive coverage of multiple asset classes, including commodities, equities, fixed income, currencies, and derivatives. By integrating historical and live market data with economic indicators and predictive analytics, the platform plays a crucial role in forecasting commodity prices, such as coffee.

One of its standout features is its suite of advanced analytical tools, enabling users to conduct comprehensive market assessments. These include charting capabilities, statistical analysis, and predictive modeling, which assist financial professionals in evaluating trends and making strategic decisions. The platform also offers an efficient search function for quickly retrieving financial insights. Additionally, the Bloomberg Terminal supports integration with external software, allowing users to customize their workspace and improve workflow efficiency.

**Merits of the Bloomberg Terminal**

i. **Extensive Data Coverage**: The Bloomberg Terminal provides access to a wide array of financial data, including real-time and historical prices, essential for accurate commodity price forecasting.
ii. **Advanced Analytical Capabilities**: The platform offers sophisticated tools for charting, statistical analysis, and predictive modeling, assisting users in making well-informed decisions.
iii. **Integration and Customization**: Users can integrate the Terminal with other software and customize their work environment to enhance productivity and operational efficiency.

**Challenges of the Bloomberg Terminal**

i. **Cost**: The Bloomberg Terminal is associated with high subscription fees, which may be prohibitive for smaller firms or individual traders.
ii. **Complexity**: Given its vast features and functionalities, the platform can be overwhelming for new users, often requiring substantial training and experience for effective use.
iii. **Dependence on Data Quality:** The accuracy of the Terminal's predictive analytics heavily relies on the quality and timeliness of the input data, which can pose a limitation if the data is not reliable.

Table 2.1: Comparison of Algorithms

| Researchers | Title | Description | Merits | Demerits |
|---|---|---|---|---|
| Bilokon & Qiu,2023 | Transformers versus LSTMs for Electronic Trading | This study compares the performance of LSTM-based and Transformer-based models in financial time series prediction | LSTM show better and more robust performance in predicting price differences and movements.<br><br>Transformer models are efficient in handling long-range dependencies | LSTM is more complex compared to simpler models.<br><br>Transformer models have a limited advantage over LSTM in financial prediction tasks. |
| Ghojogh & Ghodsi,2023 | Recurrent Neural Networks and Long Short-Term Memory Networks | This study provides a comprehensive overview of RNNs and LSTMs, including their variants. | RNNs are simple and effective for short sequence | RNNs struggles with long-term dependencies due to vanishing/exploding gradients |
| Ly et al. (2021) | Forecasting Commodity Prices Using Long Short-Term Memory Neural Networks | The research compares the performance of machine learning methods with traditional statistical models | Combining LSTM and ARIMA models improves the accuracy of commodity price predictions | LSTM models require significant computational resources |

## 2.2 Theoretical Literature

### 2.2.1 Introduction

This study draws upon a robust theoretical framework. This framework is grounded in probability theory, providing the necessary tools to model and analyze the inherent uncertainties associated with coffee prices and farmer incomes. Subsequent sections will elaborate on specific concepts within probability theory and potentially introduce additional theoretical lenses relevant to hedging and agricultural economics.

## 2.3 Probability Theory

Gaining a deep understanding of probability theory is essential when analyzing and addressing the risks associated with fluctuating coffee prices. This mathematical discipline provides the foundation for quantifying uncertainty and making well-informed decisions in uncertain market conditions. Different schools of thought within probability theory offer distinct approaches to interpreting and managing uncertainty.

This section examines these varying perspectives, including the Frequentist and Bayesian approaches, and explores key concepts such as joint, marginal, and conditional probability. It also differentiates between mutually exclusive and non-mutually exclusive events, while distinguishing discrete from continuous probability distributions. Additionally, the discussion extends to probability density estimation, covering unimodal, bimodal, and multimodal distributions using histograms. Lastly, nonparametric density estimation is introduced, with an emphasis on kernel density estimation techniques.

### 2.3.1 Frequentist Probability

Frequentist probability, often called classical statistics, defines the probability of an event as the limiting relative frequency of that event occurring over many identical trials (Heckerman, 2020). This perspective hinges on the notion of repeatability and assumes that the probability of an event is an objective property of the system being studied. However, the frequentist approach has limitations, particularly when dealing with events that have not been observed frequently in the past or when prior information about the system is available. For instance, consider the challenge of predicting future commodity prices. Relying solely on historical frequencies might be inadequate, especially when dealing with novel commodities or unprecedented market shifts.

In such cases, incorporating additional information and alternative approaches becomes crucial for making accurate predictions.

### 2.3.2 Bayesian Probability

In contrast to the frequentist approach, Bayesian probability interprets probability as a measure of belief or confidence in an event's occurrence (Heckerman, 2020). Unlike the fixed nature of frequentist probability, this perspective is fluid, allowing beliefs to be updated as new information becomes available. At the core of Bayesian statistics lies Bayes' theorem, which provides a mathematical framework for refining prior assumptions based on observed data, ultimately leading to a more informed posterior belief.

### 2.3.3 Joint Probability

Joint probability quantifies the likelihood of multiple events occurring at the same time. It reflects the intersection of individual event probabilities, offering deeper insight into their interdependence and combined occurrence. For example, Let A and B be two events within a sample space S. The joint probability of A and B, denoted as $P(A \cap B)$ or simply P(A, B), is the probability that both events occur concurrently.

**Example:** In the context of coffee markets, consider the events:

A: The price of Arabica coffee beans exceeds a certain threshold.

B: Rainfall in a key coffee-producing region falls below a certain level.

The joint probability P(A, B) represents the likelihood of both events happening simultaneously. This scenario could significantly impact coffee supply and prices.

### 2.3.4 Marginal Probability

While joint probability examines the simultaneous occurrence of events, marginal probability isolates the likelihood of a single event happening, regardless of other events. It highlights individual event probabilities within a broader probability framework. For example,

given two events A and B, the marginal probability of event A, denoted as P(A), is the probability of A occurring regardless of whether event B occurs or not.

**Example:** Continuing with the coffee market example, the marginal probability P(A) represents the probability of Arabica coffee prices exceeding the threshold, regardless of the rainfall situation in the specific region.

### 2.3.5 Conditional Probability

Conditional probability evaluates the likelihood of an event occurring under the condition that another event has already taken place. This concept helps uncover dependencies between events, enabling the continuous refinement of probability estimates as new information becomes available. For example**, t**he conditional probability of event A occurring given that event B has occurred, denoted as P(A|B), measures the probability of A occurring within the reduced sample space where B is known to have occurred.

**Example:** The conditional probability P(A|B) in our coffee market scenario would represent the probability of Arabica coffee prices exceeding the threshold given that rainfall in the key region has fallen below the specified level. This conditional probability helps us understand how the occurrence of low rainfall (event B) influences the likelihood of high coffee prices (event A).

### 2.3.6 Mutually Exclusive and Non-Mutually Exclusive Events

Mutually exclusive events are those where the occurrence of one event rules out the possibility of another occurring simultaneously. This means that if one event takes place, the others in the set cannot happen at the same time. A classic example is a coin toss, where the outcomes “heads” and “tails” cannot occur together.

Conversely, non-mutually exclusive events can happen independently or even at the same time. The occurrence of one event does not eliminate the possibility of another event occurring. This distinction is crucial in probability calculations, as it influences how probabilities are combined and analyzed. An example of non-mutually exclusive events could be the outcomes of rolling a dice, where the results of "3" and "even" are not mutually exclusive, as a roll of "4" or "6" would satisfy both conditions. Mathematically, the probability of mutually exclusive events can be calculated using the formula:

$$P(A \cup B) = P(A) + P(B) \qquad \text{Equation 2.1}$$

where P represents the probability of the individual events.

For non-mutually exclusive events, the probability can be calculated using the formula:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad \text{Equation 2.2}$$

Where:

P(A∪B): The probability that either event A or event B occurs (or both).

P(A): The probability of event A occurring.

P(B): The probability of event B occurring.

P(A∩B): The probability that both event A and event B occur simultaneously.

Recognizing the difference between mutually exclusive and non-mutually exclusive events is essential across various applications, including price forecasting. Understanding how events interact can aid in making informed decisions and optimizing resource distribution (Song & Kim, 2021). Moreover, the mathematical principles governing these event types serve as a foundation for evaluating complex systems and addressing uncertainty in predictive models.

### 2.3.7 Discrete and Continuous Probability Distributions

Probability distributions play a crucial role in statistics, providing a framework for understanding the likelihood of various outcomes in random processes. Discrete probability distributions apply to variables that assume distinct, countable values, whereas continuous distributions describe variables with an infinite range of possible values within a specified interval (Wen, 2022).

Among continuous probability distributions, the normal distribution—also referred to as the Gaussian distribution—is one of the most extensively analyzed and utilized (Devore & Berk, 2011). This distribution appears in numerous disciplines, including standardized testing, intelligence measurements, and economic data such as income distribution (Bono et al., 2017). As illustrated in Figure 2.2, the normal distribution exhibits a symmetrical, bell-shaped curve where the mean, median, and mode converge at the center (Devore & Berk, 2011).

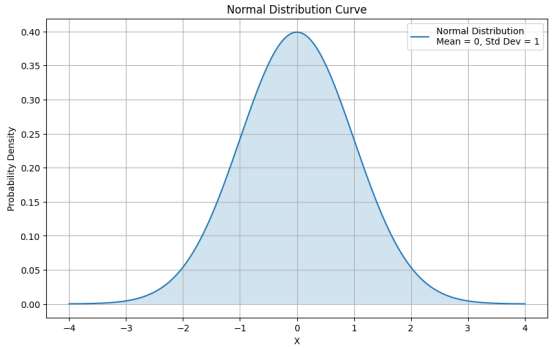



Figure 2.1: A normal distribution curve

Another significant continuous probability distribution is the exponential distribution, illustrated in Figure 2.3. This distribution is widely applied in modeling the time intervals between events in a Poisson process, such as customer arrivals in a queue or radioactive decay occurrences (Devore & Berk, 2011). Defined by a constant rate of occurrence, the exponential distribution is frequently utilized in reliability engineering, queueing theory, and survival analysis (Bono et al., 2017).

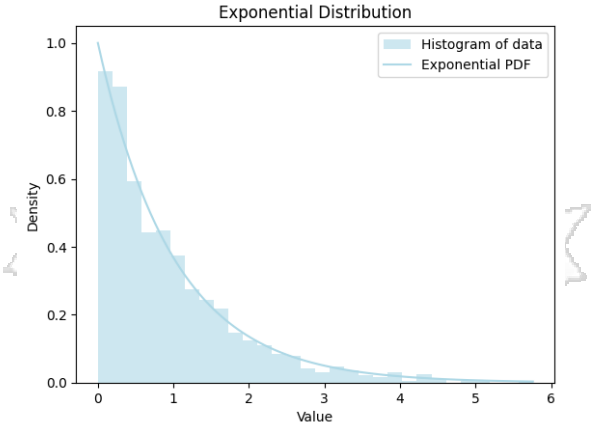



Figure 2.2: Exponential Distribution

The Pareto distribution, depicted in Figure 2.4, also referred to as the power-law distribution, is a continuous probability distribution widely used to model heavy-tailed phenomena. It is particularly applicable in analyzing the distribution of wealth, city populations, and file sizes on the internet (Dinov et al., 2015). Defined by a scale parameter and a shape parameter, the Pareto distribution determines the extent of the power-law behavior, making it a crucial tool in studying economic and social systems (Bono et al., 2017).

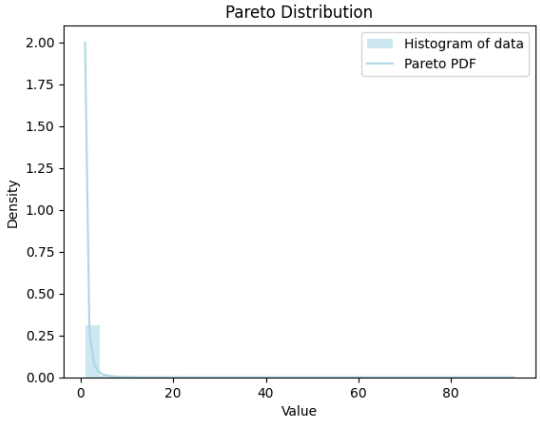



Figure 2.3: Pareto Distribution

Continuous probability distributions are essential for analyzing and understanding complex systems and processes across diverse fields, including social sciences, engineering, and natural sciences (Dinov et al., 2015). A thorough understanding of their properties and characteristics enables researchers and practitioners to develop more precise models, enhance predictive accuracy, and gain deeper insights into the underlying phenomena they seek to study.

### 2.3.8 Probability Density Estimation

Probability density estimation is a crucial technique in data analysis, helping researchers identify the underlying distribution of a dataset. A commonly used tool for this purpose is the histogram, as shown in Figure 2.5, which effectively reveals the number of peaks or modes in a probability density function.

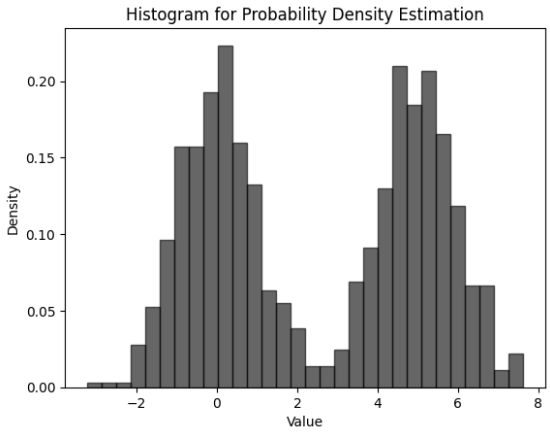



Figure 2.4: Histogram for probability density estimation

In probability density estimation, the objective is to approximate the true probability density function of a random variable using a finite set of observed data points. Histograms are a widely used approach, offering a visual representation of the data's distribution and highlighting key characteristics, such as the presence of a single peak (unimodal distribution), two peaks, or multiple peaks (Mishra & Datta-Gupta, 2018).

### 2.3.9 Unimodal Distributions

For a unimodal distribution, the histogram presents a single dominant peak, signifying that the data is concentrated around one central value. This pattern is commonly linked to the normal or Gaussian distribution, which exhibits a symmetric, bell-shaped curve (Nuzzo, 2019).

### 2.3.10 Bimodal Distributions

A bimodal distribution, unlike a unimodal one, features two distinct peaks in its histogram, as illustrated in Figure 2.6. This pattern indicates that the dataset consists of two subpopulations or modes, potentially signifying the influence of separate underlying processes or phenomena (Sainani, 2012).

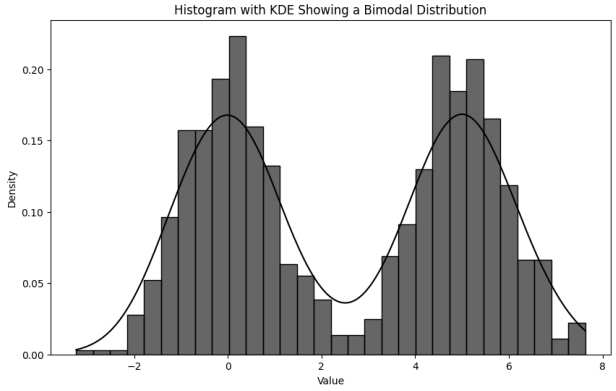



Figure 2.5: Histogram showing Bimodal Distribution

### 2.3.11 Multimodal Distributions

A distribution is classified as multimodal when its histogram displays more than two distinct peaks. For instance, Figure 2.7 illustrates a case with three peaks. Such patterns often emerge from a combination of multiple underlying distributions or the presence of several distinct subpopulations within the dataset (Nadarajah, 2008; Muñoz, 2014). Recognizing the number of

peaks in a probability density function is essential for selecting suitable statistical models and analytical techniques. Proper identification of the distribution aids in making informed choices regarding data transformation, hypothesis testing, and result interpretation.

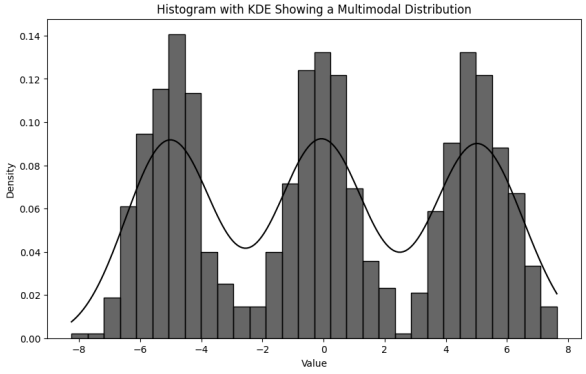



Figure 2.6: Histogram showing Multimodal Distribution

### 2.3.12 Nonparametric Density Estimation: Kernel Density Estimation

Nonparametric density estimation is an essential statistical approach that enables researchers to analyze and model the probability distribution of a continuous random variable without imposing strict assumptions about its functional form. One of the most widely applied techniques in this category is Kernel Density Estimation (KDE), which generates a smooth and continuous approximation of the probability density function using a limited set of observed data points.

Kernel Density Estimation operates by applying a kernel function to each data point, which assigns varying weights to neighboring values based on their distance from the point of interest. Data points closer to the estimation location receive higher weights, while those farther away contribute less, effectively smoothing the density estimate. The degree of smoothing is controlled by the bandwidth parameter, where a larger bandwidth produces a more generalized and less variable density estimate, while a smaller bandwidth preserves finer details but may introduce higher variability.

Due to its flexibility and capacity to model complex distributions without restrictive assumptions, Kernel Density Estimation is widely used in fields such as finance, biology, and

environmental science. Its ability to generate a smooth and continuous probability distribution makes it a valuable tool for exploratory data analysis and statistical inference.

### 2.3.13 Linear Regression

Linear regression is a core statistical method used to model the relationship between a dependent variable and one or more independent variables. It operates by fitting a linear equation to observed data, where the dependent variable is expressed as a function of the independent variables. A widely used technique for estimating the parameters of a linear regression model is Maximum Likelihood Estimation (MLE). This approach seeks to determine parameter values that maximize the probability of observing the given dataset, ensuring the best fit to the underlying data distribution.

The general form of a linear regression model with multiple independent variables can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \qquad \text{Equation 2.3}$$

Where y is the dependent variable, x1, x2, ..., xn are the independent variables, beta0, beta1, ..., beta n are the regression coefficients, and epsilon is the random error term. The method of maximum likelihood estimation seeks to find the values of the regression coefficients beta0, beta1, ..., beta_n that maximize the probability of observing the given data.

## 2.4 Derivatives Pricing Theory

The theory of derivatives pricing plays a critical role in modern finance, with applications extending across various sectors, including agriculture (Chen et al., 2021). Derivatives are financial instruments whose value is derived from an underlying asset, such as commodities, equities, or interest rates (Campbell & Diebold, 2005). Their primary function is to facilitate risk management, allowing market participants to hedge against uncertainty and potentially enhance returns (Campbell & Diebold, 2005). Given the agricultural sector's susceptibility to price fluctuations caused by factors like weather conditions and global supply and demand dynamics, derivatives can offer significant benefits (Erhardt & Smith, 2014).

Kenyan coffee farmers, who frequently experience volatile coffee prices, can utilize financial derivatives such as futures and options to protect against unfavorable market movements and ensure more stable earnings (Manfredo & Richards, 2009). However, the effectiveness of derivatives in risk management is often complicated by the complexity of pricing models, which can pose challenges for some participants (Gyamerah et al., 2019). These pricing models often rely on intricate mathematical frameworks, making them less accessible to individuals without specialized financial expertise (Campbell & Diebold, 2005).

Advancements in weather modeling and the integration of machine learning techniques present new opportunities to enhance derivatives pricing, particularly in agriculture (Gyamerah et al., 2019). By incorporating weather variables and leveraging data-driven methodologies, these innovations aim to refine pricing accuracy and support more informed hedging strategies for farmers and agricultural stakeholders (Gyamerah et al., 2019). Among various derivative products, futures contracts play a particularly vital role in managing price risks for agricultural commodities like coffee. This section explores the core principles of derivatives, focusing on the mechanics and pricing of futures contracts to illustrate their function in mitigating price volatility.

### 2.4.1 Futures Contracts

A futures contract is a legally binding agreement between two parties to buy or sell a specified quantity of an underlying asset—such as coffee beans—at a predetermined price, known as the futures price. This agreement has a fixed expiration date, at which the transaction must be executed. Unlike options, which provide the right but not the obligation to buy or sell, futures contracts impose a binding commitment on both parties, offering a level of certainty in unpredictable markets (Hull, 2001).

To maintain market transparency and liquidity, futures contracts traded on exchanges follow strict standardization. The contract specifies essential details such as the asset type, quantity, quality, delivery location, and expiration date, eliminating ambiguity (Hull, 2001). Additionally, to reduce counterparty risk, both buyers and sellers are required to deposit a margin—a portion of the contract's value—through a clearinghouse. This margin acts as collateral and undergoes daily adjustments based on market price fluctuations (Kim et al., 2022).

Upon reaching the expiration date, futures contracts can be settled in one of two ways. The first method, physical delivery, requires the seller to transfer the asset to the buyer at the agreed-upon terms (Hull, 2001). However, most financial futures opt for cash settlement, where the

contract is settled by paying or receiving the difference between the initial futures price and the price at expiration (Hull, 2001).

The pricing of futures contracts is influenced by multiple factors. While the current spot price of the underlying asset plays a central role, expectations regarding future supply and demand, storage costs, and interest rates also impact pricing (Ameur et al., 2021). The theory of storage suggests that futures prices reflect the spot price plus carrying costs, including storage and financing expenses (Ameur et al., 2021). Disparities between cash and futures prices may arise due to transaction costs, market frictions, or constraints on physical storage (Aramonte & Todorov, 2021).

Market volatility in futures trading is also shaped by elements such as time to maturity and trading volume. As a contract approaches expiration, its price movements often become more volatile, reflecting increased sensitivity to new information about the underlying asset (Matsui et al., 2022). Trading volume serves as a key indicator of market activity and liquidity, with higher volume generally leading to more stable pricing and efficient price discovery (Matsui et al., 2022).

### 2.4.2 Options Contracts

While futures contracts play a crucial role in managing commodity price risks, options provide an alternative set of financial instruments with distinct advantages. Understanding the fundamentals of options is essential for effectively navigating the complex world of financial derivatives (Gay & Hull, 1990). Unlike futures, options contracts grant the holder the right—but not the obligation—to buy or sell an underlying asset at a predetermined price within a specified period (Gay & Hull, 1990). This key distinction makes options a highly versatile tool for both hedging and speculative strategies, as they allow market participants to adjust their risk exposure while maintaining flexibility (Sánchez-Verdasco, 2018).

Options are classified into two main types: call options and put options (Gay & Hull, 1990). A call option provides the holder with the right to buy the underlying asset at a specific price, known as the strike price, on or before the expiration date (Gay & Hull, 1990). Conversely, a put option grants the right to sell the underlying asset at the strike price within the contract period (Gay & Hull, 1990). The price of an option, known as the premium, reflects its intrinsic value and is influenced by factors such as the asset's market price, volatility, time to expiration, and prevailing interest rates (Gay & Hull, 1990).

#### 2.4.2.1 Options Pricing Models

Valuing options is a complex process that relies on various mathematical techniques to determine their fair price (Sánchez-Verdasco, 2018). One of the most widely used models for option pricing is the Black-Scholes model (Tong & Reuer, 2007). While this model provides a foundational framework for understanding the factors that influence option prices, its assumptions may not always hold in real-world markets (Tong & Reuer, 2007).

#### 2.4.2.2 Options for Hedging

While our primary focus is on futures contracts, options also play a crucial role in hedging strategies. A coffee producer concerned about potential price declines can purchase put options on coffee futures, providing downside protection and limiting losses in unfavorable market conditions. Understanding the fundamentals of options contracts, their pricing mechanisms, and their role in hedging offers a more comprehensive view of risk management tools in the volatile commodities market.

### 2.4.3 Risk and Expected Returns

While derivatives provide powerful tools for managing risk, understanding their inherent risks and their impact on expected returns is crucial. Damodaran (2012) defines risk as "the likelihood that in life's games of chance, we will receive an outcome that we will not like." This definition, though simple, underscores the uncertainty inherent in any investment, including coffee farming. For example, the risk of driving too fast could result in a speeding ticket or, worse, an accident. Risk is not just about potential negative outcomes but also the probability of those outcomes occurring.

Investment risks vary widely, ranging from firm-specific risks—such as the performance of individual projects—to broader market risks like exchange rate fluctuations and political instability. Understanding this spectrum of risks highlights the complexity of risk management and the need to consider multiple factors when evaluating potential investment outcomes.

(Damodaran A. , 2012)

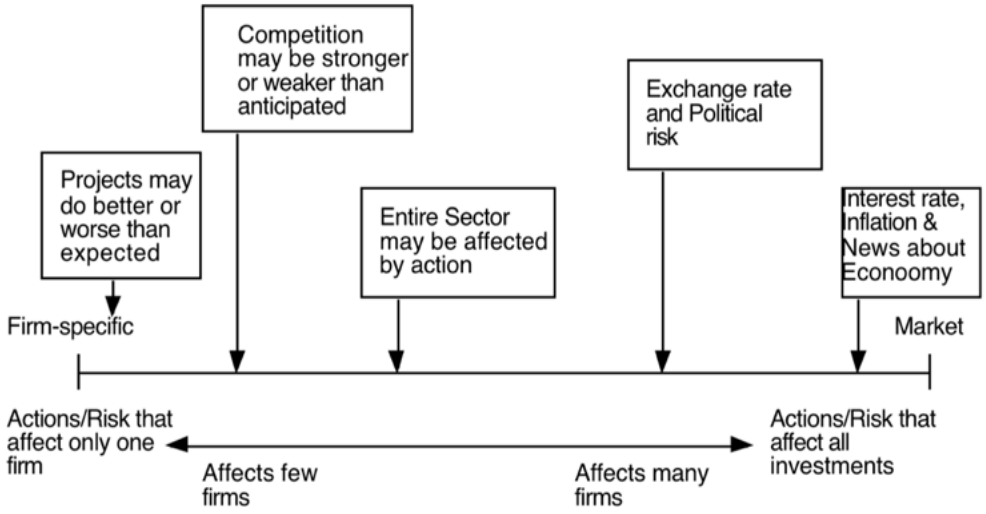



Figure 2.7: A breakdown of risk

In the context of coffee farming, various risks can impact a farmer's income:

i. **Price Risk:** Fluctuations in global coffee prices due to supply and demand changes.
ii. **Weather Risk:** Droughts, floods, or pests can severely impact crop yields.
iii. **Political Risk:** Government policies or instability can disrupt coffee production or trade.

One framework for understanding the relationship between risk and return, particularly relevant in financial markets, is the Capital Asset Pricing Model (CAPM).

### 2.4.4 The Capital Asset Pricing Model

The CAPM, a cornerstone of modern finance theory, provides a model for determining the expected return on an asset based on its risk. This model posits that investors expect to be compensated for taking on risk, and this compensation is reflected in the expected return of an asset. The higher the risk, the higher the expected return. Central to CAPM is the concept of Beta (β), a measure of a stock's volatility in relation to the overall market. A Beta of 1 indicates that the asset's price will move in the same direction as the market, while a Beta greater than 1 suggests higher volatility, and a Beta less than 1 implies lower volatility.

While coffee itself isn't a stock, the principles of CAPM can be applied to understand the risk and return dynamics of coffee futures contracts. A higher Beta for coffee futures would

indicate greater price volatility, making a case for hedging to mitigate potential losses from price swings. The CAPM equation expresses this relationship mathematically:

$$E(R_i) = R_f + \beta_i[E(R_m) - R_f] \qquad \text{Equation 2.4}$$

Where:

$E(R_i)$: $Expected\ return\ on\ asset\ i$

$R_f$: $Risk - free\ rate\ of\ return$

$\beta_i$: $Beta\ of\ asset\ i$

$E(R_m)$: $Expected\ return\ on\ the\ market\ portfolio$

Essentially, CAPM proposes that an asset's anticipated return consists of the risk-free rate plus a risk premium, which depends on the asset's Beta and the market risk premium (the gap between the market's expected return and the risk-free rate). Although CAPM does not directly determine the pricing of derivatives such as futures contracts, it serves as a useful model for analyzing the relationship between risk and return in financial markets. This perspective is particularly important when evaluating hedging strategies, as mitigating risk through hedging can limit potential gains if market prices move favorably.

### 2.4.5 Risk Management and Hedging

Throughout this discussion, we have highlighted how derivatives, particularly futures contracts, serve as essential instruments for managing price risk, a major challenge for coffee farmers. By utilizing futures contracts, farmers can secure a predetermined price for their future harvests, thereby reducing uncertainty and shielding themselves from unfavorable price fluctuations. However, it is important to recognize that hedging with derivatives does not eliminate risk entirely—it primarily shifts it to entities better equipped or more willing to handle it.

Successful risk management requires a holistic approach that accounts for market conditions, risk tolerance, and the trade-offs associated with various hedging strategies. When used strategically and with a thorough understanding of their complexities, derivatives can be effective

tools for mitigating price volatility in the coffee industry. By carefully evaluating these factors, coffee producers can make well-informed decisions that strengthen their ability to navigate the ever-changing market landscape.

## 2.5 Graphical Summary of the Theoretical Framework

The graphical summary (Figure 2.9) visually represents the integration of key concepts from probability theory, derivatives pricing theory, and artificial intelligence. This framework highlights the flow of knowledge from fundamental probability concepts to the complexities of derivatives pricing and the application of AI in financial markets.

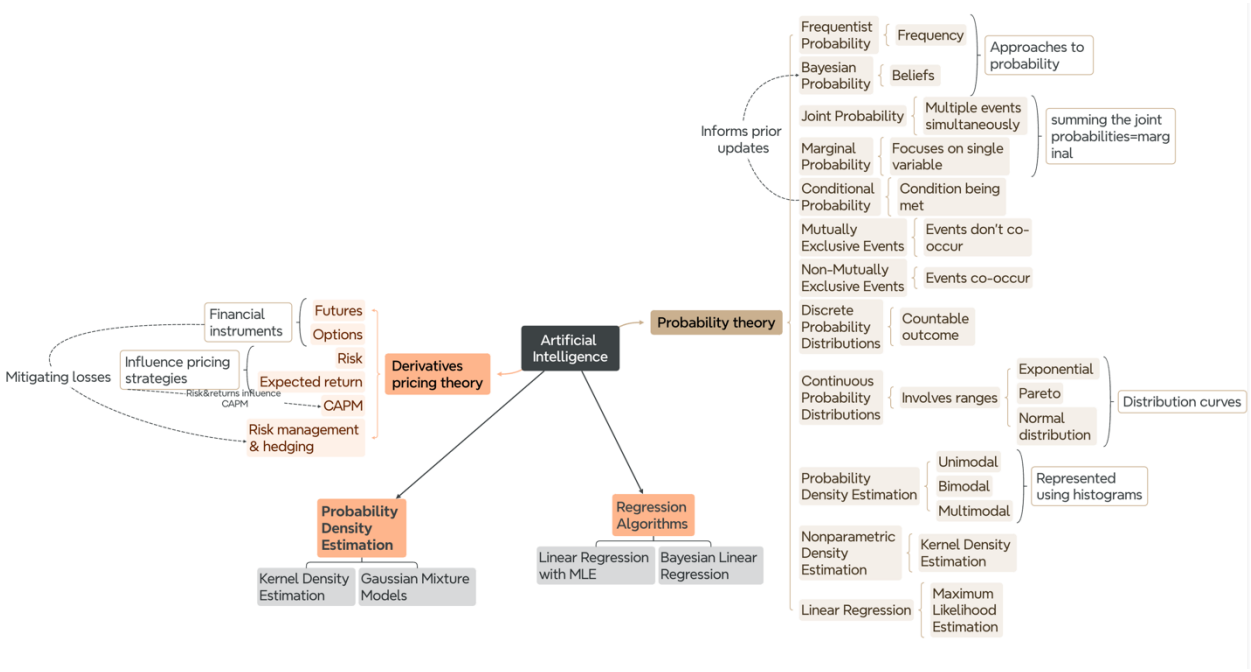



Figure 2.8: Graphical Summary of the Theoretical Framework

## 2.6 Existing Algorithms

Predictive modeling provides a robust framework for analyzing complex patterns and guiding decision-making (Zambom & Dias, 2012). In the context of stabilizing the income of Kenyan coffee farmers, these models hold the potential to uncover valuable insights and inform strategic interventions. This section presents a concise overview of commonly used algorithms in predictive modeling and data analysis. These algorithms, broadly classified into regression models and probability density estimation techniques, offer a foundational understanding and serve as a

basis for developing innovative approaches tailored to enhancing income stability for Kenyan coffee farmers.

### 2.6.1 Regression Algorithms

Regression algorithms play a fundamental role in predictive modeling by quantifying relationships between a dependent variable and one or more independent variables. These techniques are particularly useful for assessing how variations in input factors impact the target variable, facilitating accurate predictions and informed decision-making (Stulp & Fauré, 2015). This section explores two widely used regression methods: Linear Regression with Maximum Likelihood Estimation and Bayesian Linear Regression. Each approach provides unique advantages in modeling and analyzing data relationships, contributing to a deeper understanding of complex patterns.

#### 2.6.1.1 Linear Regression with Maximum Likelihood Estimation

Linear Regression with Maximum Likelihood Estimation (MLE) is a fundamental statistical learning technique used to model the relationship between a dependent variable and one or more independent variables. This method assumes that the relationship can be approximated by a linear function and aims to find the model parameters that maximize the likelihood of observing the given data. In essence, MLE seeks the parameter values that make the observed data most probable.

(See algorithm 2.1 for an illustration of Linear Regression with Maximum Likelihood Estimation).

**Algorithm 1** Linear Regression with Maximum Likelihood Estimation

**Input :** Dataset $X = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where $x_i \in R^d$ (features) and $y_i \in R$ (target)

**Output:** Model parameters $\theta = (\beta, \sigma^2)$

**Function** *LinearRegressionMLE(X)*

  Initialize $\beta$ randomly  Initialize $\sigma^2$ randomly or with a guess

  **while** *not converged* **do**

    /* Calculate predictions */

    $\hat{y} \leftarrow X\beta$

    /* Calculate log-likelihood */

    $L(\theta) \leftarrow -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

    /* Update parameters using gradient ascent */

    $\beta \leftarrow \beta + \alpha\frac{\partial L(\theta)}{\partial \beta}$

    /* Update $\sigma^2$ (optional, depending on approach) */

    $\sigma^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

  **return** $\theta = (\beta, \sigma^2)$

Algorithm 2.1: Linear Regression with Maximum Likelihood Estimation

#### 2.6.1.2 Bayesian Linear Regression

Bayesian Linear Regression takes a probabilistic approach to modeling relationships between variables. Unlike traditional methods that identify a single optimal estimate for model parameters, this technique integrates prior knowledge and refines it using observed data. The result is a posterior distribution of parameters, which captures the uncertainty in the estimates and provides a more nuanced understanding of the model's behavior.

(see Algorithm 2.2 for a visualization of Bayesian Linear Regression).

**Algorithm 2** Bayesian Linear Regression

**Input :** Dataset $\{(x_i, y_i)\}_{i=1}^{N}$, prior parameters $\mathbf{\Lambda}_0$, $\sigma^2$
**Output:** Posterior parameters $\mu_N$, $\mathbf{\Lambda}_N$

**Function** *BayesianLinearRegression*$(\mathbf{X}, \mathbf{y}, \mathbf{\Lambda}_0, \sigma^2)$
  /* Compute the posterior precision matrix */
  $\mathbf{\Lambda}_N \leftarrow \mathbf{\Lambda}_0 + \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$
  /* Compute the posterior mean */
  $\mu_N \leftarrow \mathbf{\Lambda}_N^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}\right)$
  **return** $\mu_N, \mathbf{\Lambda}_N$

Algorithm 2.2: Bayesian Linear Regression Algorithm

#### 2.6.1.3 Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric technique for approximating the probability density function of a random variable. Unlike parametric approaches that rely on predefined functional forms, such as the Gaussian distribution, KDE remains flexible by not imposing any specific shape on the data distribution (Dias & Zambom, 2012). It works by positioning a smooth kernel function at each data point and aggregating these functions to construct an estimate of the overall density. The degree of smoothness in the density estimate is influenced by the choice of kernel function and the bandwidth parameter (Silverman, 2018). KDE is extensively used in exploratory data analysis, visualization, and as a foundational tool for various statistical applications.

**Algorithm 3** Kernel Density Estimation

**Input :** Data points $\{x_i\}_{i=1}^{N}$, bandwidth $h$, kernel function $K$
**Output:** Estimated density function $\hat{f}(x)$

**Function** *KernelDensityEstimation*($\{x_i\}_{i=1}^{N}, h, K$)
    /* Initialize the density function */
    $\hat{f}(x) \leftarrow 0$
    /* Sum the contributions of each data point */
    **for** $i \leftarrow 1$ **to** $N$ **do**
        $\hat{f}(x) \leftarrow \hat{f}(x) + \frac{1}{Nh} K\left(\frac{x - x_i}{h}\right)$
    **return** $\hat{f}(x)$

Algorithm 2.3: Kernel Density Estimation Algorithm

## 2.7 Models and Frameworks

Predicting coffee prices involves a multidimensional approach that integrates various models and frameworks to account for the multiple factors influencing the coffee market.

For Kenyan coffee a predictive model needs to assess past price patterns plus macroeconomic factors along with weather reports as well as global trade flows. This section outlines the primary models and frameworks that can be employed to predict Kenyan coffee prices, providing farmers with the tools necessary to make informed hedging decisions.

### 2.7.1 Models

#### 2.7.1.1 Time Series Analysis

Time series analysis plays a crucial role in forecasting coffee prices by leveraging historical data to identify trends, seasonal patterns, and underlying fluctuations. This approach operates on the assumption that past price movements provide valuable insights into future price behavior. By capturing the temporal relationships between observations, time series models can effectively reflect market dynamics. Given the cyclical nature of agricultural markets, these methods are particularly well-suited for predicting coffee price variations over time.

##### 2.7.1.1.1 Autoregressive Integrated Moving Average

Time series forecasting has become widely used across the globe, with ARIMA models standing out as one of the most effective statistical techniques. This model analyzes past observations and trends while accounting for short-term variations to uncover the underlying structure of price data (Boyeena & Kumar, 2024). ARIMA consists of three key components—autoregressive (AR), integrated (I), and moving average (MA)—which allow it to handle various types of time series data, including those with seasonal trends, irregular fluctuations, or non-stationary behavior. When applied to coffee price forecasting, ARIMA helps capture historical price correlations, making it suitable for short-term predictions. It also models different factors

contributing to price volatility, such as periodic price cycles and sudden market shocks (Box, Jenkins & Reinsel, 2015).

#### 2.7.1.1.2 Seasonal Decomposition of Time Series

Seasonal decomposition of time series (STL) is a technique that breaks down time series data into three key components: trend, seasonality, and residual variation. This method is particularly beneficial for analyzing coffee prices, as it helps isolate the effects of factors such as harvest cycles, global demand, and climatic conditions (Wen, Zhang, Li, & Sun, 2020). By distinguishing seasonal fluctuations from broader trends, STL enhances the accuracy of coffee price forecasts (Wen et al., 2020). Additionally, its flexibility allows it to account for nonlinear seasonal effects and adapt to changing market conditions over time, making it a valuable tool in price analysis (Wen et al., 2020).

### 2.7.1.2 Machine Learning Models

With the growing availability of data and enhanced computational power, machine learning models have become indispensable for predicting coffee prices. These models can process large datasets rapidly, making them particularly useful in complex, highly non-linear scenarios where traditional methods struggle to detect hidden patterns (Smith, Johnson, & Lee, 2021). One of the key advantages of machine learning is its ability to identify relationships within data without relying on predefined assumptions, allowing for more flexible and adaptive forecasting techniques (Doe, 2022).

#### 2.7.1.2.1 Support Vector Machines

Support Vector Machines (SVM) are supervised learning models used for both classification and regression tasks. When applied to coffee price forecasting, SVM regression can estimate continuous price values based on relevant input features. These models are particularly suitable for datasets with non-linear relationships, as kernel functions transform the input data into higher-dimensional spaces, facilitating separation using hyperplanes (Doe, 2022). Given their ability to capture intricate patterns and their potential to overfit, SVMs have been successfully utilized in various financial prediction applications (Johnson, 2023).

#### 2.7.1.2.2 Random Forests

Random Forests is a supervised learning algorithm which consists of several decision trees, each providing a prediction output, which is aggregated for better performance. Random Forests are useful for improving prediction accuracy, combating overfitting, and increasing robustness by

amalgamating the outputs of a plethora of decision trees. Given the multitude of interacting factors that can influence coffee prices, this model would be invaluable since it can work with many predictor variables and capture non-linear relationships. It's been proven that the forecasting power of Random Forests supersedes the power of most other machine learning models for all types of tasks, especially when factoring in complexities like weather-related events, political changes, or altering supply chains.

#### 2.7.1.2.3 Neural Networks

Long Short-Term Memory networks are a special kind of recurrent neural network that can learn sequences and time series data. LSTMs are well-suited for problems that have time sequences, like predicting coffee prices, because they can learn long-term dependencies from the data (Fofanah, 2021). Unlike standard neural networks, LSTM memory cells can retain and recall information over extended periods, thus capturing patterns and trends spanning multiple time frames, making LSTMs ideal for price-based forecasting (Fofanah, 2021). When predicting coffee prices in volatile markets, LSTMs can adjust forecasts by learning from past price data and are therefore able to provide more accurate predictions of future prices (Fofanah, 2021).

### 2.7.1.3 Econometric Models

Models of Econometrics combine economic principles with statistics to study and predict financial and economic phenomena. With respect to coffee price forecasting, these models are especially helpful since they enable the analyst to combine both coffee market fundamentals with other exogenous variables, such as interest rates, inflation, and the rest of the world's goods prices (Salaudeen, Sathasivam, & Ishak, 2024). Econometric models are also able to show how coffee prices respond to changes in other economic variables over time (Salaudeen et al., 2024).

#### 2.7.1.3.1 Vector Autoregression

Vector Autoregressive (VAR) models are a set of statistical models that consider the relationships among several time series variables. With respect to forecasting prices of coffee, VAR can be used to specify the interactions between coffee prices and other economic phenomena like forex rates, inflation rates, or global commodity prices (Fofanah, 2021). Instead of treating these variables in isolation, VAR enables one to grasp the interaction and dynamics that different economic factors engage in, thus affording a better understanding of the coffee price dynamics (Fofanah, 2021). It also helps explain how external forces such as changes in global supply chains

or changes in patterns of consumption by the leading coffee importing nations affect local coffee prices in Kenya (Fofanah, 2021).

#### 2.7.1.3.2 Error Correction Models (ECM)

Error Correction Models (ECM) are econometric models used to model the short-term and long-term dynamics between variables. In the case of coffee prices, ECMs can help model the relationship between the short-term price fluctuations and long-term equilibrium price levels. When coffee prices deviate from their long-run equilibrium, the ECM helps estimate the speed at which the prices return to equilibrium, providing insights into the persistence of price shocks. This model is valuable for understanding both the short-term volatility and the long-term trends in the coffee market (Engle & Granger, 1987).

### 2.7.2 Frameworks

The development of models is rather complex as it involves creating frameworks for data processing, training, and evaluation. These frameworks make it possible to process massive datasets, multitudes of machine learning and econometric models, as well as the verification of the predictions made by the models in a systematic manner.

#### 2.7.2.1 Python Libraries

Data scientists prefer Python as its library collection for data analysis, machine learning, and even econometrics is vast. Tools for machine learning, like SVMs, and Random Forest algorithms are available in the scikit-learn library, while skills for deep learning applications, like LSTM networks, are supplied by TensorFlow. Moreover, stats models is a potent library for econometric models' implementation, which includes ARIMA, and VAR. Thus, it allows researchers to build, as well as test, their predictive models for coffee prices in an efficient manner (Pedregosa et al., 2011; Abadi et al., 2016).

#### 2.7.2.2 R Packages

One more programming language popular for statistical modeling is R. With respect to time series analysis, the forecast package is more useful because of its functions for ARIMA modelling, seasonal decomposition, and other forecasting methods. An equally popular package in R, called caret, makes tuning and training of machine learning models much easier by providing a standard interface.

## 2.8 Algorithms

Within the context of generating forecasting models for predicting coffee prices in Kenya, two algorithms stand out, Linear Regression and Extreme Gradient Boosting (XGBoost) algorithms. The nature of these algorithms is such that they provide both complementarity and simplicity. They allow the modelling of intricate relationships within processes while allowing for easy understanding. Their use within the study bolsters the ability to significantly and positively impact the precise forecasts that these farmers rely on for their decisions.

### 2.8.1 Linear Regression

A traditional predictive analytical algorithm that stands out, and is still widely used, is Linear Regression, which is popular because of its straightforwardness. It presents a straight-line equation and also relies heavily on the relationship between one dependent variable, such as coffee prices, and one or more independent variables, including weather, production, and even global market activity (Fofanah, 2021). From the approach used by this algorithm, the coffee market is easier to understand than any other. A combination of rainfall, historical prices, and even exchange rates can all be used. Attempts to apply these variables in understanding the market can result in more accurate predictions As demonstrated, it is possible to make educated guesses on the relationships that are bound to exist between the variables and the market. Such predictions are made on the assumption that once the combination of factors is established, a clear outcome is bound to be set. The wide acceptance of using linear regression is sometimes misplaced due to its reliance on assumptions that the relied variable is linear. This creates a limit on the algorithm as it proves to be ineffective in complex markets such as the coffee market, which is characterized by trade volatility and an ever-changing structure. Additionally, even with these limitations, it can still be an effective initial model to capture basic patterns and help build more sophisticated models.

### 2.8.2 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is one of the fastest, most flexible, and powerful machine learning models known for great work in predictive modeling due to its efficacy in complex datasets. Instead of linear regression where many times it would be inefficient, XGBoost would excel in modeling nonlinear trends and interactions in data. An XGBoost model is built using an ensemble of decision trees where each tree in the ensemble is trained to fix the errors of the previous one. This process of boosting increases the generalization of the model especially in volatile markets, such as coffee trading, where prices are influenced by several factors including trade volume, currency value, and climate. XGBoost also makes it possible to determine the

relative importance of different features, which helps researchers gauge the effect of inflation, transportation costs, and weather forecasts. Furthermore, its speed in computation makes it a good choice among the rest when dealing with big data. Nonetheless, XGBoost has its own complexities. For one, because it is so complex, tuning the hyperparameters might be very needed. Also, it is tougher to interpret as opposed to other algorithms such as linear regression.

## 2.9 Conceptual Framework

This conceptual framework, illustrated in Figure 2.9, outlines a proposed pathway to stabilizing Kenyan coffee farmers' income. These forecasts can then inform farmers' decisions regarding hedging strategies using futures contracts, allowing them to mitigate price risk and stabilize income. The framework acknowledges the influence of potential mediating factors, such as farmers' knowledge of hedging strategies, which could impact their ability to effectively utilize the price forecasts.

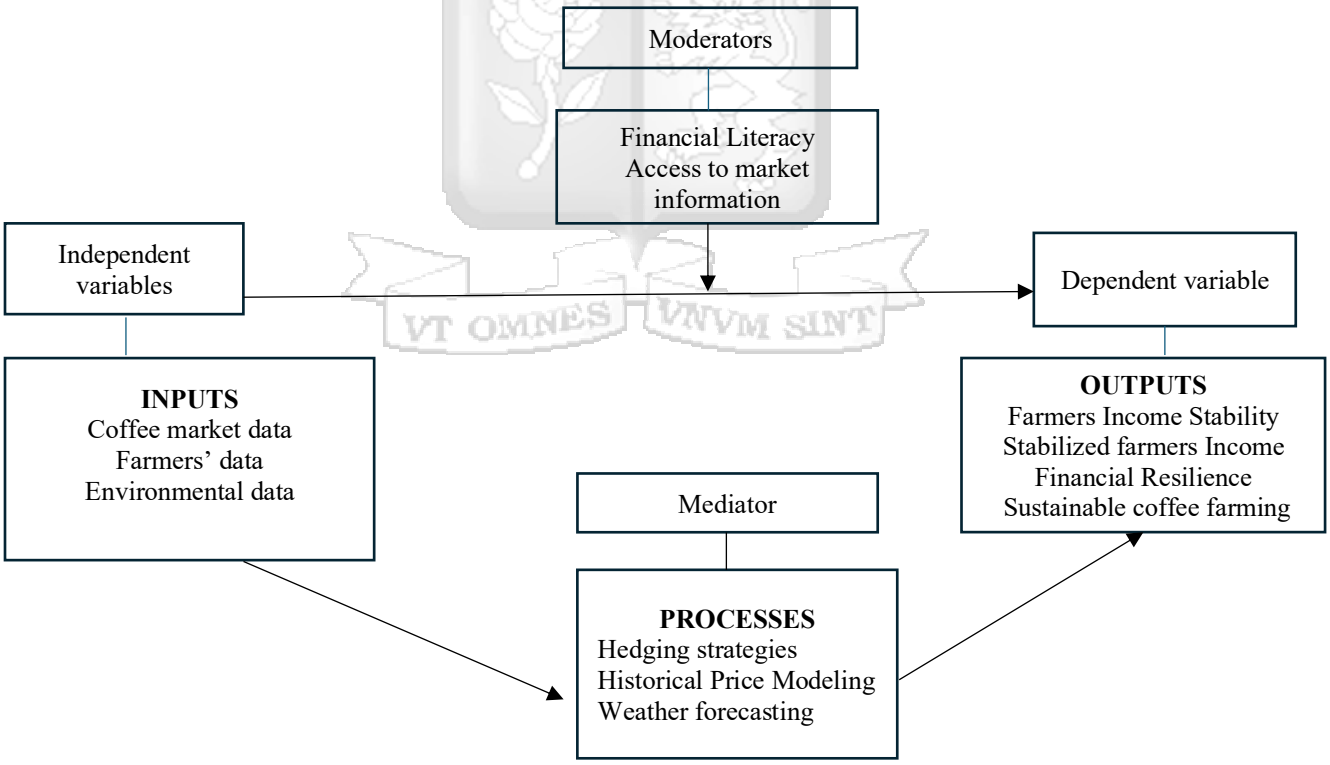



Figure 2.9: Conceptual Framework

## 2.10 Research Gap

Although there is a vast literature on the use of predictive models and machine learning algorithms in financial time series prediction and commodity price forecasting, there is a dearth of such studies in the context of agriculture in developing countries such as Kenya. For example, Bilokon and Qiu (2023) compared the performance of LSTM and Transformer-based models for electronic trading but did not generalize their study to agricultural commodities, especially where price fluctuations affect smallholder farmers. In the same vein, Ghojogh and Ghodsi (2023) presented a systematic review of RNNs and LSTMs, nevertheless, their work did not focus on specific sectors, such as models tailored to the coffee industry. Due to the fluctuating nature of coffee prices and the fact that it is a key source of income for many farmers in Kenya, there is a need to develop solutions that can help in predicting the prices in this sector.

Furthermore, Ly et al. (2021) have investigated the performance of machine learning techniques in comparison to traditional statistical models for commodity price forecasting. Although their study provided evidence on the usefulness of machine learning techniques, it was more applicable to large commodity markets and did not consider regional factors such as climate change, social issues, and financial challenges that affect smallholder farmers in Kenya. These factors are important in determining the price of coffee and should be incorporated into the models for price forecasting and risk management.

These studies underscore the potential of advanced predictive techniques but reveal an absence of solutions addressing the unique challenges faced by coffee farmers in Kenya. To bridge this gap, the current study will develop a region-specific predictive model for hedging futures contracts. the model incorporates key local factors such as historical price patterns, macroeconomic indicators, and climate conditions to improve accuracy and applicability, aiming to stabilize incomes and enhance economic resilience.

# Chapter 3: Research Methodology

## 3.1 Introduction

This research methodology involved collecting and analyzing historical coffee price data to develop a predictive model using time-series forecasting techniques. The study adopted a quantitative approach to ensure that the research outputs were objective, measurable, and replicable. This chapter provides a detailed discussion of the research design, outlines the data collection process, describes the methods used for data analysis, and explains the approach taken in developing the predictive model.

Historical pricing data was collected and analyzed to identify trends and seasonal patterns, forming the foundation for model development. Various forecasting algorithms were evaluated and fine-tuned to improve predictive performance. By adhering to a rigorous, data-driven research methodology, the study contributes to the field of agricultural price forecasting and offers practical insights for stakeholders in the Kenyan coffee sector.

## 3.2 Research Design and Philosophy

This study adopts a positivist research philosophy, which emphasizes objectivity, measurement, and the use of quantifiable data to generate knowledge. Given the technical and data-driven nature of the research, a quantitative approach was most appropriate. This approach facilitated the analysis of historical coffee price trends and enabled the development of a robust predictive model. Historical data was focusing on key variables influencing coffee prices. Various forecasting techniques were applied, including regression models and time-series approaches, to determine the most effective algorithm for accurate price prediction.

## 3.3 System Architecture and Proposed Modules

The system consists of several modules that together form the foundation for the predictive model. These modules are designed to process incoming data, apply machine learning algorithms, and provide actionable insights for coffee farmers. The system consists of the following modules.

i. **Data Ingestion Module**: This module is designed to gather and integrate data from various sources, including historical coffee prices and weather patterns.

ii. **Feature Engineering Module**: Following data ingestion, the system will process and convert raw data into meaningful features suitable for machine learning algorithms.

iii. **Predictive Modeling Module**: This module is the core of the system, where machine learning algorithms are applied to the processed data to develop a predictive model.

iv. **Evaluation and Reporting Module**: Once the model is trained, this module will evaluate its performance and generate reports detailing the accuracy of the predictions.

v. **User Interface Module**: This module will provide a simple interface for farmers to input their data (such as expected yield and operational costs) and receive predictions and recommendations.

Below is a high-level system architecture diagram illustrating the components and their interactions:

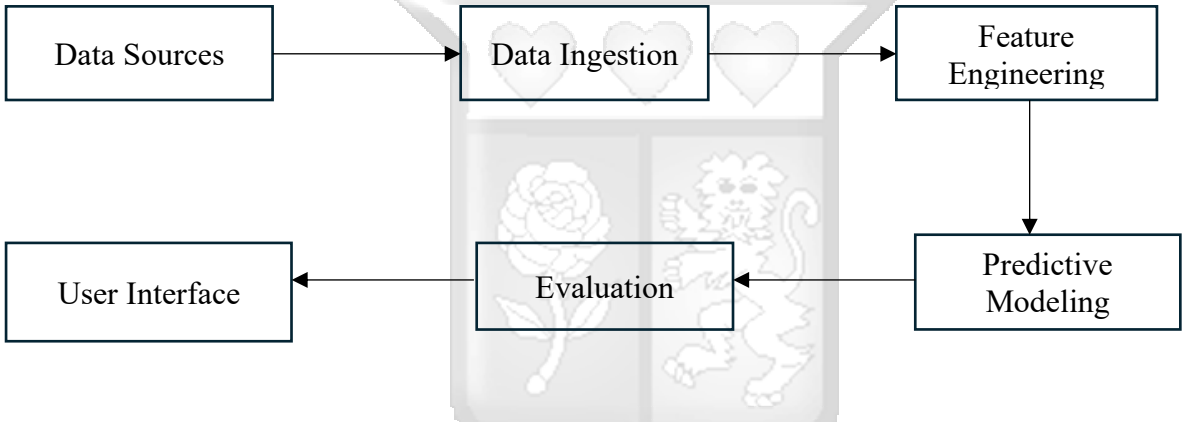



Figure 3.1: System architecture diagram

## 3.4 Population and Sampling

### 3.4.1 Target Population

The target population for this study consists of smallholder coffee farmers in Embu, Kiambu, and Kakamega counties. These regions were chosen for their substantial role in Kenya's coffee production and their varied agricultural landscapes. According to the Kenya Agricultural and Livestock Research Organization (2020), they collectively support about 88,200 smallholder coffee farmers, with Embu hosting approximately 20,000, Kiambu 56,200, and Kakamega 12,000. The total population size was estimated using secondary data from county agricultural offices and cooperative societies. These sources provided reliable figures for the number of registered smallholder coffee farmers in each county.

### 3.4.2 Sampling

This study employed a quantitative research approach, relying entirely on secondary data for analysis. Although the initial design considered a mixed-methods approach, no qualitative data collection was conducted, and thus the final methodology focused solely on quantitative techniques. The quantitative analysis utilized historical datasets obtained from the Kenya National Bureau of Statistics (KNBS). The data included key variables such as coffee prices, inflation rates, exchange rates, fuel prices, and export metrics. These datasets spanned multiple years and covered a broad range of economic indicators relevant to coffee price forecasting.

To ensure manageability and maintain representativeness across the dataset, a stratified sampling approach was applied. Stratification was based on time periods (e.g., years and months) and variable categories (e.g., price, production, and macroeconomic indicators). Within each stratum, a systematic sampling method was used, selecting every $n^{th}$ record depending on the desired sample size and total record count. This technique ensured even coverage of the data across time and reduced the risk of selection bias. Systematic sampling was deemed appropriate due to its simplicity, efficiency, and ability to ensure balanced representation across temporal and categorical dimensions of the dataset.

## 3.5 System Development Methodology

To structure the development of the prediction model, I will follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. This widely adopted approach in data science ensures a systematic and organized framework for the modelling process.

CRISP-DM consists of six key phases:

i. Business Understanding: Understanding the income volatility challenges faced by Kenyan coffee farmers and the role futures contracts can play in stabilizing income.
ii. Data Understanding: Analysing historical data on coffee prices, weather conditions, and market trends to identify key variables for the model.
iii. Data Preparation: Cleaning and transforming the data, ensuring it is suitable for modelling.
iv. Modelling: Applying machine learning algorithms to develop a predictive model that estimates income stabilization through hedging.
v. Evaluation: Evaluating the model using metrics such as R squared to ensure accuracy and reliability.

vi. Deployment: Translating the model's predictions into actionable strategies for coffee farmers.

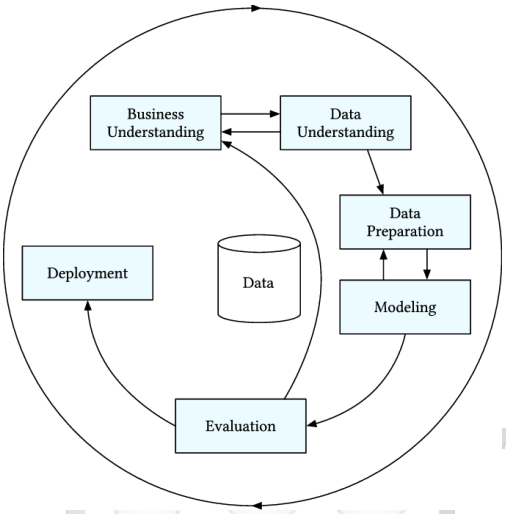



Figure 3.2: Cross-Industry Standard Process for Data Mining

## 3.6 Data Collection Methods

Both primary and secondary data collection methods will be used. For primary data interview questions, will be utilised. For secondary data, the study relied on the Kenya National Bureau of Statistics (KNBS), which provided essential information on coffee price movements and volatility across the periods under analysis. This data shall provide the platform on which to determine the level of price risk that influences farmer revenues.

Qualitative data from case studies of different Kenyan coffee farms will also complement these evaluation metrics. The case studies will provide real-life views on problems faced by farmers, the sort of data they collect, and their opinions on income volatility and futures contracts. In this light, a qualitative approach will be applied through semi-structured interviews with the coffee farmers to ascertain how they manage income fluctuations and their understanding of financial instruments, such as future contracts.

Interviews will provide in-depth insight into socio-economic factors that shape financial decision-making by farmers, as well as their acceptance of the prediction model. Besides interviews, observations on farms will record the ongoing practices in data recording, especially the data farmers keep on yields, market prices, weather conditions, and operational costs. Such

observations will assist in finding the gaps between the actual recorded data by farmers and the data required to apply the predictive model effectively.

### 3.7 Data Analysis Procedures

This study's data analysis will be conducted in two key phases: evaluating the predictive model's performance using machine learning techniques and analyzing the dataset's statistical properties. These approaches will ensure that the model is both statistically sound and aligned with real market trends. The model's accuracy will be assessed using fundamental error metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination ($R^2$). RMSE will measure the average deviation of predictions from actual values, providing an overall performance evaluation. MAE will indicate the mean magnitude of prediction errors, while $R^2$ will determine how well the model explains price variations.

These metrics will play a crucial role in evaluating different modeling techniques to determine the most effective approach. To ensure reliability, the study will employ a holdout validation strategy, where the dataset will be split into training and testing sets while maintaining chronological order. The model will be trained on historical coffee price data and tested on unseen data to measure its predictive performance. Although cross-validation methods will be considered, their use may be constrained due to the time-dependent nature of the dataset. Statistical validation will be conducted to assess the model's ability to generalize to future market conditions.

Beyond performance evaluation, descriptive statistical analysis will be utilized to examine the distribution and variability of coffee prices. Measures such as skewness, kurtosis, standard deviation, and the coefficient of variation will provide insights into price dispersion and trends. Additionally, Pearson's correlation coefficient will be applied to analyze relationships between key economic factors and coffee prices, helping to identify the drivers of price fluctuations. Unlike qualitative research, which focuses on case studies and thematic analysis, this study relies solely on quantitative data sourced from secondary materials, including auction prices, economic indicators, and market trends. This structured analytical approach ensures a thorough evaluation of the predictive model's effectiveness in forecasting coffee prices, aligning it with real-world market conditions and supporting data-driven decision-making within Kenya's coffee sector.

### 3.8 System Analysis

The system analysis phase is essential for comprehending the requirements and elements that will guide the construction of the predictive model. The dataset will be analyzed to find critical variables including coffee prices, futures contracts, meteorological patterns, and macroeconomic factors. The relationships among these variables will be evaluated to ascertain their impact on the income volatility of coffee farmers. The investigation will include evaluating the historical efficacy of futures contracts in alleviating income risks, discerning trends, and comprehending market cycles. An essential aspect of system analysis entails determining the functional needs of the prediction model. This includes being able to precisely predict coffee prices and model the impacts of various futures contract methods. The model will be scalable to accommodate new data or fluctuations in market conditions.

Non-functional requirements include efficiency, robustness, and usefulness of the model, facilitating informed decision-making for farmers utilizing the system. Ultimately, in the system analysis phase, use-case scenarios will be generated. These scenarios will outline how farmers will interact with the system and receiving predictions on optimal futures contract strategies.

### 3.9 System Testing

The developed model will be subjected to functionality and performance testing to verify that it meets specified requirements and delivers dependable, accurate predictions. Evaluation metrics such as Root Mean Squared Error (RMSE) will be used to assess prediction accuracy, while the coefficient of determination ($R^2$) will indicate how effectively the model captures price variability. Additionally, Mean Absolute Error (MAE) will serve as a direct measure of the average prediction error. By analyzing these metrics across various algorithms, the model's predictive capabilities will be refined and optimized.

Beyond performance testing, stress testing will be conducted to evaluate the system's ability to process large datasets without significant performance degradation. This is crucial for scalability, as new data on coffee prices, futures contracts, and macroeconomic conditions continuously emerge. Additionally, the model's accuracy will be assessed under various market conditions, including price surges and declines, ensuring robustness.

Finally, user acceptance testing (UAT) will be performed to validate the system's practicality for Kenyan coffee farmers. This phase will involve real-world testing, where selected farmers will input data and evaluate the system's usability, functionality, and predictive accuracy.

### 3.10 System Validation

The resulting model will undergo functional and performance testing to ensure it meets the specified objectives and provides reliable, accurate forecasts. Key evaluation metrics will include Root Mean Squared Error (RMSE) to quantify prediction errors, the coefficient of determination ($R^2$) to assess how well the model explains price variability, and Mean Absolute Error (MAE) to measure average forecast error. Comparing these metrics across multiple algorithms will help optimize predictive performance.

In addition to performance testing, stress testing will assess the system's ability to handle large datasets without significant degradation. This is essential for scalability, particularly as new data on coffee prices, futures contracts, and macroeconomic factors become available. The model's accuracy will also be evaluated across different market conditions, including price surges and downturns. Finally, user acceptance testing (UAT) will ensure the system meets the practical needs of Kenyan coffee producers.

### 3.11 Utilization of Results

Coffee farmers, cooperatives, traders, and market analysts will benefit most from the model's insights and predictions. Its primary purpose is to support informed decision-making. Farmers can use price forecasts to choose optimal hedging strategies, reducing income volatility from market fluctuations and improving financial stability.

Similarly, cooperatives and traders can leverage the model to better understand price trends, optimize supply chains, and mitigate market risks. Beyond coffee, the model provides a framework for analyzing price fluctuations in other agricultural commodities, contributing to global discussions on market stability and farmer income security.

### 3.12 Dissemination of Results

The research findings will primarily be disseminated through academic channels. The final thesis will be submitted to the university library, ensuring accessibility for future researchers and students. Additionally, presenting the study at university research seminars or graduate colloquiums will enhance its visibility within the academic community.

While workshops and training sessions for coffee farmers or cooperatives are beyond the research's direct scope, findings may be shared informally through collaborations with agricultural extension officers or summarized reports for interested cooperatives. Key stakeholders, such as

cooperative managers and local agricultural advisors, may receive a concise document or presentation highlighting actionable insights.

The dissemination strategy will prioritize both academic contribution and practical engagement, ensuring the research findings are accessible and beneficial to those involved in the Kenyan coffee sector.

### 3.13 Ethical Considerations

Ethical considerations play a crucial role in this study, particularly due to the involvement of Kenyan coffee farmers and the handling of sensitive financial data. Prior to data collection, informed consent will be obtained, ensuring participants fully understand the study's objectives, procedures, and their right to withdraw at any stage without consequences. Confidentiality will be strictly upheld by anonymizing personal identifiers and employing secure storage methods such as password-protected files and encryption to safeguard digital data.

The study will adhere to the principle of non-maleficence, ensuring that no harm comes to participants while respecting their autonomy. Farmers will have the freedom to withhold information, and participation will be entirely voluntary, free from coercion. Cultural sensitivity will be maintained throughout, ensuring interactions are respectful and non-exploitative. Ethical approval will be sought from the relevant review board to ensure compliance with institutional guidelines. Additionally, participants will have the opportunity to provide feedback and will be informed about how the research findings will be utilized, with an option to receive a summary of the results to promote transparency.

# Chapter 4: System Analysis and Design

## 4.1 Introduction

The coffee price forecasting model was created by comparing various times series coffee price forecasting techniques and machine learning algorithms to determine the most accurate predictor. A use case and a system sequence diagram were created to illustrate the interactions between the different parts of the system. In the design process of the system, functional and non-functional requirements were collected which were used to create the system design and architecture.

## 4.2 System Analysis

System analysis focuses on defining the system's characteristics, functional expectations, and operational constraints. This study aims to develop a predictive model for forecasting future Kenyan coffee auction prices based on historical data and economic indicators. This section presents the system's functional and non-functional requirements, derived from an in-depth analysis of data processing needs and user expectations.

## 4.3 Requirement Gathering

The process of gathering requirements involved an extensive review of relevant literature, an in-depth analysis of historical coffee price data, and an assessment of system capabilities aligned with the research objectives. The primary dataset for this study consists of monthly coffee prices from January 2019 to September 2024, alongside key macroeconomic variables such as inflation rates, exchange rates, fuel prices, export data, and weather conditions. This data was instrumental in understanding market trends, determining relevant inputs for the forecasting model, and establishing evaluation criteria. Additionally, input from coffee farmers was incorporated to enhance the system's practical usability and ensure its relevance to real-world applications.

## 4.4 Functional Requirements

Functional requirements define the core functionalities that the system must implement. The coffee price forecasting system is designed for two main types of users: general users like farmers and traders, who can access forecasts and analyze price trends and administrators who manage system data, oversee model execution, and maintain system performance.

The identified functional requirements for the system include:

i. Users must be able to register and log in securely to access system features.

ii. Administrators must have privileged access to manage system resources and user accounts.
iii. The system must support historical price data storage and allow updates when new auction data becomes available.
iv. The system must support coffee price forecasting by integrating a prediction model, with flexibility for testing and optimization during development.
v. Users must be able to view historical price trends and future predictions through graphical representations, including line and bar charts.
vi. The system must support interactive filtering based on time periods and economic indicators.
vii. Users must be able to generate price forecast reports with statistical summaries and export them in PDF or CSV format.
viii. Administrators must be able to manage system data, configure model execution settings, and control user access.

### 4.4.1 Non-Functional Requirements

Non-functional requirements define the performance, security, and operational constraints of the system. These are listed in table 2 below:

Table 4.1: Non-Functional Requirements

| **Requirement** | **Description** |
|---|---|
| Availability | The system must be available 99.9% of the time for users to access forecasts. |
| Reliability | The system should consistently provide accurate predictions with minimal downtime. |
| Performance | Forecast generation should be efficient, with results displayed within 10 seconds. |

| | |
|---|---|
| Usability | The system should have a user-friendly interface, accessible even to non-technical users. |
| Security | User data and price forecasts should be secured using encryption and authentication mechanisms. |
| Scalability | The system should be able to handle an increasing volume of users and expanding datasets. |
| Maintainability | The system should be modular and easy to update with new forecasting models or data. |

## 4.5 System Design

During the design phase, UML diagrams were created based on the system's functional specifications, ensuring seamless interaction between components. This section presents key diagrams, including use case, sequence, and Entity Relationship Diagrams (ERD), to illustrate system features, data processing, and user interactions.

### 4.5.1 Use Case Diagram

The use case diagram depicts the interactions and relationship between various actors and the coffee prediction system. Figure 4.1 illustrates scenarios in which external parties likes users and administrators use the system.

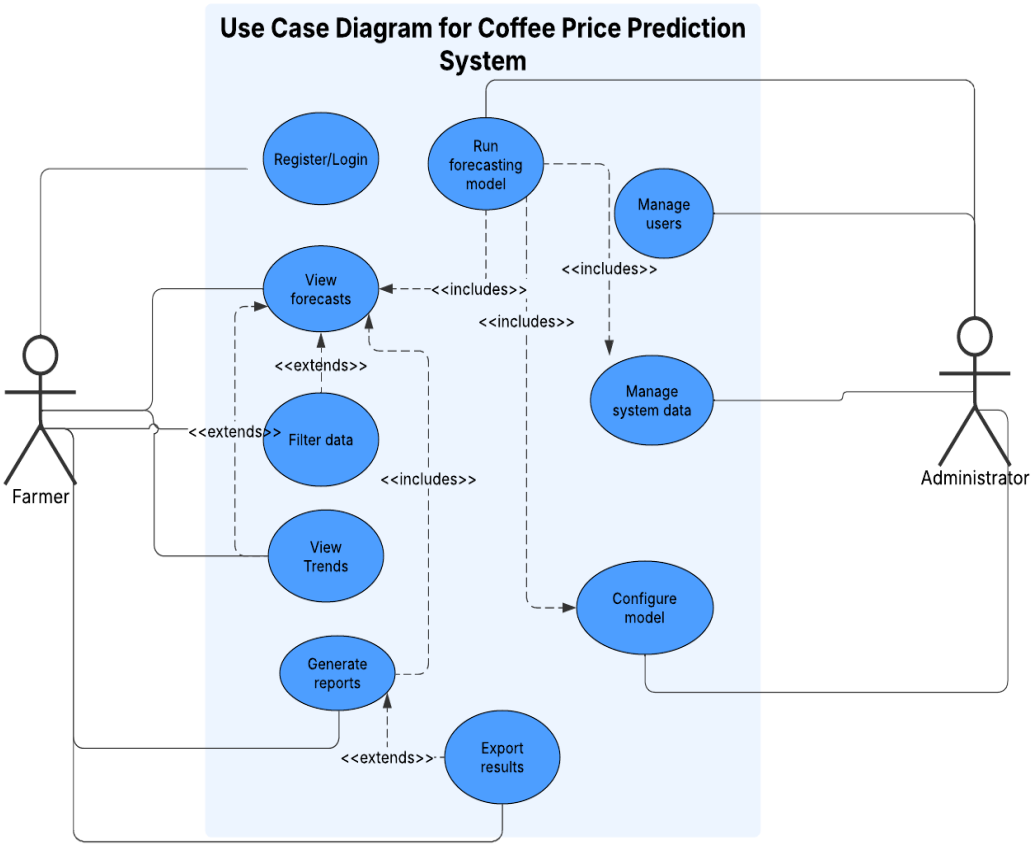



Figure 4.1: Use Case Diagram for Coffee Prediction System

### 4.5.2 Use Case Scenarios

Table 3 presents the different use case scenarios, outlining their preconditions, success scenarios, and postconditions. These descriptions correspond to the use cases illustrated in Figure 4.2.

Table 4.2: Use Case Scenarios for Coffee Price Prediction System

| Use Case | Preconditions | Success Scenario | Postconditions |
|---|---|---|---|
| **Register/Login** | User must provide valid credentials or register. | i. User enters credentials.<br>ii. System authenticates.<br>iii. User gains access. | User is logged in and can access system functionalities. |
| **View and analyze forecasts.** | User must be logged in. | i. User selects forecast view.<br>ii. System retrieves forecasted prices.<br>iii. Forecast is displayed. | Forecast is displayed to the user. |
| **Generate Reports** | User must have accessed forecast/trend data. | i. User selects report type.<br>ii. System generates a report.<br>iii. Report is displayed. | User can view or export the report. |
| **Run Forecasting Model** | The forecasting model must be accessible via an API. | i. Admin triggers model execution.<br>ii. Google Colab processes data and generates forecasts.<br>iii. Forecasting results are stored. | Forecast results are ready for display. |

| | | | |
|---|---|---|---|
| **Manage Users and Data** | Admin must be logged in. | i. Admin manages users and updates coffee price data. | ii. User records are updated in the system.<br>iii. Model forecasts are updated |

## 4.6 Sequence Diagram

Figure 4.2 illustrates the sequence of events in the coffee price forecasting system, from user authentication to the generation and presentation of price predictions. The diagram outlines how a user logs in, requests a forecast, and how the system processes the request through an API. The Web Application forwards the request to the API Service, which interacts with the forecasting model. The model analyzes historical data and generates predictions, which are then returned to the user. This process ensures that farmers, researchers, and traders can efficiently access accurate and up-to-date coffee price forecasts in an intuitive format.

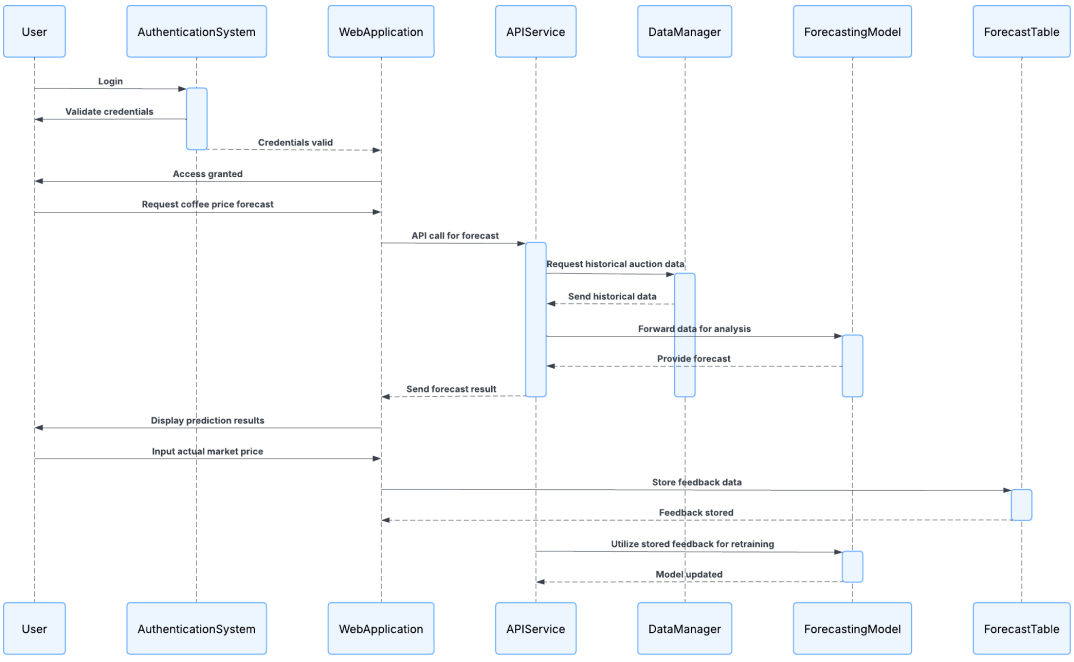



Figure 4.2: Sequence Diagram

## 4.7 ERD Diagram

Figure 4.3 illustrates the relationships between different entities in the Coffee Price Forecasting System. It outlines the key tables, their attributes, and how they interact within the database. The diagram provides a clear structure for managing users, forecast requests, reports, and API request logs, ensuring efficient data storage and retrieval.

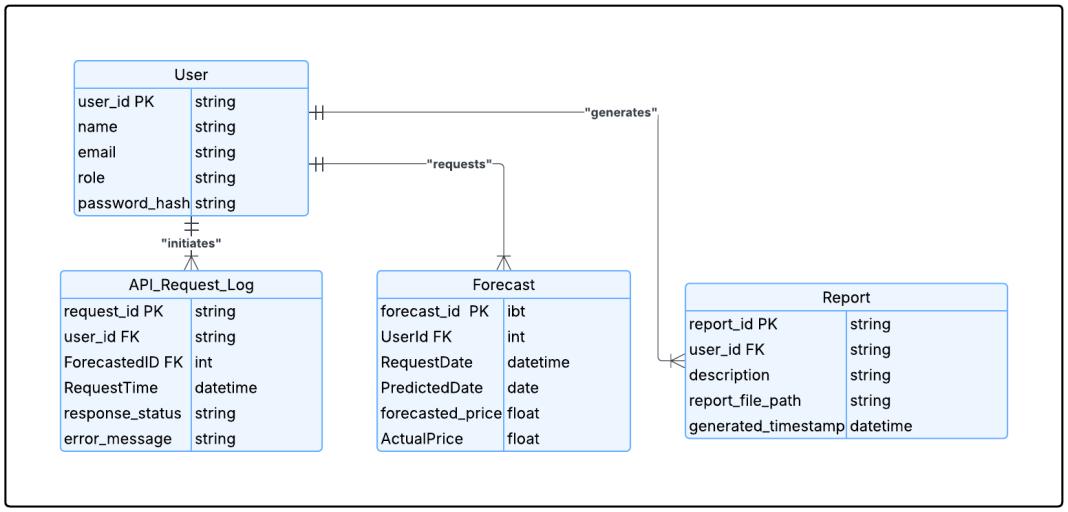



Figure 4.3: Entity Relationship Diagram

## 4.8 System Architecture

The system architecture follows a client-server model, integrating a statistical time series forecasting approach with a user-friendly web interface. The frontend enables farmers to view recent coffee price trends and select a future date to receive a price prediction. This interface is designed to be responsive and intuitive, minimizing user input requirements. The backend, developed using Python, is responsible for handling user requests, data processing, and communication with the forecasting engine. The core predictive model is implemented using SARIMAX.

The choice of SARIMAX over machine learning models like LSTM is driven by three key considerations:

i. Seasonality Handling: Coffee prices in Kenya exhibit clear seasonal patterns influenced by harvesting cycles and market trends. SARIMAX is well-suited for capturing such seasonality in time series data.

ii. Integration of Exogenous Variables: SARIMAX allows the inclusion of exogenous variables such as exchange rates, inflation, fuel prices, and export volumes—enhancing the model's predictive accuracy while retaining transparency and interpretability.

iii. Reduced User Input Complexity: Unlike deep learning models that often require multiple feature inputs during inference, SARIMAX can generate reliable forecasts with minimal user input. This allows the system to predict prices for a selected date without requiring the farmer to manually input all relevant features, improving usability and accessibility.

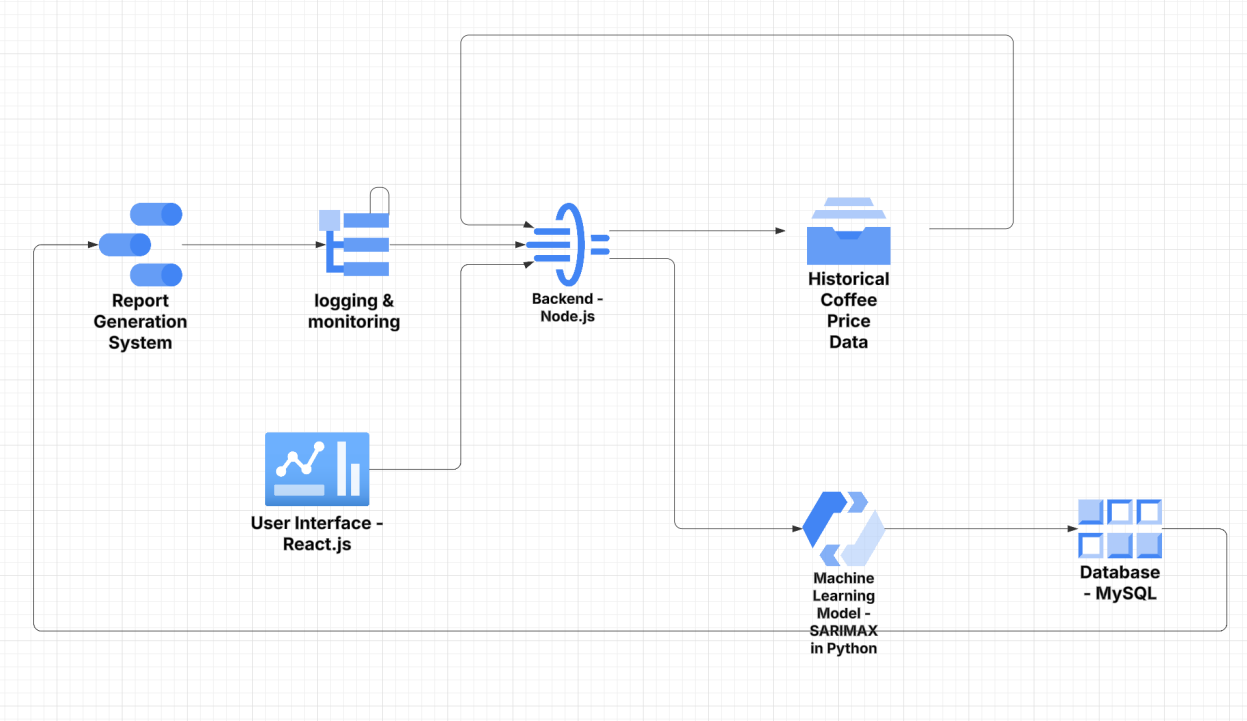



Figure 4.4: System Architecture

# Chapter 5: Implementation and Testing

## 5.1 Introduction

The system implementation and testing phase focused on developing a robust coffee price forecasting model by integrating historical auction prices, macroeconomic indicators, and time-series methodologies. This chapter outlines the model development process, covering data collection, preprocessing, feature engineering, model selection, and evaluation. A comparative analysis was conducted, assessing various forecasting techniques, including traditional statistical models, machine learning approaches, and deep learning models such as Long Short-Term Memory (LSTM) networks. The effectiveness of these models was evaluated using standard performance metrics to ensure accurate and reliable coffee price predictions.

## 5.2 Model Development

The model development phase aimed to build a reliable forecasting system for coffee prices using historical auction data and economic indicators. This section details the key steps involved in implementing the predictive model, including data preprocessing, feature engineering, model selection, and evaluation to ensure accuracy and robustness.

### 5.2.1 Data Collection

This project utilizes data from the Kenya National Bureau of Statistics (KNBS), including historical coffee production figures, auction prices, inflation rates, exchange rates, and other macroeconomic factors. The dataset spans 2019 to 2023, focusing on key variables influencing coffee prices.

| | Month | Year | Inflation_Rate | Exchange_Rate | Motor Gasoline Premium (KSh per Litre) | Light Diesel Oil | Export Quantity(KG) | Export value(KSH) | Price(USD/KG) | Price(Ksh/KG) | Precipitation(Inches) | Temperature(Degrees Fahrenheit) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | January | 2019 | 4.7 | 115.95 | 104.99 | 103.1 | 3,469,390.00 | 1,498,720,000.00 | 4.46 | 452.57 | 0.08 | 69.77 |
| 3 | February | 2019 | 4.14 | 100.23 | 101.13 | 96.83 | 4,567,370.00 | 1,903,180,000.00 | 4.49 | 449.47 | | 71.18 |
| 4 | March | 2019 | 4.35 | 100.36 | 102.13 | 97.47 | 4,351,340.00 | 2,256,180,000.00 | 2.97 | 298.47 | 0.09 | 73.48 |
| 5 | April | 2019 | 6.58 | 101.07 | 107.57 | 103.09 | 4,551,540.00 | 2,500,940,000.00 | 2.01 | 203.37 | | 73.22 |
| 6 | May | 2019 | 5.49 | 101.14 | 112.79 | 105.23 | 5,489,900.00 | 2,699,810,000.00 | 1.98 | 200.58 | | 67.63 |
| 7 | June | 2019 | 5.7 | 101.69 | 115.82 | 105.57 | 4,548,700.00 | 1,963,940,000.00 | 1.89 | 191.97 | 0.17 | 65.63 |
| 8 | July | 2019 | 6.27 | 103.16 | 116.14 | 104.74 | 5,115,450.00 | 1,713,180,000.00 | 1.91 | 196.94 | | 65.68 |
| 9 | August | 2019 | 5 | 100.3 | 111.7 | 104.92 | 3,931,890.00 | 1,461,510,000.00 | 2.1 | 216.51 | 0.15 | 65.77 |
| 10 | September | 2019 | 3.83 | 103.8 | 113.57 | 103.9 | 3,144,730.00 | 1,112,650,000.00 | 2.24 | 232.6 | 0.20 | 67.63 |
| 11 | October | 2019 | 4.95 | 103.67 | 108.83 | 102.82 | 3,985,760.00 | 1,390,060,000.00 | 2.51 | 259.85 | 0.28 | 67.23 |
| 12 | November | 2019 | 5.56 | 102.39 | 110.99 | 105.1 | 3,664,410.00 | 1,175,660,000.00 | 3.24 | 331.71 | 0.36 | 67.87 |
| 13 | December | 2019 | 5.82 | 101.5 | 109.91 | 102.28 | 1,905,940.00 | 633,740,000.00 | 4.29 | 435.44 | 0.49 | 67.35 |
| 14 | January | 2020 | 7.1 | 101.08 | 110.61 | 102.81 | 2,639,380.00 | 985,320,000.00 | 4.34 | 438.95 | 0.28 | 68.29 |
| 15 | February | 2020 | 7.17 | 100.79 | 112.58 | 105.37 | 3,168,500.00 | 1,686,820,000.00 | 4.24 | 427.28 | 0.36 | 69.31 |
| 16 | March | 2020 | 5.84 | 103.74 | 112.07 | 102.93 | 4,604,420.00 | 2,410,160,000.00 | 4.06 | 421.92 | 0.31 | 69.71 |
| 17 | April | 2020 | 6.01 | 106.41 | 94.09 | 98.84 | 4,395,520.00 | 2,590,200,000.00 | 2.75 | 294.7 | 0.34 | 68.27 |
| 18 | May | 2020 | 5.33 | 106.68 | 84.58 | 79.67 | 4,312,750.00 | 2,279,120,000.00 | 2.59 | 276.1 | 0.02 | 67.42 |
| 19 | June | 2020 | 4.59 | 106.4 | 90.34 | 75.88 | 5,414,080.00 | 2,956,330,000.00 | | | 0.39 | 64.83 |
| 20 | July | 2020 | 4.36 | 107.27 | 101.37 | 92.81 | 3,546,250.00 | 1,799,260,000.00 | 3.32 | 357.53 | 0.03 | 63.71 |
| 21 | August | 2020 | 4.36 | 108.32 | 104.83 | 95.57 | 3,181,820.00 | 1,484,150,000.00 | 4.86 | 525.29 | | 65.29 |
| 22 | September | 2020 | 4.2 | 108.41 | 106.3 | 95.45 | 3,391,490.00 | 1,606,840,000.00 | 4.47 | 484.47 | 0.22 | 65.60 |
| 23 | October | 2020 | 4.84 | 108.64 | 108.13 | 93.85 | 2,732,150.00 | 1,322,080,000.00 | 4.85 | 526.8 | 0.11 | 68.77 |
| 24 | November | 2020 | 5.33 | 109.25 | 106.72 | 91.64 | 3,594,290.00 | 1,836,610,000.00 | 5.2 | 568.36 | 0.10 | 68.29 |
| 25 | December | 2020 | 5.62 | 110.59 | 107.69 | 92.75 | 2,405,440.00 | 1,284,840,000.00 | 5.98 | 660.07 | 0.26 | 69.08 |
| 26 | January | 2021 | 5.69 | 109.83 | 107.86 | 97.33 | 2,129,360.00 | 1,342,450,000.00 | 6.34 | 697.01 | 0.13 | 69.50 |
| 27 | February | 2021 | 5.78 | 109.68 | 116.03 | 102.84 | 3,481,130.00 | 2,161,120,000.00 | 6.06 | 664.49 | 0.71 | 68.92 |
| 28 | March | 2021 | 5.9 | 109.73 | 123.66 | 108.58 | 6,065,070.00 | 4,556,940,000.00 | 4.96 | 544.14 | 0.08 | 70.65 |
| 29 | April | 2021 | 5.76 | 107.95 | 123.66 | 108.58 | 3,337,010.00 | 2,307,120,000.00 | 4.04 | 435.74 | 0.29 | 68.86 |
| 30 | May | 2021 | 5.87 | 107.43 | 127.21 | 108.58 | 4,430,330.00 | 3,010,490,000.00 | | | 0.50 | 66.39 |
| 31 | June | 2021 | 6.32 | 107.81 | 127.98 | 108.58 | 3,436,880.00 | 2,271,770,000.00 | 5.1 | 550.7 | | 63.40 |
| 32 | July | 2021 | 6.55 | 108.14 | 127.98 | 108.58 | 2,696,290.00 | 1,763,900,000.00 | 6.23 | 674.11 | 0.04 | 63.32 |
| 33 | August | 2021 | 6.57 | 109.24 | 127.98 | 108.58 | 2,504,380.00 | 1,657,820,000.00 | 6.26 | 683.88 | | 64.81 |

Figure 5.1: Raw data from KNBS

### 5.2.2 Data Preparation and Preprocessing

Data preprocessing was essential to ensure the model received clean and reliable data for training. This process included handling missing values, detecting and managing outliers, and applying necessary transformations to prepare the data for time-series and machine learning models. Specifically, missing precipitation values were interpolated based on corresponding temperature data, ensuring continuity and usability for the models.

```
# Find rows where Precipitation is missing
missing_precipitation = df[df['Precipitation(Inches)'].isnull()][['Month', 'Year', 'Temperature(Degrees Fahrenheit)']]
print(missing_precipitation)
```

```
        Month  Year  Temperature(Degrees Fahrenheit)
1    February  2019                        71.178571
3       April  2019                        73.217391
4         May  2019                        67.625000
6        July  2019                        65.677419
19     August  2020                        65.290323
29       June  2021                        63.400000
31     August  2021                        64.806452
32  September  2021                        66.933333
44  September  2022                        65.666667
45    October  2022                        68.516129
46   November  2022                        67.206897
48    January  2023                        68.935484
49   February  2023                        71.857143
50      March  2023                        70.935484
55     August  2023                        66.354839
```

**Fill missing values based on the month**

Kenya has two main rainy seasons: "Long Rains" (March – May) 🌧️ "Short Rains" (October – December) ⛅

```
# Identify dry and rainy months
dry_months = ['January', 'February', 'July', 'August']
rainy_months = ['April', 'May', 'September', 'October', 'November']

# Fill missing values: Dry months → 0, Rainy months → Interpolate
df.loc[df['Month'].isin(dry_months) & df['Precipitation(Inches)'].isnull(), 'Precipitation(Inches)'] = 0

# Interpolate for rainy months
df['Precipitation(Inches)'] = df['Precipitation(Inches)'].interpolate()
```

Figure 5.2: Handling missing values.

### 5.2.3 Feature Engineering

Feature engineering played a crucial role in enhancing the model's predictive capability. As illustrated in Figure 5.3, time-based features such as month, quarter, and year were extracted from the dataset to enable the model to recognize seasonal trends in coffee prices. Additionally, lag features were created to incorporate past values of coffee prices and macroeconomic indicators, allowing models to learn from historical trends.

```
# Create lag features (Previous month's values)
df['Price(Ksh/KG)_Lag1'] = df['Price(Ksh/KG)'].shift(1)  # Lag of 1 month
df['Exchange_Rate_Lag1'] = df['Exchange_Rate'].shift(1)
df['Inflation_Rate_Lag1'] = df['Inflation_Rate'].shift(1)

# Drop rows with NaN values (since first row won't have previous month's data)
df.dropna(inplace=True)

# Display sample data
print(df[['Price(Ksh/KG)', 'Price(Ksh/KG)_Lag1', 'Exchange_Rate', 'Exchange_Rate_Lag1']].head().to_string())
```

```
            Price(Ksh/KG)  Price(Ksh/KG)_Lag1  Exchange_Rate  Exchange_Rate_Lag1
Date
2019-02-01         449.47              452.57         100.23              115.95
2019-03-01         298.47              449.47         100.36              100.23
2019-04-01         209.94              298.47         101.07              100.36
2019-05-01         209.94              209.94         101.14              101.07
2019-06-01         209.94              209.94         101.69              101.14
```

Figure 5.3: Creating Lag Features for Coffee Prices

To capture long-term trends, rolling mean features were introduced, averaging coffee prices over a specified time window, as illustrated in Figure 5.4. This approach smoothed short-term fluctuations while preserving meaningful trends. Furthermore, feature selection was performed to remove highly correlated variables, reducing redundancy and improving model interpretability.

**Add Moving Averages (Rolling Means)**

Moving averages smooth out fluctuations in the data.

```
# Rolling mean of past 3 months
df['Price(Ksh/KG)_Rolling3'] = df['Price(Ksh/KG)'].rolling(window=3).mean()

# Display sample data
print(df[['Price(Ksh/KG)', 'Price(Ksh/KG)_Rolling3']].head(10).to_string())
```

```
            Price(Ksh/KG)  Price(Ksh/KG)_Rolling3
Date
2019-02-01         449.47                    NaN
2019-03-01         298.47                    NaN
2019-04-01         209.94             319.293333
2019-05-01         209.94             239.450000
2019-06-01         209.94             209.940000
2019-07-01         209.94             209.940000
2019-08-01         216.51             212.130000
2019-09-01         232.60             219.683333
2019-10-01         259.85             236.320000
2019-11-01         331.71             274.720000
```

Figure 5.4: Adding moving averages.

For the machine learning models, Figure 5.5 shows where feature scaling was performed using StandardScaler to ensure that the numerical features were on a similar scale, enabling the model to learn effectively.

```
from sklearn.preprocessing import StandardScaler
# Make a copy of the original unscaled data
df_unscaled = df.copy()  # Keep this before applying scaling
# Initialize scaler
scaler = StandardScaler()

# Define columns to scale (excluding the target variable & cyclical month features)
features_to_scale = df.columns.difference(['Price(Ksh/KG)', 'Month_sin', 'Month_cos'])  # Exclude target and cyclical features

# Apply scaling
df[features_to_scale] = scaler.fit_transform(df[features_to_scale])

# Verify scaled data
df.head()
```

| | Year | Inflation_Rate | Motor Gasoline Premium (KSh per Litre) | Export Quantity(KG) | Price(Ksh/KG) | Precipitation(Inches) | Temperature(Degrees Fahrenheit) | Month_sin | Month_cos |
|---|---|---|---|---|---|---|---|---|---|
| 0 | -1.455305 | -0.907236 | -1.076036 | -0.336556 | 452.57 | -0.491877 | 0.847066 | 0.500000 | 8.660254e-01 |
| 1 | -1.455305 | -1.261089 | -1.175792 | 0.479660 | 449.47 | -1.021265 | 1.437694 | 0.866025 | 5.000000e-01 |
| 2 | -1.455305 | -1.128394 | -1.149948 | 0.319068 | 298.47 | -0.403646 | 2.407215 | 1.000000 | 6.123234e-17 |
| 3 | -1.455305 | 0.280697 | -1.009359 | 0.467893 | 209.94 | -0.234536 | 2.295144 | 0.866025 | -5.000000e-01 |
| 4 | -1.455305 | -0.408052 | -0.874456 | 1.165450 | 209.94 | -0.065426 | -0.056802 | 0.500000 | -8.660254e-01 |

Figure 5.5: Feature Scaling

### 5.2.4 Model Selection: Time Series vs. Machine Learning

With the data prepared and features selected, the next step was to determine which type of model to use for forecasting. Given that the problem involves predicting prices over time, both time series models and machine learning models were considered.

#### 5.2.4.1 Time Series Modeling

The first model tested was SARIMA (Seasonal AutoRegressive Integrated Moving Average), which is a well-established approach for handling time series data with seasonal and trend components. SARIMA models were configured with various orders (p, d, q) to account for seasonality and trends in the data. The initial selection of the parameters (p, d, q) was based on plots of the AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF), which provided insights into the temporal dependencies in the data. These plots helped to guide the selection of appropriate values for the autoregressive (AR), differencing (I), and moving average (MA) components.

```
# Plot ACF and PACF
fig, axes = plt.subplots(1, 2, figsize=(14, 5))

# ACF plot (for q and Q selection)
plot_acf(df['Price(Ksh/KG)'].diff(2).dropna(), ax=axes[0], lags=30)
axes[0].set_title("Autocorrelation Function (ACF)")

# PACF plot (for p and P selection)
plot_pacf(df['Price(Ksh/KG)'].diff(2).dropna(), ax=axes[1], lags=30)
axes[1].set_title("Partial Autocorrelation Function (PACF)")

plt.show()
```

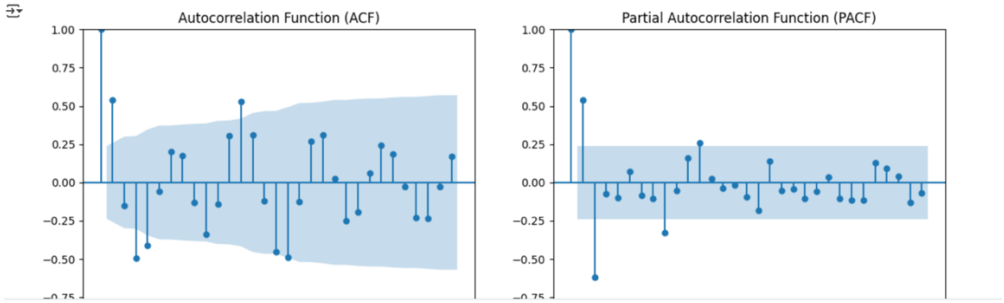



Figure 5.6: ACF and PACF plot

To optimize performance, grid search was conducted to identify the best combination of SARIMA hyperparameters.

```
# Grid search over all parameter combinations
for param in pdq:
    for seasonal_param in seasonal_pdq:
        try:
            model = sm.tsa.SARIMAX(
                df["Price(Ksh/KG)"],
                order=param,
                seasonal_order=seasonal_param,
                enforce_stationarity=False,
                enforce_invertibility=False
            )
            results = model.fit()
            if results.aic < best_aic:
                best_aic = results.aic
                best_params = (param, seasonal_param)
            print(f"SARIMA{param}x{seasonal_param} - AIC: {results.aic}")
        except:
            continue

print(f"✅ Best SARIMA Model: {best_params} - AIC: {best_aic}")
```

Figure 5.7: Fine tuning SARIMA using grid search

#### 5.2.4.2 Machine Learning Models

Several machine learning models were explored, including Linear Regression, Random Forest, XGBoost and LSTM. The dataset was split into training and testing sets, with machine learning models trained using both scaled and unscaled features.

```
# Scaled dataset for Linear Regression
X_scaled = df[selected_features]  # Scaled features
y_scaled = df["Price(Ksh/KG)"]

# Unscaled dataset for tree models
X_unscaled = df_unscaled[selected_features]  # Unscaled features
y_unscaled = df_unscaled["Price(Ksh/KG)"]

# Train-test split (80-20 split)
X_train_scaled, X_test_scaled, y_train_scaled, y_test_scaled = train_test_split(
    X_scaled, y_scaled, test_size=0.2, random_state=42
)
X_train_unscaled, X_test_unscaled, y_train_unscaled, y_test_unscaled = train_test_split(
    X_unscaled, y_unscaled, test_size=0.2, random_state=42
)
```

Figure 5.8: Splitting data into train and test

Random Forest and XGBoost, both ensemble methods, were particularly useful in capturing nonlinear relationships in the data. Hyperparameter tuning was conducted using RandomizedSearchCV to optimize model parameters, improving predictive accuracy.

```
# Define hyperparameter grid
param_grid_rf = {
    "n_estimators": [50, 100, 200],  # Number of trees
    "max_depth": [None, 10, 20, 30],  # Max depth of trees
    "min_samples_split": [2, 5, 10],  # Min samples to split a node
    "min_samples_leaf": [1, 2, 4],  # Min samples per leaf
}
```

Figure 5.9: Hyperparameter tuning Random Forest using RandomizedSearchCV

A sequential LSTM model was constructed with multiple layers, incorporating dropout regularization to prevent overfitting. The model was trained using the Adam optimizer and mean squared error as the loss function. After training, predictions were generated and inverse-transformed to obtain actual price values.

```
# Define LSTM model with more units, layers, and dropout
lstm_model = Sequential()
lstm_model.add(LSTM(units=100, return_sequences=True, input_shape=(X_train_lstm.shape[1], X_train_lstm.shape[2])))
lstm_model.add(Dropout(0.2))
lstm_model.add(LSTM(units=50, return_sequences=False))
lstm_model.add(Dropout(0.2))
lstm_model.add(Dense(units=1))

# Compile the model
lstm_model.compile(optimizer='adam', loss='mean_squared_error')

# Train the model
history = lstm_model.fit(X_train_lstm, y_train_lstm, epochs=100, batch_size=32, validation_data=(X_test_lstm, y_test_lstm), verbose=1)
```

Figure 5.10: Training LSTM model

### 5.2.5 Model Evaluation

Each model was evaluated based on standard error metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared ($R^2$) to assess predictive accuracy.

### 5.2.6 Linear Regression Performance

The linear regression model served as a baseline for comparison. Figure 5.11 shows that while it provided a reasonable fit, it struggled with capturing complex price fluctuations, resulting in relatively high RMSE and moderate $R^2$ scores.

```
🔷 Linear Regression Performance:
MAE: 76.4335
MSE: 7458.4006
RMSE: 86.3620
R²: 0.7680
```

Figure 5.11: Linear Regression Performance

#### 5.2.6.1 Random Forest and XGBoost Performance

Random Forest and XGBoost outperformed the linear regression model by leveraging ensemble learning to improve prediction accuracy. Figure 5.12 shows models were fine-tuned using hyperparameter optimization, leading to improved performance with lower RMSE and

higher $R^2$ values. XGBoost demonstrated a slight improvement over Random Forest in capturing subtle price variations.

```
# Define a more focused hyperparameter grid for XGBoost
param_grid_xgb_refined = {
    "learning_rate": [0.01, 0.1, 0.2],  # Narrow learning rate range
    "max_depth": [5, 6, 7],  # Narrow max_depth range
    "n_estimators": [100, 200],  # Keep a smaller number of estimators
    "subsample": [0.8, 1.0],  # Focus on higher values for subsample
    "colsample_bytree": [0.8, 1.0],  # Similar with colsample_bytree
    "gamma": [0, 0.1],  # Narrow gamma range
    "min_child_weight": [1, 3]  # Narrow min_child_weight range
}
```

Figure 5.12: Fine-tuning XGBoost

```
🔷 Random Forest Performance:
MAE: 56.2394
MSE: 4475.5512
RMSE: 66.8996
R²: 0.8608
```

Figure 5.13: Random Forest Performance before fine-tuning

```
Best Refined XGBoost Params: {'subsample': 0.8, 'n_estimators': 200, 'min_child_weight': 1, 'max_depth': 7, 'learning_rate': 0.1, 'gamma': 0, 'colsample_byt
{'Model': 'Refined XGBoost (RandomizedSearchCV)',
 'MAE': 50.61689238823786,
 'MSE': 4318.8350440595505,
 'RMSE': np.float64(65.71784418298846),
 'R²': 0.865679344001496}
```

Figure 5.14: Fine-tuned XGBoost



### 5.2.6.2 SARIMAX

The SARIMAX model demonstrated strong performance in capturing the seasonal patterns and underlying trends in coffee price data. By incorporating exogenous variables such as exchange rates, fuel prices, and inflation rates, the model provided more accurate and explainable forecasts compared to baseline univariate models. Figure 5.15 presents the model's diagnostic plots, confirming that residuals followed a white noise pattern, indicating a good model fit.

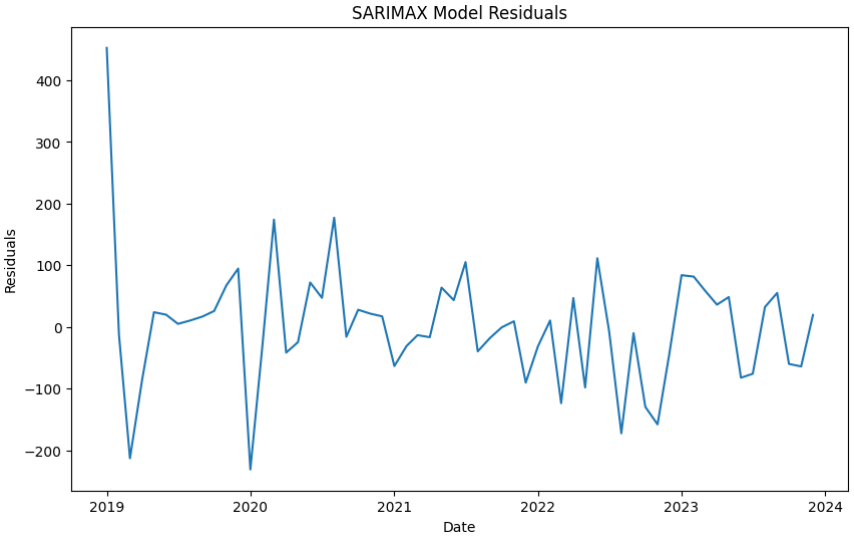



Figure 5.15: SARIMAX Model Residuals

The forecast accuracy was evaluated using RMSE and MAPE, with results showing the model's suitability for short-term forecasting.

```
MAE: 78.2290184280752
RMSE: 92.87231408708371
MAPE: 12.6333203897586666%
R²: 0.9826
```

Figure 5.16: SARIMAX Evaluation Metrics

## 5.3 Conclusion

The implementation of a coffee price forecasting model involved several critical steps, including data preparation, feature engineering, and model selection. By comparing time-series models, machine learning techniques, and deep learning approaches, an optimized forecasting system was developed. SARIMAX was selected as the final model due to its ability to capture seasonal trends and incorporate exogenous variables, enabling accurate and interpretable forecasts without requiring extensive user input. The final model provides an effective tool for stakeholders in the coffee industry, offering valuable insights into price trends. The next chapter discusses the conclusions drawn from this research and potential areas for future improvement.

# Chapter 6: Discussions of Results

## 6.1 Introduction

This chapter aims at discussing the findings of the study considering the research objectives. The purpose of this paper is to assess the performance of the predictive model that has been used to forecast coffee prices and determine whether it meets the needs of Kenyan smallholder coffee farmers. The research sought to develop and test and user-friendly forecasting model using machine learning algorithms, namely LSTM networks, to predict coffee prices using past data and economic factors. The results were then used to assess the performance of the model in generating accurate price forecasts that can be useful to farmers in their financial planning.

In addition, the chapter assesses the extent to which the research objectives were achieved, and the challenges faced in the study. The discussion also highlights some of the areas that could be developed further and improved in the future to make the model more useful for smallholder coffee farmers.

## 6.2 Study Results

In this study, multiple forecasting models were explored to predict coffee prices, including Linear Regression, Random Forest, XGBoost, and LSTM. After evaluating their performance using key error metrics, SARIMAX was selected as the final model due to its strength in handling seasonality and integrating relevant external factors. Its ability to generate accurate, interpretable forecasts made it well-suited for the price prediction task in this context.

### 6.2.1 Performance Evaluation of Models

#### 6.2.1.1 Linear Regression

Linear regression provided a baseline model for comparison. Despite its simplicity, it showed a relatively high mean absolute error (MAE = 76.43) and root mean squared error (RMSE = 86.36), with an $R^2$ score of 0.768. These results indicate that linear regression was not well-suited for capturing the complexities in coffee price movements, which are influenced by multiple factors, including economic indicators and seasonal variations.

#### 6.2.1.2 Random Forest

Random Forest performed significantly better than linear regression, achieving an MAE of 56.24 and an RMSE of 66.90, with an $R^2$ score of 0.8608. The model benefitted from its ability to

capture nonlinear relationships in the data. However, Random Forest does not inherently handle temporal dependencies, making it less suitable for time-series forecasting.

#### 6.2.1.3 XGBoost

XGBoost, a powerful gradient boosting technique, provided performance close to that of Random Forest, with an MAE of 56.21 and RMSE of 67.97. While it performed well in capturing nonlinearities, its reliance on tree-based learning meant it could not explicitly model sequential dependencies, which are essential in predicting future price trends.

### 6.2.2 Hyperparameter Tuning Results

Hyperparameter tuning was done to improve the performance of both Random Forest and XGBoost models. The tuned Random Forest model had an RMSE of 67.40 and an $R^2$ of 0.8587, which was slightly better than the previous model. Likewise, the tuning of XGBoost led to an improved RMSE of around 65.74, which also improved the predictive power of the model. However, these improvements were still insufficient to capture the sequential nature of coffee price movements in both models.

### 6.2.3 Predictive Model Performance and Insights

Random Forest and XGBoost proved effective in static, feature-based regression tasks. However, since coffee prices exhibit temporal and seasonal patterns, time-series models were more appropriate. SARIMAX was ultimately preferred for its ability to model both seasonality and external economic factors, such as exchange rates and fuel prices. The validation process using historical data showed that SARIMAX provided accurate forecasts with a small error margin, as illustrated in Figure 6.1. Its interpretability and capacity to incorporate exogenous variables without requiring extensive feature input made it more practical and adaptable than machine learning-based models for this use case.

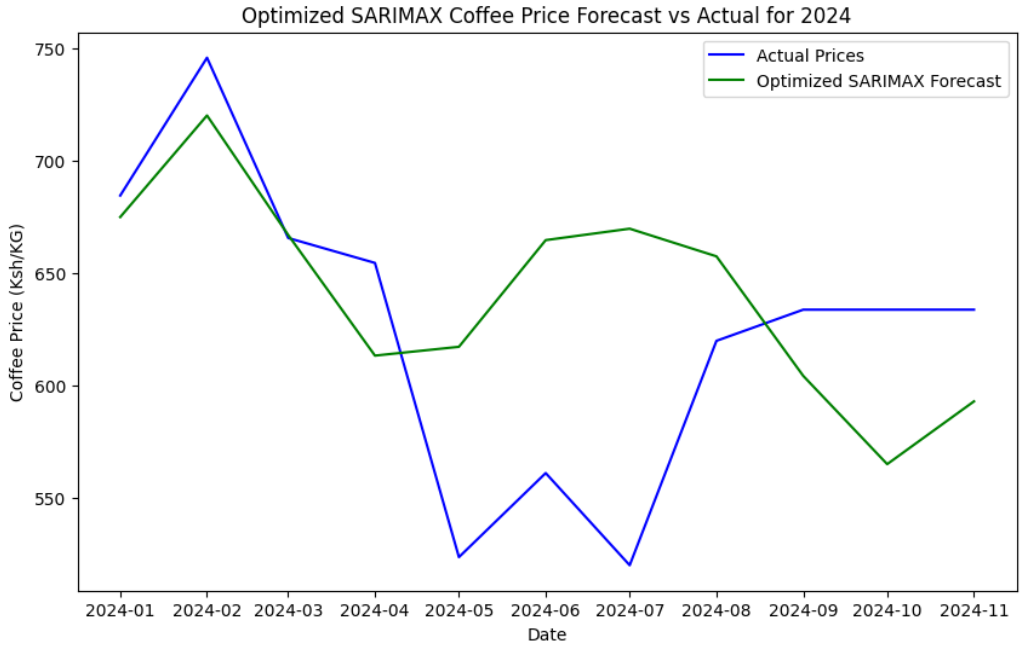



Figure 6.1: SARIMAX predicted vs actual prices

### 6.2.4 Implications for Smallholder Coffee Farmers

The implications of the study are important for smallholder coffee farmers in Kenya who are faced with price risks and market volatility. The forecasting model is a useful tool that can help farmers to make the right decisions on when to sell their produce. In this way, the model can help in managing risks, especially when entering into contracts and using derivatives to hedge against price volatility. Furthermore, the findings show that AI has the potential to revolutionize the agricultural sector, and the need to embrace technology in farming to enhance financial viability and market access. However, for successful adoption, it may be necessary to incorporate training and awareness programs so that farmers can understand and have confidence in the predictions made by the model.

### 6.2.5 Limitations and Areas for Improvement

Although the forecasting model achieved promising results, certain limitations must be acknowledged. One of the primary challenges is its dependence on historical data, which makes it less effective in predicting sudden market shocks caused by external factors such as government policy changes or global trade disruptions. Additionally, the availability and reliability of data remain a concern, as some macroeconomic indicators are reported with delays, potentially affecting the model's real-time applicability. Another limitation is the need for continuous

retraining to ensure that the model remains relevant and accurate over time. Addressing these challenges could involve incorporating hybrid models that combine deep learning techniques with external market intelligence, allowing for improved forecasting capabilities.

# Chapter 7: Conclusions and Recommendations

## 7.1 Conclusion

The coffee price prediction model developed in this study has demonstrated its usefulness in forecasting coffee prices with a reasonable degree of accuracy. By leveraging historical auction prices and key economic indicators such as inflation rates, exchange rates, and fuel costs, the model provides actionable insights for coffee farmers, traders, and policymakers. Its strength lies in the ability to account for seasonal patterns and exogenous variables, enabling stakeholders to anticipate price volatility and make informed financial decisions.The SARIMAX model was selected as the final forecasting approach due to its capacity to handle seasonality and incorporate external economic factors without requiring the user to manually input these variables. This made it a practical and effective choice for this application. After thorough data preprocessing, feature engineering, and parameter tuning, the model consistently produced accurate predictions, thereby validating its effectiveness in the context of Kenya's coffee market.

In addition, the integration of the SARIMAX model with a mobile-friendly web-based interface enhances accessibility for coffee farmers and other end-users. By allowing users to input a future date and receive a forecasted price, the platform empowers farmers to better plan the timing of their sales or storage decisions. This system contributes to reducing economic vulnerability among smallholder producers by equipping them with data-driven tools for navigating market fluctuations.Looking ahead, advances in AI and machine learning can further improve the accuracy and responsiveness of the predictive model. To maintain high forecasting performance over time, the model will need to be periodically updated with new data, retrained, and fine-tuned to reflect evolving market conditions.

## 7.2 Recommendations

Based on the findings of this study, the following recommendations are proposed to enhance the adoption and effectiveness of the coffee price prediction model:

i. The model should be integrated with coffee trading platforms and cooperatives to provide real-time price predictions, enabling farmers to make more strategic sales decisions.
ii. Regular updates of the training dataset should be conducted to ensure that the model captures emerging trends in the coffee market, including external economic factors affecting price movements.

iii. Efforts should be made to educate coffee farmers and stakeholders on the use of the forecasting tool. Training sessions or workshops can be organized to improve adoption and utilization.

iv. Government agencies and agricultural organizations should consider incorporating predictive analytics into their policy frameworks to enhance price stabilization strategies and farmer support programs.

v. Additional features, such as weather conditions, coffee grading, and demand fluctuations in the global market, should be incorporated into the model to improve prediction accuracy.

### 7.3 Suggestions for Future Work

i. Future studies can investigate other machine learning models, such as Transformer-based architectures, to determine if higher accuracy can be achieved compared to SARIMAX.

ii. Expanding the model to include data from other coffee producing regions can enhance its applicability across different market conditions.

iii. Research can focus on integrating factors such as supply chain disruptions, climate change impacts, and logistical constraints into the forecasting model.

iv. The development of a real-time price forecasting system using streaming data sources can enhance the responsiveness of the model to immediate market fluctuations.

v. Future studies can evaluate the financial benefits of adopting predictive analytics in coffee pricing by assessing its impact on farmer income and market efficiency.

By implementing these recommendations and extending the research, the coffee price forecasting model can be further optimized to provide more accurate, reliable, and user-friendly predictions, ultimately benefiting coffee farmers and industry stakeholders.

# References


Ameur, H B., Ftiti, Z., & Louhichi, W. (2021, July 27). Revisiting the relationship between spot and futures markets: evidence from commodity markets and NARDL framework. Springer Science+Business Media, 313(1), 171-189. https://doi.org/10.1007/s10479-021-04172-3

Aramonte, S., & Todorov, K. (2021, April 12). Futures-based commodity ETFs when storage is constrained. https://ideas.repec.org/p/bis/bisblt/41.html

Bennun, L., & Njoroge, P. (1999, January 1). Important bird areas in Kenya /. https://doi.org/10.5962/bhl.title.87589

Gay, G D., & Hull, J. (1990, March 1). Options, Futures, and Other Derivative Securities.. Wiley, 45(1), 312-312. https://doi.org/10.2307/2328826

Gay, G D., & Hull, J. (1990, March 1). Options, Futures, and Other Derivative Securities.. Wiley, 45(1), 312-312. https://doi.org/10.2307/2328826

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Grigsby, J., Wang, Z., Nguyen, N., & Qi, Y. (2021). *Long-Range Transformers for Dynamic Spatiotemporal Forecasting*. https://arxiv.org/pdf/2109.12218

Hull, J. (2001, July 20). Fundamentals of Futures and Options Markets. https://www.gbv.de/dms/zbw/625299647.pdf

Kalele, D., Ogara, W., Oludhe, C., & Onono, J O. (2021, July 1). Climate change impacts and relevance of smallholder farmers' response in arid and semi-arid lands in Kenya. Elsevier BV, 12, e00814-e00814. https://doi.org/10.1016/j.sciaf.2021.e00814

Kenya at a glance | FAO in Kenya. (2023, January 1). https://www.fao.org/kenya/fao-in-kenya/kenya-at-a-glance/en

Kiani, A K., Sardar, A., Khan, W., He, Y., Bilgiç, A., Kuşlu, Y., & Raja, M A Z. (2021, August 25). Role of Agricultural Diversification in Improving Resilience to Climate Change: An Empirical Analysis with Gaussian Paradigm. Multidisciplinary Digital Publishing Institute, 13(17), 9539-9539. https://doi.org/10.3390/su13179539

Kim, H., Kim, H., & Park, Y. (2022, January 1). Perpetual Contract NFT as Collateral for DeFi Composability. Cornell University. https://doi.org/10.48550/arxiv.2208.06472

Kinoti, K D. (2018, January 1). Dynamics of Climate Change Adaptations on Horticultural Land Use Practices around Mt. Kenya East Region. Science Publishing Group, 7(1), 1-1. https://doi.org/10.11648/j.ajep.20180701.11

Li, Y., Lu, X., Xiong, H., Tang, J., Su, J., Jin, B., Dou, D., & Research, B. (2023). *Towards Long-Term Time-Series Forecasting: Feature, Pattern, and Distribution*. https://arxiv.org/pdf/2301.02068

Manfredo, M R., & Richards, T J. (2009, January 1). Hedging with weather derivatives: a role for options in reducing basis risk. Chapman and Hall London, 19(2), 87-97. https://doi.org/10.1080/09603100701765166

Matsui, T., Alali, A., & Knottenbelt, W J. (2022, May 2). On the Dynamics of Solid, Liquid and Digital Gold Futures. https://doi.org/10.1109/icbc54727.2022.9805528

Nyang'au, J O., Mohamed, J H., Mango, N., Makate, C., & Wangeci, A. (2021, April 1). Smallholder farmers' perception of climate change and adoption of climate smart agriculture practices in Masaba South Sub-county, Kisii, Kenya. Elsevier BV, 7(4), e06789-e06789. https://doi.org/10.1016/j.heliyon.2021.e06789

ODSC Community . (2020, December 15). *Understanding the Mechanism and Types of Recurrent Neural Networks*. Open Data Science - Your News Source for AI, Machine Learning & More. https://opendatascience.com/understanding-the-mechanism-and-types-of-recurring-neural-networks/

Ollerton, R L. (2015, January 29). A unifying framework for teaching probability event types. Taylor & Francis, 46(5), 790-794. https://doi.org/10.1080/0020739x.2015.1005702 Song, G S., & Kim, M C. (2021, February 10). Mathematical Formulation and Analytic Solutions for Uncertainty Analysis in Probabilistic Safety Assessment of Nuclear Power Plants. Multidisciplinary Digital Publishing Institute, 14(4), 929-929. https://doi.org/10.3390/en14040929

Sánchez-Verdasco, J. (2018, January 1). The Use of Derivatives to Hedge Market Risk in Corporate Financing. RELX Group (Netherlands). https://doi.org/10.2139/ssrn.3230485

Srivastava, S. (2022, August 25). *AI/ML Techniques for Commodity Price Forecasting | PriceVision*. PriceVision Blog; PriceVision Blog. https://pricevision.ai/blog/commodity-price-forecasting-artificial-intelligence-ai-ml-techniques/

Tong, T., & Reuer, J J. (2007, June 19). Real Options in Strategic Management. , 3-28. https://doi.org/10.1016/s0742-3322(07)24001-x

Wen, X. (2022, January 1). Analysis of Relationships between Common Distributions Based on Computing Science. , 5(12). https://doi.org/10.25236/ajcis.2022.051207 Mishra, S., & Datta-Gupta, A. (2018, January 1). Distributions and Models Thereof. Elsevier BV, 31-67. https://doi.org/10.1016/b978-0-12-803279-4.00003-1

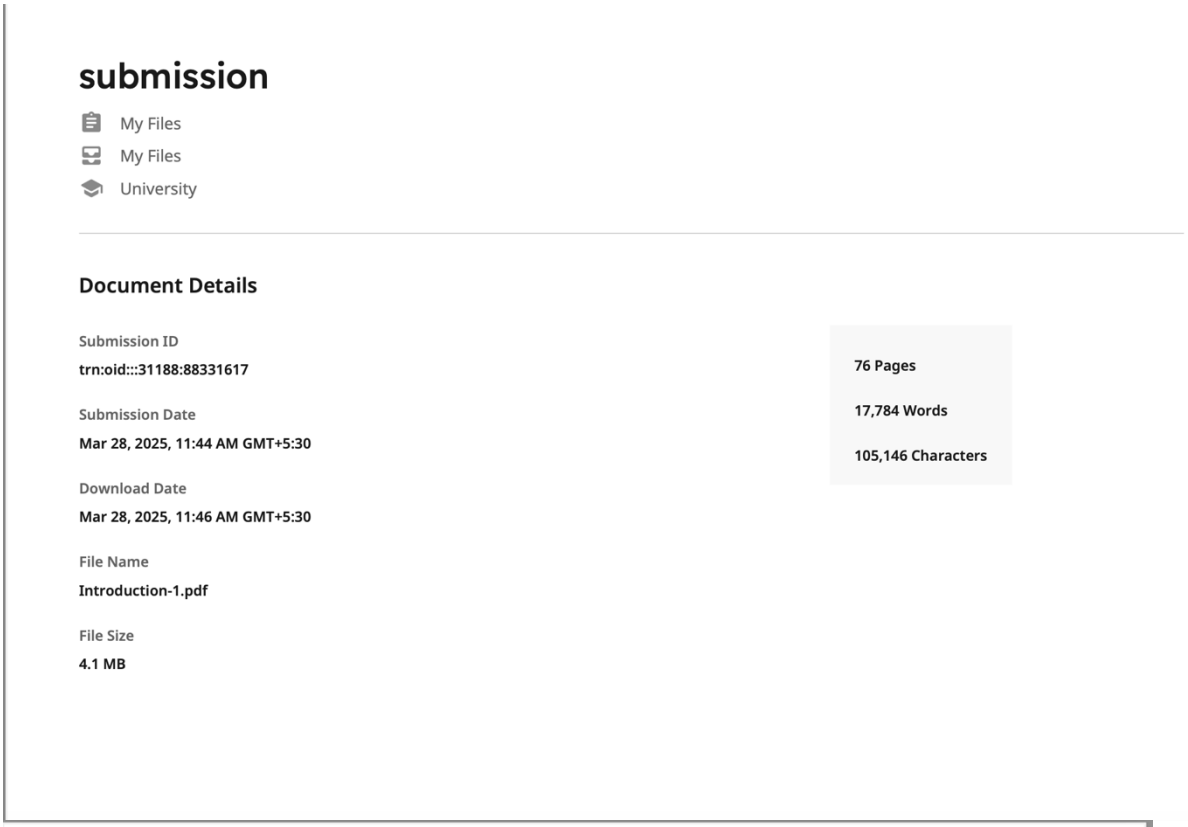
Wu, H., Xu, J., Wang, J., & Long, M. (2021). *Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting*. https://arxiv.org/pdf/2106.13008

Yang, F., Liu, J., Zhang, R., & Yao, Y. (2023). *Diffusion characteristics classification framework for identification of diffusion source in complex networks*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285563

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325

# Appendices

## Appendix A: Similarity Report



submission

My Files

My Files

University

Document Details

Submission ID

trn:oid:::31188:88331617

Submission Date

Mar 28, 2025, 11:44 AM GMT+5:30

Download Date

Mar 28, 2025, 11:46 AM GMT+5:30

File Name

Introduction-1.pdf

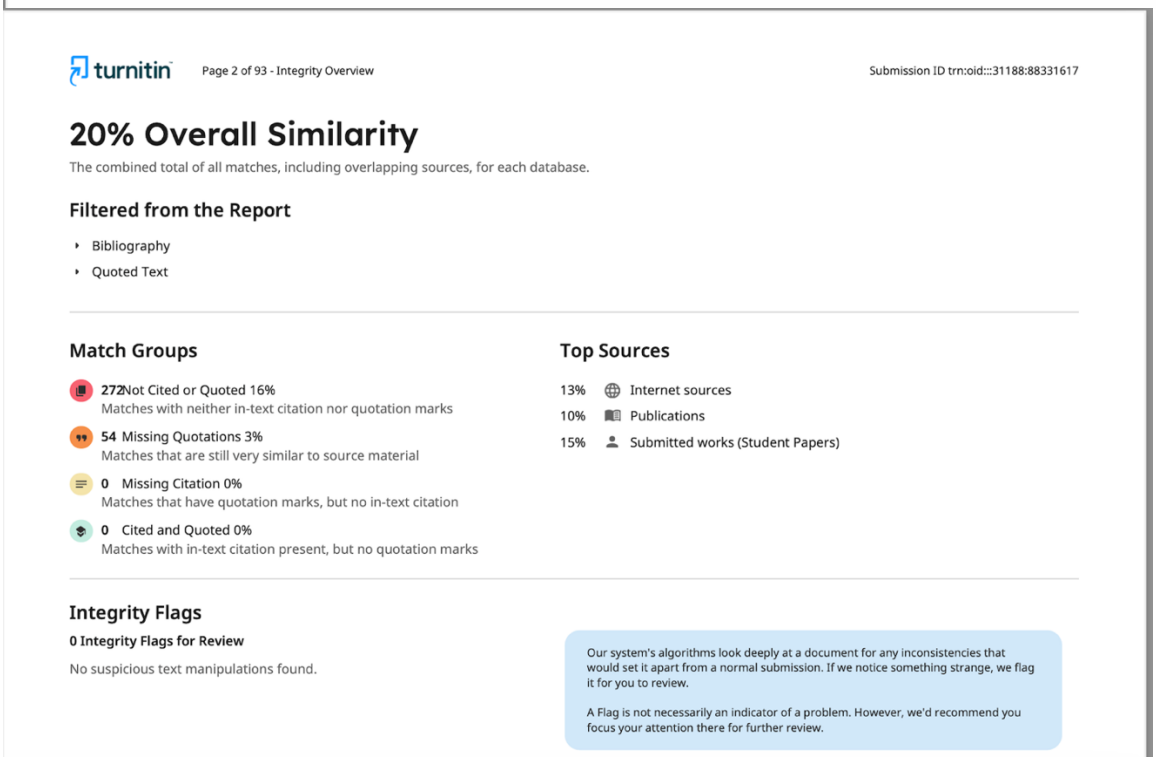
File Size

4.1 MB

76 Pages

17,784 Words

105,146 Characters



turnitin Page 2 of 93 - Integrity Overview Submission ID trn:oid:::31188:88331617

20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

272 Not Cited or Quoted 16%
Matches with neither in-text citation nor quotation marks

54 Missing Quotations 3%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

13% Internet sources

10% Publications

15% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

**11th February 2025**

Ms Runyiri Joan,
joan.runyiri@strathmore.edu

Dear Ms Runyiri,

**RE: A Predictive Model for Hedging Futures Contracts to Stabilize Kenyan Coffee Farmers' Income**

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2642/25.** The approval period is from **11th February 2025 to 10th February 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

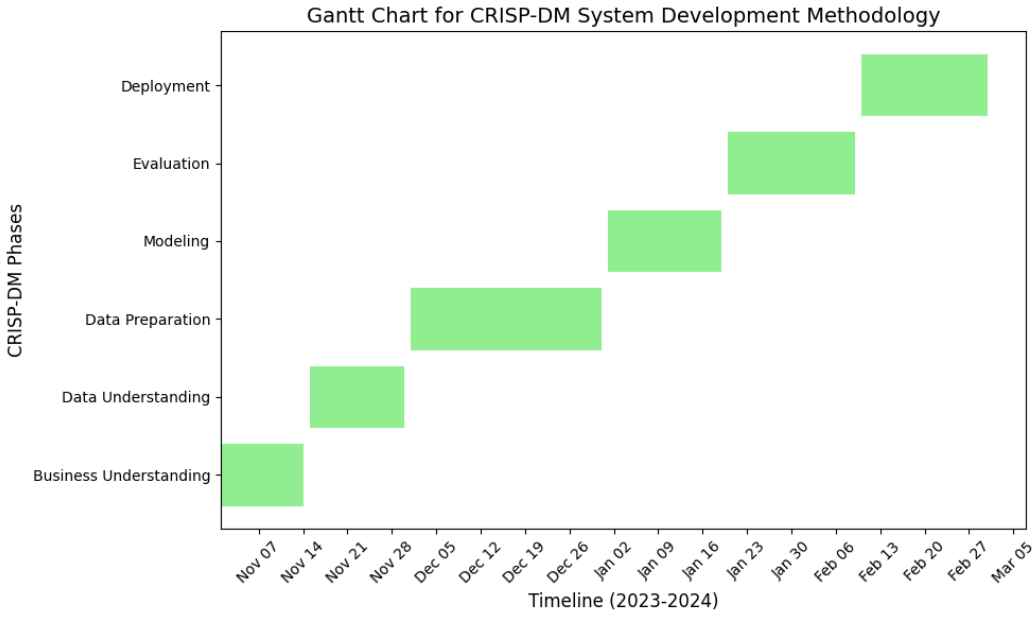
Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

**Mr Ambrose Rachier,**
**Chairperson; SU-ISERC**

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

## Appendix C: CRISP-DM Gantt Chart



**Appendix D: Consent Form and Data Collection Tool.**

**Research on Income Stabilization for Kenyan Coffee Farmers Using Futures Contracts - Interview Guideline**

Greetings and welcome to the Coffee Farmers' Interview. I am Joan Runyiri,a graduate student in the School of Computing and Engineering Sciences at Strathmore University. The objective of my study is to explore how futures contracts can be used to stabilize the income of Kenyan coffee farmers by developing a predictive model to predict coffee prices. Your participation in this interview is voluntary and will take approximately 15-20 minutes of your time.

Please provide responses to the best of your knowledge. There will be no penalty for non-participation, and if you decide to withdraw at any point, please inform me. Your responses will be kept confidential, and only those involved in the project will have access to the data. To ensure anonymity, your responses will be anonymized.Feel free to skip any questions you do not wish to answer. If you have any inquiries, please contact me at joan.runyiri@strathmore.edu. Thank you for your valuable contribution to my research.

**Disclaimer**: The data collected in this interview is strictly for academic research purposes and aims to understand the financial challenges faced by Kenyan coffee farmers. Participation is voluntary, and your responses will be treated confidentially. By participating, you acknowledge that you have read and understood this disclaimer.

**Interview Questions**

1. **Farm Operations and Income Volatility**
    i. What are the primary challenges you face in managing the income from your coffee farming operations?
    ii. How does the fluctuation in coffee prices impact your income, and what strategies do you currently use to manage this volatility?
2. **Understanding of Futures Contracts**
    i. Have you heard of or used futures contracts before to manage price risks? If yes, can you describe your experience with them?
    ii. If not, how familiar are you with the concept of futures contracts, and what are your initial thoughts on using them to stabilize your income?

3. **Current Data Collection Practices**
    i. What types of data do you currently collect related to your coffee farming operations (e.g., yield, costs, market prices)?
    ii. How do you record and use this data to make decisions regarding your farming practices or financial planning?
4. **Perceptions of Financial Instruments**
    i. Besides futures contracts, are there other financial tools or strategies you have considered or currently use to protect your income from price fluctuations?
    ii. What are the barriers that prevent you from using more advanced financial instruments like futures contracts?
5. **Challenges with Hedging Strategies**
    i. What would be your main concerns or challenges if you were to adopt futures contracts as a tool for income stabilization?
    ii. How do you believe futures contracts might help (or not help) with stabilizing your income, based on your current knowledge of how they work?
6. **Improving Decision-Making**
    i. In your current decision-making process regarding the sale of coffee, in which areas would you like more support or insights?
    ii. How do you think forecasting of coffee prices could assist you in making more informed decisions about selling your coffee or engaging in futures contracts?
7. **Adoption of Technology and Predictive Tools**
    i. How comfortable are you with using technology or software tools to help predict coffee prices or manage futures contracts?
    ii. What features would you expect or find most helpful in a system that could assist you in using futures contracts to stabilize your income?
8. **Suggestions for Improvement**
    i. What improvements or tools do you believe could help Kenyan coffee farmers manage price volatility more effectively?
    ii. Are there any specific features or resources you would like to see included in a tool designed to help you manage your coffee farm's financial risks?

**Conclusion**

Thank you for participating in this interview. Your insights and experiences are invaluable and will contribute to the development of tools aimed at stabilizing income for Kenyan coffee farmers. Your feedback will help shape an effective system for managing the financial risks associated with coffee farming.

**Appendix E: Repository for Source Code, Data, and other Artifacts**

Link to GitHub repository: https://github.com/joanrunyiri/PredictiveBrew

## Appendix F: Research Budget

| Category | Description | Month 1 (October) | Month 2 | Month 3 | Month 4 | Month 5 | Month 6 (March) | Total |
|---|---|---|---|---|---|---|---|---|
| Travel and Accommodation | E.g., Data collection: Transportation to the 3 case study sites (2 visits per site) | | Ksh.10,000 | | Ksh.26,000 | | | 36,000.00 |
| Participant Compensation | Compensation for participants involved in the research | | Ksh.3,000 | | Ksh.1,500 | | | Ksh.4,500.00 |
| Materials and Supplies | Computers | | | | | | | 0.00 |
| Materials and Supplies | Specialized hardware required for modelling and simulation | | | | | | | 0.00 |
| Materials and Supplies | Specific IoT devices (list each on its own line) | | | | | | | 0.00 |
| Materials and Supplies | Software licenses | | | Ksh.1,920 | Ksh.1,920 | | | 3,840.00 |
| Materials and Supplies | Access to datasets | | | | | | | 0.00 |
| Publication | Article Processing Charges (APC) for journals or conferences | | | | | | | 0.00 |
| Education | Required online classes and workshops | Ksh. 6,300 | | | | | | 6,300.00 |
| | | | | | | | **Total** | 50,640.00 |
| | | | | | | | | |
| **Indirect Costs** | | | | | | | | |
| **Category** | **Description** | **Month 1 (October)** | **Month 2** | **Month 3** | **Month 4** | **Month 5** | **Month 6 (March)** | **Total** |
| Facilities and Administrative (F&A) | Rental cost of using a makerspace lab | | | | | | | 0.00 |
| Facilities and Administrative (F&A) | Rental cost of using cloud services | | | | | | | 0.00 |
| Facilities and Administrative (F&A) | University library, workspaces, study spaces, lab maintenance, and | | | | | | | 0.00 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | journal subscriptions *(paid through the tuition fee for the thesis/dissertation course)* | | | | | | | |
| IT Infrastructure | University IT infrastructure available to graduate students: Internet, computer networks, electricity, software licenses, printers, and general computing resources available in computer labs *(paid through the tuition fee for the thesis/dissertation course)* | | | | | | | 0.00 |
| Institutional Compliance | Ethical Clearance from Strathmore University Institutional Scientific and Ethical Research Committee (SU-ISERC) *(paid through the tuition fee for the thesis/dissertation course)* | | | | | | | 0.00 |
| Institutional Compliance | Research permit from the Kenya National Commission for Science, Technology and Innovation (NACOSTI) | Ksh.1,000 | | | | | | 0.00 |
| Utilities | Internet | Ksh. 2,500 | Ksh. 2,500 | Ksh. 2,500 | Ksh. 2,500 | Ksh. 2,500 | Ksh. 2,500 | 15,000.00 |
| Utilities | Electricity | | | | | | | 0.00 |
| | | | | | | | **Total** | 15,000.00 |
| | | | | | | | | |
| | | | | | | | **Grand Total (Direct + Indirect Costs)** | 65,640.00 |