# APPLICATION OF MACHINE LEARNING IN ESTABLISHING DETERMINANTS OF GROWTH IN THE HORTICULTURAL EXPORT SUB-SECTOR IN KENYA.

By
Yvonne Juliana Dima Odera
145615


**Master of Science in Data Science and Analytics**

**2004**

# APPLICATION OF MACHINE LEARNING IN ESTABLISHING DETERMINANTS OF GROWTH IN THE HORTICULTURAL EXPORT SUB-SECTOR IN KENYA

Yvonne Juliana Dima Odera
145615

Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Institute of Mathematical Sciences

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 dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.



Signature… …… … ……. Date…28th March 2025

## Approval

The Dissertation of Yvonne Juliana Dima Odera was reviewed and approved by the following:

Signature… …… Date…28th March 2025

Dr. John Olukuru,

Strathmore Institute of Mathematical Sciences,

Strathmore University.

## Abstract

The export industry is key in any country's economic growth (Furuoka, Harvey&Munir, 2019) of any country. The export industry plays a crucial role in a country's economic growth, yet factors influencing its stability and contribution to economic development remain areas of concern. In Kenya, the horticultural sub-sector has experienced slow growth over the past decade, prompting questions about the key challenges affecting its performance. This study aimed to identify factors influencing the growth of horticultural exports and explore ways to enhance the industry's contribution to employment, foreign exchange, and overall economic stability.Using an augmented gravity model, the study analyzed variables such as exchange rates, agricultural GDP, interest rates, climate, trade distance, and preferential trade policies to assess their impact on Kenya's horticultural trade. The findings underscore the importance of monitoring climatic conditions, as they significantly affect production, labor force participation, and economic stability. The results highlight the predictive model's economic significance in shaping GDP, employment, trade balance, and market growth. These insights can guide stakeholders including Policymakers and farmers in making strategic decisions regarding high-value horticultural products, market investments, and effective marketing strategies to drive revenue growth



**Key Terms: Gravity Model, Horticultural Product Code, predictive modelling**

## Table of Contents

**List of Figures**

**List of Tables**

## List of Abbreviations

| | |
|---|---|
| ARMA | Autoregressive moving average |
| COMESA | Common Market for East and Southern Africa |
| CRISP-DM | Cross Industry Standard Process of Data Mining |
| EAC | East African Community |
| EU | European Union |
| FAO | Food and Agriculture Organisation |
| GDP | Gross Domestic Product |
| IMF | International Monetary Fund |

| | |
|---|---|
| PTA | Preferential Trade Agreement |
| OLS | Ordinary Least Squares |
| RCA | Relative Comparative Advantage |
| SHAP | SHapley Additive exPlanations |
| UK | United Kingdom |
| USA | United States of America |
| VAR | Vector Autoregression Model |
| VECM | Vector Error Correction Model |




# Chapter 1. Introduction

## 1.1 Background of the Study

Globalization leading to economic growth has become a primary focus for many policymakers interested when it comes to a nation's economy, trade and growth policies (Sikobi, 2021). Economists in Kenya, by taking similar action to review and use the correlation of exports and economic growth monitor our GDP and thus understand the various

levels of economic growth (Kalaitzi, 2015). In essence countries, which export goods to a larger market can enlarge their gross margins when compared to those that concentrate on building only local markets (Shetty, 2021).

Exports, defined as the movement of goods from one Country to a foreign destination, enable the exporting country to achieve high and sustainable levels of economic growth (Shafiullah & Navaratnam, 2016) and also acquire foreign currency (Ndemo, 2020), thus improving economic development, employment, and trade. By understanding the factors responsible for their variation (Ireen, 2008) one can forecast a Countries economic growth and put in measures to improve it. Some export products include, but are not limited to agricultural goods, livestock, fish, nuts, flowers, among others. It is of importance to mention that the growth in the Agricultural sector in Kenya provides an income rate of over 50% of Kenya's employable population (Ndemo, 2020) as indicated in the chart below.

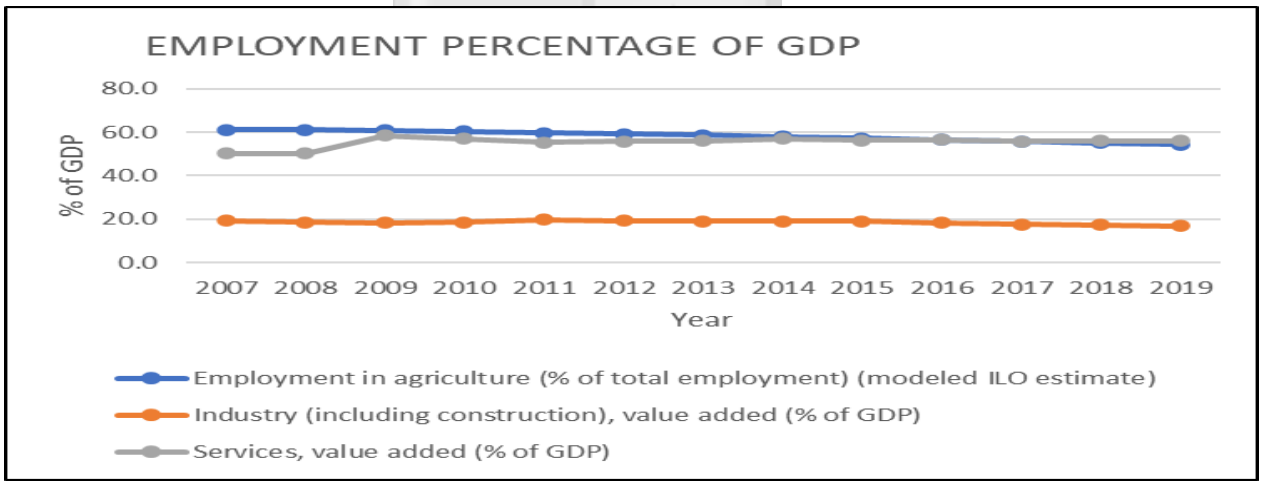



*Figure 1: Employment GDP Graph in Kenya 2007-2019*

The countries' major horticultural exports include flowers, vegetables, tea, fruits, coffee and macadamia nuts, accounting for $5.9B, in 2019(WITS), as illustrated below.

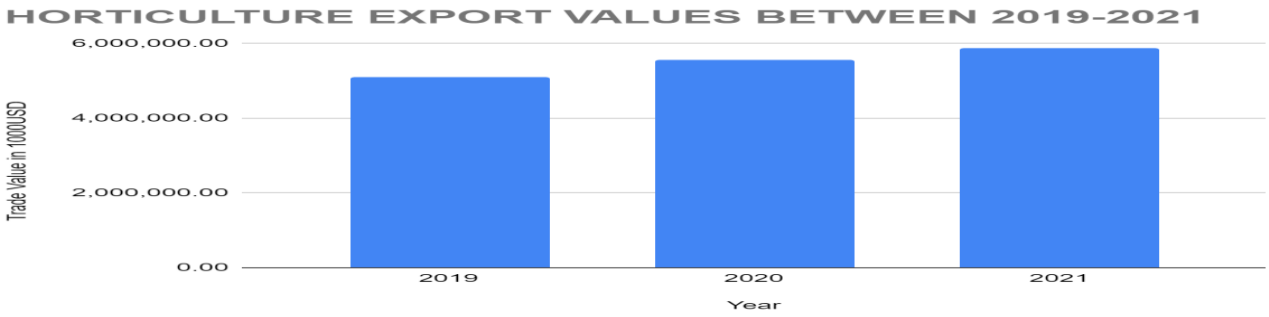



*Figure 2:Horticultural Exports between 2019-2021*

A recent report on product distribution from the Kenya National Bureau of Statistics indicated that fruit and vegetables exports have had a tremendous growth in the past, peaking at 13% as of the beginning of 2021.

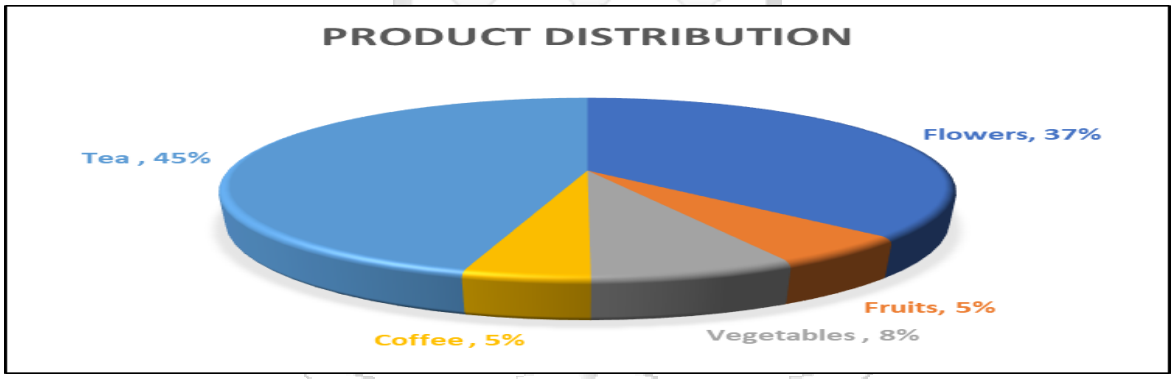



*Figure 3: Export Product Distribution in Kenya*

Studies have revealed that developing and developed countries open to global or international trade, not only accelerate growth in the long run but also are exposed to better investment experiences which enable them to grow their industries (Cipamba, 2012) and thus build their export industries. This has been well illustrated by how the European markets support the developing countries. The higher the external demand for diversified locally produced goods, the higher the economic growth. This in turn shapes the policy direction of many countries which protects these industries (Usman&Landry, 2021) resulting in periodic boom and bust cycles in commodity prices.

## 1.2 Export Industry

Ignorance of the types of commercial industries, unnecessary prohibitions of exports so to preserve a countries market with regard to what it thinks is necessary to preserve its economy is common worldwide given the larger the quantity of any product exported (Hume, 1994), the higher the impact it has in the source country which allows the industries and economies to grow and vice versa. Regarding the effect of exports on economic growth, it has been noted in various studies that for developing countries whose reliance on primary product exports is quite high, it can slow down their economic growth given various determining factors. In contrast, expanding diversified exports can positively and significantly affect the growth of any economy (Kalaitzi, 2015).

There is a debate whether there is a causal link between export and economic growth. Some indicate no causal link (Kalaitzi, 2015) while other explain that there is the causal link between exports and economic growth, giving an example of a review of Asian Countries such as Taiwan, South Korea, Hongkong, and Singapore, showing their growth base on exports as the leading determinant, to rapid growth and continuous growth in GDP which exceeds those of developing countries (Sulaiman&Saad, 2009). Thus, indicating no consensus on whether exports cause economic growth. Ngumi (2009) on a study in Kenya about whether manufactured exports did or didn't lead to significant economic growth, acknowledge that, generally, exports played a role in economic growth. However, this goal, of economic growth continues to be a mirage of a number of African countries.

In Kenya, it is notable that part of the Kenya Vision 2030 (2007) aims, include, among others, increasing agricultural exports to play a critical role in pushing the economy to a 10 percent growth. However, to achieve this, the factors which determine this increased, need to be determined and taken into account so as to achieve this vision in the remaining few years. It is also noteworthy that other countries worldwide, with an emphasis in Africa and South America, work at achieving growth in these industries based on reliance with trade partners, of a wider base than just their former colonial powers. (Usman&Landry, 2021)

In this paper, factors influencing economic growth included the income status of the countries, their economic development, (Usman&Landry, 2021) climatic changes among others. In addition, Cross-country trade works such brought out by Rwenyagila (2013) and Kaltazi (2015) set the pace for supporting a positive relationship in this regard.

The high sustained economic growth in East Asia's economies was been partly due to its exports industries remarkable high and sustained growth record thus supporting the argument that when one concentrates on production in these industries by adopting superior technologies, creation of employment and specialization, in the long run, the economies will eventually grow.

### 1.2.1 Global Production of Horticultural Products

The significant increase in the production of horticulture globally based on the FAO (2021) report indicates that about 0.8B Metric Tonnes of fruits and 1B Metric Tonnes of Vegetables are produces annually with China, the UK, the USA, Netherlands, and Israel being the lead consumers.

Regarding the floriculture industry, the Netherlands accounts for about 55% of the world's exports ($5.77 billion) while Kenya though exporting to Netherlands only accounts for 6.67% ($725 million), and is in the third position in this sector, Kenya was top in the industry, (Trend Economy, 2021) raising queries on what strategies has Netherlands put in place to be a market leader.

In the last 20 years, Kenya has posted impressive growth in the exports markets with horticultural exports becoming one of the biggest foreign exchanges earner together with tourism (KNBS 2017). Major contributors to the growth and success of the Kenya's horticulture and exports industries include but are not limited to, the favorable climate conditions, ensuring year-round production, foreign investors, good policies, and an expanding domestic market. However, there are some challenges, such as the high cost of inputs, low compliance levels with the market requirements, poor quality materials and seeds, poor quality control for local markets, pests, and diseases, dependency on rain-fed agriculture, and reduced technical support smallholder farmers a primary determinant in this industry (Irandu, 2019). In addition, Kenya has some negative factors affecting this industry such as competition from neighboring countries such as Egypt due to its relatively high production costs, increasing dollar rate and deforestation which is causing dramatic climate changes and

the formal and informal horticultural setups have a multiplier effect (Meme, 2013). Handling these two aspects can lead to improving Kenya's competitive advantage, improve incomes and foreign exchange.

### 1.2.2 The challenges Facing the Export Industry

#### 1.2.2.1 The Market Share

For any industry to be built understanding the challenges specific to it is quite important (Shetty,2021). For the export industry, the right market can make or break a business model thus placing emphasis on the need to obtain relevant data and research platform to help them make data-driven business decisions. In the last five years, Kenya's horticulture sector has lost its market share with a growth rate well below the benchmarked countries. In addition, there is considerable pressure for the gross margins for exporters given the rapid growth in domestic consumption cost for raw materials, as well as labor, water, land, seeds, and others such as regulatory frame works which prevent us from being at par with other countries (Laws of Zambia).

The customs duties and tariffs levied on certain products in one country at time have a similar impact on the same exports in destination countries, double taxation, thus increase the total costs for the export products and affecting the viability of a country's products (Tenhoff, 2014). Factors that also make or break an export business is the Quality standards and regulations in different markets. Taking Kenya as an example, due to standards, Kenya is ranked 41st in the EU (RASFF, 2022) yet is the third highest exports provider of flowers to the EU. The pricing strategy also impacts exporters' competitiveness in global markets causing trade imbalances in countries and their partner countries such as the East African Community (EAC).

Thus, for any Company to grow, they need to scale up to meet global needs (Shetty, 2021) by considering the benefits and challenges of the industry and establishing what outweighs the other so as to invest in it, with biggest benefit being tax rebates and value added tax exemptions.

### 1.2.2.2 Exports Diversification

One of the earliest discoveries in the economic development literature shows that the degree of specialization and diversification in production and trade structure determines countries growth and development (Hodey, 2013). In addition, for developing countries, Hodgey argues that, their dependence on the production and export of primary commodities exposes them to adverse external shocks which subsequently slows down growth.

However, export diversification focusing on major trade sectors and partners, increases the range of goods and services offered by an economy as exports. Thus, to grow and cushion oneself from economic shocks this would entail a country moving from the exporting one or a few primary commodities to more comprehensive sets of manufactured goods and services to more significant number of specific destination countries (Usman&Landry, 2021). This increase in trading partners, types of exported goods and, new global value chains, enables domestic firms become globally more competitive through technological transfers. There are also efficiency gains which foster private investment and expanded a Country's insertion in the global market. Kenya's current export areas are illustrated as indicated below

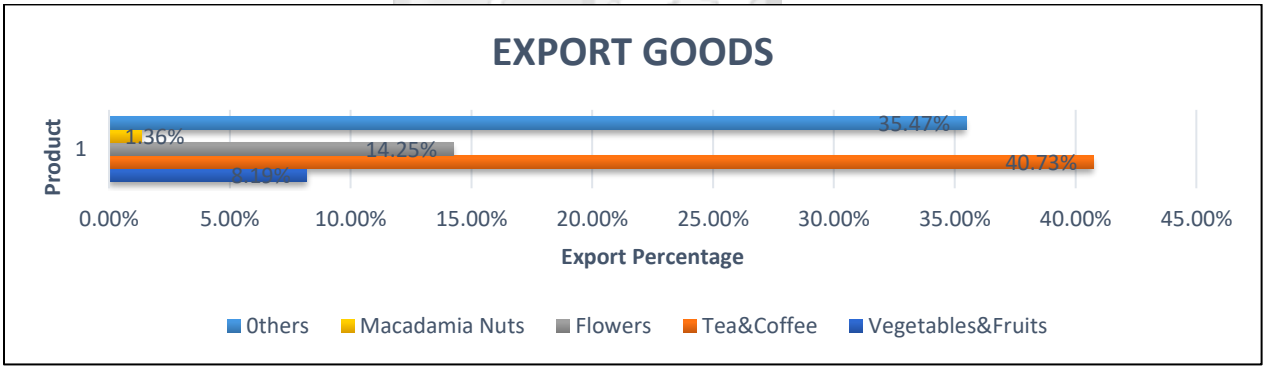



*Figure 4: Export Goods in Kenya*

## 1.3 Kenya's Comparative Advantage Viz-a-Vie its Top Exports

For many Countries, their production and exchange of commodities is based on their relative comparative advantage in the market. For Kenya, any export it makes using the minimum possible opportunity costs gives it a comparative advantage given that the costs of imports are low. Statistics shown that, Kenya's exports are generally agricultural goods (labor-intensive),

while imports are capital goods (capital-intensive), with Kenyan exports coffee and tea exports leading worldwide. The question, however, remains, with which other export products in our new export goods diversification areas can we as a country create an RCA.

## 1.4 Justification of Research/Statement of the problem

Justification of concentration on horticultural exports as a key determinant towards economic growth have been a significant debate issue in Kenya over time as well as the impact of how its improved performance would be critical to macroeconomic growth and enhancing account balance (Rwenyagila, 2013). It is, however, notable that although there is satisfactory performance of exports in some countries, there are countries that have had minimal economic growth with exports. Also, in the Kenyan context, research has shown that growth based on exports only is viable only in the short run (Muhoro, 2012). Thus, the question remains whether, given the critical nature of economic diversification, do different export products and the expansion of the horticultural industry in particular, genuinely lead to economic growth.

Given the several bilateral and multilateral trade agreements that Kenya has entered into and the policy formulation and implementation this has not been a determinant in our trade performance where Kenya's export volume and value performance has remained poor over the years. The USAID competitiveness report (2012) indicated that Kenya's horticultural exports only accounted for 5% of its local production. Based on this, over-reliance on the domestic market, import demands which have continued increasing, climaxing to about Ksh.35.9B in September 2022 (KNBS) and Kenya’s failure to exploit the foreign market it has resulted in low prices and foreign currency thus high exchange rates.

The study developed a machine learning analytical model, that can contribute towards analyzing factors that influence the growth of the horticultural sector and generate insights that could enable growth of the sector based on making more intelligent industry goals.

## 1.5 Research Objectives

The study sought to integrate a machine learning framework by understanding the influencers of growth in the horticultural exports thereby establishing critical ways to improve it as a source of income generation, a source of employment increase and foreign exchange earner.

### 1.5.1 Specific Research Objectives

The specific research objectives of the study were to:

i. Determine the factors that influenced the horticultural sectoral growth
ii. Identify key horticultural subsectors that could be concentrated on based on their growth levels.
iii. Investigate country destinations to which Kenya has unrealized export potential.
iv. Recommend policies for improvement of the horticultural export sub-sector based on the results.

### 1.5.2 Research Questions

Based on the above specific research objectives, the study aimed to address the following questions:

i. What unexploited export areas can further be utilized?
ii. What factors influenced the horticultural subsector?
iii. Which were the best subsectors in the horticultural industry that could be harnessed on based on their growth levels?

## 1.6 Significance of the study

Agriculture is one of the key sources of economic growth on the Kenyan economy with the horticultural sector being one of the leading subsectors for income generation. A close study of the industry and the factors which positively or negatively affect its performance, through this model, would help identify ways to improve it. It would also provide a tool that could be transferable to the Public and Private sectors to enable them easily identify these factors that contribute towards growth in the export market as well as suggest policy change formulation.

## 1.7 Scope and Source of Data for the Study

The study analyzed the impact of exports on economic national growth, comparing Kenya to various international countries based on bilateral trade flows to Kenya valued in US Dollars (US$). It covered types of products and quantities of exports, imports, labor force variables, foreign exchange rates, economic growth rates, climate change data among others.

## 1.8 Contribution and Limitation of the study

The study examined the relationship between exports and economic growth and established a forecasting machine learning model that enabled discovery of modes of export diversification and ways of improving economic growth. It was however limited by availability of data and time frame covered.

### 1.8.1 Dissemination of the study

The target audience for this model and the eventual website was students, policy makers and export businesspersons by raising awareness on various determinants of the success of the horticultural business beyond those commonly known, mainly supply and demand. It also was a means by which we could educate the public on methods of quick market analysis which ensure profits during different seasons in a year based on different determinants.

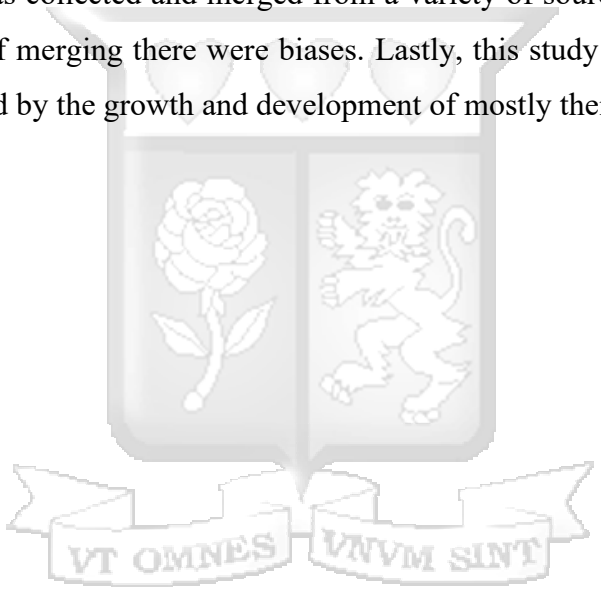
For the application, a viable product approach was used where based on available market data and as well as feedback and data inputted by the clientele in the system, market forecasting and market analysis was provided. The application will then in future be flexible to create new parameters based on audience feedback.

The main methods of disseminating the results of this study was through creation of an application software dashboard, presented to students and individuals upon approval, followed by an online journal article upon approval. The dashboard in future would be available to the public through subscriber application. To measure the success of the study would be through continuous monitoring the site databases to check usage logs and evaluate

presentation sessions feedback through completed evaluation questionnaire. In addition a final stage would be augmentation of the study through partnership with policy maker or other interested parties in seeking opportunities that can help me disseminate the research findings to new or wider audiences and creation of follow-up research projects.

### 1.8.2 Limitation of the study

There were, however, several limitations. Limited data availability, for the given period. Secondly, the data was collected and merged from a variety of sources, thus for the current study data, because of merging there were biases. Lastly, this study was limited to external trade which is affected by the growth and development of mostly their internal economies.




## Chapter 2: Literature Review

### 2.1 Introduction

In export trade there are two factors, namely, domestic factors, such as a natural conditions, economic conditions, trade policies; and global factors such as international market demand changes and economic growth. The relationship between these influencing factors among others in the horticultural exports sector is nonlinear and thus affected by issues such as

random, complex and nonlinear factors with biases. Due to this, using simple algorithms in modelling predictive analytics is not possible, (Dabin, Zhu Hou, & Jingguang, 2009).

## 2.2 Theoretical Literature Review

When it comes to establishing determinants of exports there can never be a unifying theory as was noted by Orindi (2011). However, based on many demand and supply chain theories, time series models have been presented in several literature reviews based on volumes between traded between Countries. The study reviewed some of these models in an attempt to determine the best approach.

### 2.2.1 Revealed Comparative Advantage

Revealed Comparative Advantage (RCA) model measures a country's relative competitiveness in exporting different goods using trade patterns, differences in relative costs, and other factors not related to prices. As an index, this would be consistent regarding changes of relative factors such as productivity. It can be used to establish the potential commodity and loss-making products or zones. Thus, this is a key model in visually capturing the feasibility of a country's expansion into new product areas or service industries. This is however handicapped by data quality.



### 2.2.2 Cobb-Douglas Production Model

For long run tests to establish relationships between actual exports and economic growth, models based on the cointegration test could be best used (Sulaiman&Saad, 2009). One of these models is the Cobb-Douglas Production Model. In this model it indicates that the revenue in an economy is determined by its technology, capital, labour and imported inputs.

In addition to the above, It, is assumed that the exporting sector also augments productivity in the non-export sector through technological spillovers, based on the proportion of exports volumes (Wamalwa & Were ,2019). Given that the exporting sectors are more innovative than local industries due to exposure and risks in the export market, productivity differences may

exist thus, influencing output growth. However, time yields, competitiveness resulting in productivity gains among others should also be factored in and the model amended.

In the linearization of this model, one can use linear logarithms as demonstrated by Nyasula (2013), in his study. To cater for limitations such as unquantified variables, proxies were ideally taken for respective items. This model is good for determining the relation between exports and economic growth with the actual individual impacts and influences of other essential factors. The limitation is that it cannot capture random disturbance terms (μ) or biases.

### 2.2.3 Estimation Techniques

A third option that this paper considered was the existence of a relationship between two variables in the long and short term using the Granger Causality Test or the Vector Error Correction Model (VECM) an econometric model used to analyze the long-term equilibrium and short-term dynamics between multiple time series variables, particularly when those series that are cointegrated

#### 2.2.3.1 Testing for Granger Causality (GC)

In definitive testing of the causal relationship between variables, the Granger causality test suggested by Granger (Granger, 2001) is a good option. This is a test where the dependent variable is regressed on the other variable based on a lag of the dependent variable. A more recent and robust test for causality is the block homogeneity Wald test (Enders, 2008). The main purpose of these tests is to evaluate whether the lags that one variable causes to other variables are essential. Using the null hypothesis for this test, one variable's lags can eventually be excluded from the equation. Thus, when one rejects the null for the equation it suggests that the variable is endogenous.

This test is good at establishing the existence of short-term equilibrium relationships between real exports and economic growth which influence productivity or output. However, it should be noted that causality is implied, not controlled for other explanatory variables. For a dual analysis of the long-term and short-term relationship between economic growth and exports, the vector error correction model is preferable.

The limitation of this causal link relationship is that it is can only best used for long-term behavioral relationship tests because these require estimation techniques appropriate for long-run. For this to happen the system must be tested for cointegration before testing for Granger causality (Ngumi, 2009).

#### 2.2.3.2 Vector Auto-regression Model and VECM Techniques

As a generalization of the Autoregressive Moving Average (ARMA) model, for one to forecast interrelated time series data systems and analyze the dynamic impact of random disturbances on the system of variables, the vector autoregression (VAR) is the most appropriate. It assumes that the future values in time series data are linearly related to the past values of multiple time series data (Gunathilaka &Tularam, 2016). Given the relationships between different exports and the economic growth rate, this model can be used to estimate the relative effects of these variables on one another using impulse responses and variance decomposition functions. It is however limited in that it does not tell the existence of a long-run relationship or the adjustment of the short-term to long term equilibrium. It also does not provide a sufficient dynamic specification that identifies all these relationships of all variables which may appear on both the left and right sides of equations. Based on these limitations, there was a need for an alternative, non-structural approaches for the estimation and analysis of errors.

## 2.3 Empirical Literature Review

Several regional and global phases have examined the factors used to determine trade flows between different countries such as the Gravity Model. This model, when applied to international trade, indicates that trade flow between two countries can be measured based on their economic size, distance, trade volumes, income sizes, population sizes, exchange rates among others (Lien, Feng, & Fei, 2019, Tinbergen, 1962, Pöyhönen, 1963 and Linnemann, 1966).

To augment this model to one can consider whether a country is a Member Of COMESA, EU, or had an American Embassy as factors that influence export levels, Orindi (2011) developed an augmented gravity model using the OLS technique and panel data. This study proved that in general the distance of a country from Kenya did indeed affect the volumes of exports given transportation costs. However, its limitation was that it did not narrow down to which subsectors were most affected. In addition, factors such as Foreign Direct Investments are a good addition to this model (Rutto, Odhiambo, Obange & Enock, 2019)

Karamuriro and Karukuza (2015) noted that given the augmented gravity model, the impact of other factors could also be established. Such as the GDP per capita income of the importing Country, language as a dummy, existence of a common border among others variables. The test revealed that sharing a common language, as well as a common border could enhance trade in some instances.

Further in 2019, by adding the importers income and exchange rates, Lien, Feng, & Fei (2019) noted that in the augmented gravity model developed for China in the agricultural subsector the fluctuations in the exchange rate negatively impacted the total exports, and the Country's general income. However, it was also noted that increases in production lowered exports, contrary to what was expected, thus revealing a need to moderate production to reap maximum benefits.

Ndemo (2020), in reviewing the export of horticultural tea from Kenya using the gravity model, added variables such as population and whether the Country was a World Trade Organization (WTO) Members among others. Using the OLS with dummy variables as the regression model, he noted that the fixed effect model had no significant value and was therefore not considered explanatory. When it came to the random effect model, it was noted that when the population increased, Kenya's trade Partners decreased their demand of tea exports. Another negative and significant impact was the GDP of the importing Country.

## 2.4 Overview of Literature Review

From the literature review it was noted that a challenge that has been experienced in establishing a model that could best be used in establishing factors that affect the economic trade in the horticultural industry which takes into account both the fixed and random variables. In addition, establishing parameters for growth based on dynamic factors through nonstructural approaches has also been a challenge when the OLS approach is used. This approach usually negates the studies by indicating the nonstructural variables have no significant effect on trade. Thus, a model that correctly captures all these variables id required.




# Chapter 3: Research Methodology and Data

## 3.1 Introduction and Methodology

Chapter 3 explained in detail the steps that were covered to attain and accomplish the research objectives in chapter one of the study. It outlined the research design and methodology that was used to achieve the objectives of the study which were to establish the top selling horticultural goods, unexploited export areas, factors that influence the horticultural subsector

and best subsectors in the horticultural industry that can be harnessed on based on their growth levels; To develop a model for identifying factors that influence the horticultural subsector, based on insights from historical data.

An experimental research design approach was used to establish a cause-and-effect relationship by manipulating one or more independent variables in the case factors affecting the horticultural goods sub-sector while establishing their effect on a dependent variable. This study focused on the factors affecting the horticultural values based on historical horticultural export data including climate, population, economic trade and labour force dynamics.

### 3.2 Research Design

The research adopted the Cross Industry Standard Process for Data Mining (CRISP-DM) Model, which offers a methodical, structured way to manage different data mining tasks. It makes it possible to comprehend the underlying facts at a deeper level. The CRISP-DM Model comprises six distinct stages that were followed as represented below.


*Figure 5: CRISPM-DM Model*

The first stage of CRISP-DM methodology was a business understanding stage where an understanding of the project's requirements and objectives that clearly defined its success criteria were established. The process then moved to data understanding which drove the focus to identifying, collecting and analyzing data sets. In this stage, initial data that is collected is examined, explored and verified for quality. Data preparation, the third stage, involved preprocessing and data transformation A comprehensive data preparation and preprocessing process was performed. This included cleaning the data, inspecting and exploring the data, handling missing data, data type conversion and transformation. After good exploration of the data was done, the data was modelled for a data output in the CRISP-DM cycle. The last two stages involve evaluation of the model and deployment respectively marking the end of the cycle of CRISP-DM methodology.

### 3.2.1 Business Understanding

The business case explored was the analysis of the horticultural export industry over a period

based on historical performance. This analysis was key in understanding the top horticultural goods, unexploited export areas, factors that influence the horticultural subsector and best subsectors in the horticultural industry that can be harnessed on based on their growth levels which can enable expansion of the horticultural industry in particular, genuinely leading to economic growth

### 3.2.2 Data Understanding

The study used secondary data sourced from various open sources such as the United Nations Statistics Division, IMF Database, and the World Bank's World Integrated Trade Solutions (WITS) Online Database, Central Bank of Kenya, Infotrade Website, OkoaMombasa Shipping Data, The Kenya National Bureau of Standards. The sample coverd 23 years, from 2000 to 2022, consisting of import data to Kenya imports and export data of horticultural export goods to all countries, labour force statistics for all countries, climate change data, dollar exchange rates in Kenya over the year, economic growth rate of world countries and the foreign exchange rate data among other variables.

The study employed bilateral trade (import) data referencing from United Nations (UN) Comtrade website for the several horticultural commodities. UN Comtrade provides data using several nomenclatures, based on the Standard International Trade Classification (SITC) Revision 1 classification as this classification features data with the longest possible timeframe, i.e. from 1962 for some commodities. Specific tariff data are obtained from the UNCTAD Trade Analysis Information System (TRAINS) database as follows:

*Table 3:1 Goods Description and Tariff Codes*

| **Tariff Code** | **Description** | **Tariff Code** | **Description** |
|---|---|---|---|
| 603 | Cut flowers and flower buds | 802 | Other nuts, fresh or dried, whether |
| 709 | Other vegetables, fresh or chilled | 807 | Melons (including watermelons) |
| 803 | Bananas, including plantains, fresh | 808 | Apples, pears and quinces, fresh. |
| 804 | Dates, figs, pineapples, avocados | 810 | Other fruit, fresh. |
| 806 | Grapes, fresh or dried | 902 | Tea, whether or not flavoured |

Various sets of data were considered for analysis based on historical data. These data sets provided different features which were used to create the final sample dataset. Below is a table showcasing the data distribution of the initial data collected.

*Table 3:2 Raw Horticultural Goods Subsector Data*

| **Data Set Name** | **Years** | **Row** | **Column** | **Column Description** |
|---|---|---|---|---|
| Statistics Dataset | 2000-2022 | 2000 | 35 | ['EC_Code','ProductCode', 'Exporter'x', 'IC_Code', 'Importer', 'Year', 'TradeValue in 1000 USD', 'ProductDescription', ImportRegion', 'Quantity', 'QtyUnit', TradeFlowName', 'ExporterRegion', 'Country','Temperature','CO2 Emissions', 'Sea Level Rise', 'Precipitation', 'Humidity', 'Wind Speed', 'day', 'month', 'year', 'Country.1','C_Code', 'City', 'lat', 'lng', 'Year.1', 'Diesel_Fuel_per_L','Air_Transport', 'Road_Sea_Transport', 'Train_Sea_Transport', 'Region_Road_Transport'] |
| I_country codes | n/a | 240 | 3 | ['Country', 'I_Code', 'Code'] |
| Climate Kenya | | 46 | 11 | ['Country','Temperature','CO2Emissions', 'Sea Level Rise', 'Precipitation', 'Humidity', 'Wind Speed', 'day', 'month', 'year', 'id'] |
| Climate_sep | 2000-2020 | 4010 | 6 | ['Date', 'temperature', 'humidity', 'windspeed', 'Rainfall', 'Month'] |
| World GDP | 2000-2022 | 220 | 26 | ['Country Name', 'Country Code', 'Indicator Name', '2000', '2001', '2002', '2003', '2004', '2005', '2006', '2007','2008','2009','2010','2011', |

| | | | | |
|---|---|---|---|---|
| | | | | '2012', '2013', '2014', '2015', '2016', '2017','2018','2019','2020','2021', '2022'] |
| Country codes | n/a | 249 | 3 | ['Country', 'E_Code', 'Code'] |
| CO2 Emissions | 2000-2020 | 21 | 6 | ['year', ‘methaneEmissions_CO2_equivalent_kt', 'greenhouse_gas_emissions_CO2_equivalent_kt', 'CO2Emissions_solidfuel_consumption_kt', 'CO2Emissions_liquidfuel_consumption_kt', 'CO2Emissions_kt'] |
| Countries Distance | n/a | 244 | 4 | ['Country', 'C_Code', 'lat', 'lng'] |
| Town Elevations | n/a | 63 | 3 | ['Town', 'lat', 'lng'] |
| Kenya Imports | 2000-2023 | 200000 | 10 | ['Nomenclature', 'ReporterISO3', 'ProductCode', 'ReporterName', 'PartnerISO3', 'PartnerName', 'Year', 'TradeFlowName',’TradeFlowCode', 'TradeValue in 1000 USD'] |
| Compiled Exchange Rates | 2001-2024 | 24 | 6 | ['Year', ' Qtr1 ', ' Qtr2 ', ' Qtr3 ', ' Qtr4 ', ' AvgDollarRate '] |
| Distance Costs | n/a | 244 | 11 | ['Country',‘C_Code','Distance_from_Kenya_km','Train_Msa','Road_Border_port','Port_Handling_20FT', 'Port_Handling_40FT','Entry_Clearance','Entry_Clearance_EAC','distance_costs', 'C02_Cost_kt'] |

| | | | | |
|---|---|---|---|---|
| Exports Kenya | 2000-2021 | 1000 | 12 | ['id','EC_Code','ProductCode', 'Exporter', 'IC_Code', 'Importer', 'Year', 'TradeValue in 1000 USD', 'ProductDescription','ImportRegion', 'Quantity', 'QtyUnit'] |
| World Labourforce | 2000-2022 | 5405 | 4 | ['Country Name', 'Country Code', 'Year', 'labourforce'] |
| PTA Agreements | 2000-2021 | 750 | 4 | ['Year', 'Country', 'PTA', 'Country_Code'] |
| Kenya Labour Force | 2000-2022 | 23 | 4 | ['Year', 'Country Code', 'Unnamed: 2', 'ke_labourforce'] |
| Yearly Climate Amended | 2000-2022 | 23 | 8 | ['Year', 'temperature', 'humidity', 'windspeed', 'Rainfall','temperature_celcius',‘avg_temp_celcius', 'avg_humidity_mm'] |

### 3.2.3 Data Preparation

This is a crucial step in research, and what follows extraction is pre-processing and transformation which includes checking for completeness, consistency, accuracy, then cleaning the data cleaning in readiness for modeling. The individual 17 datasets were initially preprocessed in excel using the data manipulation inbuilt in-built library in to ensure they were reliable and complete. These were then transformed further in python to merge the relevant fields from the different data sets once each was cleaned.

*Table 3:3 Transformed Data*

| **Data Set Name** | **Transformation** | **Output** | **Software Used** |
|---|---|---|---|
| Statistics Dataset | *Renamed Columns<br>*Selected a new suitable of Kenya export data containing 2000 rows and 13 columns<br>*Missing values were removed reducing the rows to 1922<br>*The time frame covered is reduced to 2000 to 2018<br>The data was filtered to consist of only Kenya Export data per year to various countries for only horticultural goods. | Data filtered to include only Kenya exports to various countries by Horticultural Value per year for the period 2000 to 2018 | Python |
| I_country codes | *Merged with the previous data set so as to assign the data to the United Nations Statistics 3-digit numeric codes (e.g. Kenya is 254) and the International Standards Organization 3-digit alphabetic codes (e.g. Kenya is KEN)<br>*This increased above set to 15 columns and 1922 rows<br>*Renamed I_Code to ImporterCode and dropped all null values<br>*Outliers treated<br>*Float data converted to integers | Standardized so as to analyze based on international data | Python |
| Climate Kenya | *Column names renamed to remove spaces | Dropped because of limited data | Python |

| | | | |
|---|---|---|---|
| World GDP | *World GDP data merged with Country codes data<br>*Column names standardized to match export column names namely Country and IC_Code<br>*Data types Changed and Outliers treated<br>*Kenya data was filtered and the rows were transposed into two columns with Year and GDP<br>*World GDP data transposed with year and country codes as columns<br>*Merged Kenya Export data with Kenya GDP<br>*Merge the above merged Kenya data with the transposed world GDP data | *Standardized for further analysis<br>*World GDP and Kenya Exports dataframe created | Python |
| Country codes | *Some Countries which were recently recognized such as South Sudan Codes were updated into the above data set | Updated merged data frame to ensure all countries have the accurate Statics and Country Codes | Python |
| CO2 Emission | *Dropped null values<br>*Handled Outliers<br>*Year column converted to an integer<br>*World GDP and Kenya Exports merged with the emission data based on the Year<br>*Only greenhouse emissions and other emissions are allowed. Methane, solid fuel and liquid fuel are dropped | Data frame including Kenya horticultural goods exports, CO2 emissions and GDP created | Python |
| Climate Kenya | *Date converted into datetime format<br>*NumPy data converted to pandas data frame<br>*Outliers removed | The data frame including Kenya horticultural goods exports, CO2 Emissions, | Python |

| | | | |
|---|---|---|---|
| | *Climate data added to the data frame created above | GDP and climate variables was created | |
| Countries Distance | *Using the folium database and Kenya as the base country, the Haversine function was used to establish the distances of countries to Kenya using the Latitude and Longitude<br>*Columns 'lat', 'lng' and 'distance_to_next' dropped<br>*This table is transposed and concatenated with the above to include the distance to each country from Kenya | Obtained distances between countries in Kilometers<br>Updated data frame to incorporate country distances. | Python |
| Town Elevation | Using the API Key the elevation of towns based on latitudes and longitudes was established | Town elevations established | Python |
| Kenya Imports | *Trade value converted to millions of dollars<br>*Columns were renamed to standardize with the Export/GDP/Climate Conditions data frame<br>*Kenya Import Values is added to the above merged data frame | Standardized to analyze based on international data | Python |
| Compiled Exchange Rates | * The Qtr1, Qtr2, Qtr3, Qtr4 columns were dropped<br>*AvgDollarRate column name renamed to remove spaces<br>*The Exchange rates were merged with the Imports data frame. | A data frame including Kenya horticultural goods exports, CO2 emissions, GDP , climate variables, Kenya Import Values and Average dollar rates in Kenya is created | Python |

| | | | |
|---|---|---|---|
| Distance Costs | *The distance costs and C02_Cost_kt were added to the data modeled for analysis | Data frame including Kenya horticultural goods exports, CO2 emissions, GDP , climate variables, Kenya Import Values, Average dollar rates, distance costs and C02_Cost_kt in Kenya is created | Python |
| Exports Kenya | *Trade value converted to millions of USD<br>*Export Values-Columns dropped to a scaled data set with 7 columns<br>*Export Quantity-Columns dropped to a scaled data set with 8 columns<br>*Data separated with product tariff/code | Analysis data frame above enhanced to include product codes/tariffs, export quantities from Kenya for each tariff to specific Countries and the Horticultural Values | Python |
| World Labourforce | *The world labour force data was added to the analysis data frame based on the Country code and the year | Enhanced Analysis dataframe | Python |
| PTA Agreements | *The world trade agreements information with Kenya were included in the data frame in a categorical manner based on the Country code and the year | Enhanced Analysis dataframe | Python |
| Kenya Labour Force | *Kenyan labour force a key determinant in any trade activity is also added to augment the data based on the year | Enhanced Analysis dataframe | Python |
| Yearly Climate Amended | *Given the outliers detected in the climate values, the temperature and humidity values are averaged on a yearly values are updated on the table then merged onto the dataframe that will be used for analysis. | Completed analysis dataframe | Python |

#### 3.2.3.1 Data Cleaning

The final raw dataset was processed using python to understand its contents using the pandas library by ensuring all variables were in the correct format, the columns were properly labelled, missing values were handled among other cleaning processes. The main objective was to ensure that the analytical dataset was reliable and complete and could be used for exploration. The analytical data set consisted of 374,294 rows and 25 columns. The data set types consisted of mainly float and integer type of data. There were on average 88,517 missing values under Product Code, Quantity and Horticultural Value. The following steps were included in the cleaning process:

i. The rows with missing values null values in the columns Product Code, Quantity and Horticultural Value were dropped given these were the determinants of the study and could not be guessed. If there was no record of the horticultural good exported through product code, the data was not required.
ii. When dropped the missing values reduced to 422 under Kenya Import Values and two (2) under the International Labour force Column. Where the labour force was missing it was established to be in the year 2022 for Iran and Cote D'Ivoire. These were filled using the average increasing rate per year. The Kenya import values were filled with zero because these needed to be factual and if not provided cannot be guessed.
iii. Categorical columns created based on the Horticultural Product Code and converted to Boolean format of 0 or 1 to indicate if the export values refer to that Commodity code or not. This increased the columns to 29 from 25.
iv. With all missing values handled the rows decreased from 374,294 to 285,777.

### 3.2.4 Data Exploration

The research used Matplotlib, Seaborn and Plotly to analyze the data and to understand patterns, detect any anomalies such as outliers and summarize key insights primarily through visualizations. Below is a graph shot to represent the data exploration process with python:

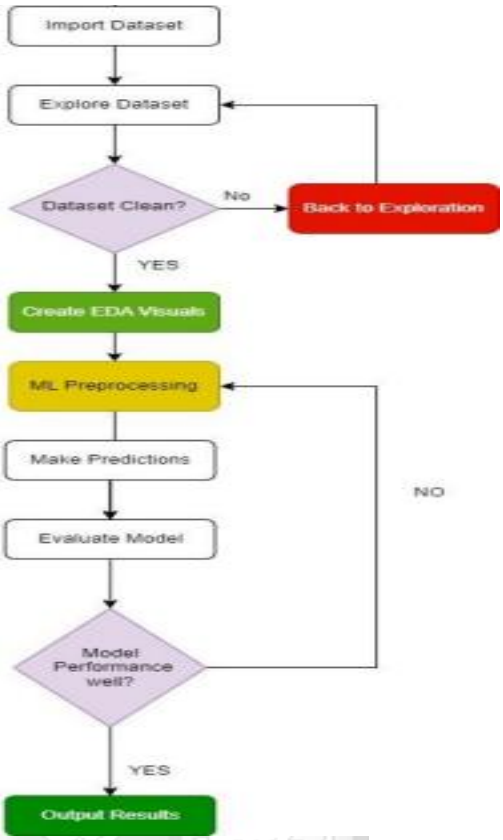



*Figure 6: Exploratory Data Analysis Process*

### 3.2.5 Modelling

The research concentrated on building a predictive model to estimate trade values based on factors such as exporter, importer, product type, transport mode, and economic indicators. The modeling process involved several key steps, starting with data preprocessing, where we handled missing values, encoded categorical variables, and standardized numerical features. This step ensured that our dataset was clean and suitable for machine learning algorithms.

The dataset was then split into training (2019 and prior) and testing sets (2020-2022), ensuring that the model could learn from historical data while being evaluated on unseen data. For the modeling phase, multiple machine learning algorithms were tested

### 3.2.5.1 Model Architecture

The study employed a combination of machine learning models to develop a predictive system that provided accurate estimates for trade values. The architecture below illustrates the workflow, from data preprocessing to model training and evaluation, ensuring robust predictions aligned with the study's objectives. Beginning with a Linear Regression model as a baseline. While this provided initial insights, to further enhance the model, we used hyperparameter tuning via GridSearchCV, optimizing factors such as the number of estimators in the Random Forest model. The study finally improved performance by implementing a Random Forest Regressor, which better captured non-linear relationships in trade data.

The machine learning models utilized include Random Forest, Lasso Regression, and Gradient Boosting, each contributing uniquely to the predictive accuracy. Random Forest built multiple decision trees, aggregating their outputs to enhance their accuracy while reducing overfitting. When the Lasso Regression was applied, on the other hand, it shrank coefficients of less significant features to zero which was especially useful for feature selection and improving model interpretability. Lastly, Gradient Boosting sequentially built trees, corrected errors from the previous ones, and thus resulted in a highly accurate but computationally intensive model that required careful tuning.

### 3.2.5.2 Model Implementation

In developing a robust predictive model for trade and economic factors, feature selection was guided by an augmented gravity model framework. The gravity model explains trade flow between countries by incorporating economic size, distance, and additional influential factors. This model was extended by including variables such as GDP, exchange rates, trade agreements, climate-related factors, and sector-specific influences like horticultural type of goods. The model takes the form

$$T_{ij} = \propto \left( \frac{GDP_{it}^{\beta_1} \times GDP_{jt}^{\beta_2}}{D_{ij}^{\beta_3}} \right) \quad (1)$$

$T_{ij}$ $-Trade\ flow\ between\ country\ i\ and\ j$

Taking natural logarithms for equation (1) gives the following basic gravity model:

$$lnT_{ijt} = \beta_0 + \beta_1 lnGP_{it} + \beta_2 lnGP_{jt} - \beta_3 lnDT_{ij} + \mu_{ijt} \quad (2)$$

Where;

$T_{ijt}$ is the total value of trade flows between country $'i'$ and country $'j'$ at time $t$,

GP is the size of the economy,

$DT_{ij}$ is the distance between the capital centers of different countries,

$\beta_0$ $(ln \propto)$ is the gravitational constant,

and $\beta_1$, $\beta_2$, and $\beta_3$ are model coefficients to be estimated.

$ln$ and $\mu_{ijt}$ denotes the natural logarithm operator and the stochastic (error) term, respectively.

To optimize the model performance in determining Horticultural Values, log transformation was applied to critical numerical features, ensuring a more normalized data distribution and improving model interpretability. Key transformed features included GDP, greenhouse gas emissions, exchange rates, labor force size, and climate variables like temperature and humidity. The transformation was performed using the natural logarithm function to mitigate skewness and ensure consistent scale across variables.

$$lnHV_{ijt} = \beta_0 + \beta_1 lnGDP_{it} + \beta_2 lnKE\_GDP_{jt} - \beta_3 lnD_{ij} + \beta_4 lnLabourForce_{it} + \beta_5 lnEX_R + \beta_6 lnPTA_{ijt} + \beta_7 lnGreenhouse\_Emmissions_{ijt} + \beta_8 lnTemperature_{ijt} + \beta_9 lnHumidity + \beta_{10} lnHortiProductCode\_X + \beta_{11} lnCountry\_Code + \beta_{12} lnKE\_LabourForce_{jt} + \beta_{13} lnRainfall_{jt} + \beta_{14} lnKEN\_ImportValues_{jt} + \beta_{15} lnKEN\_ExportValues_{jt} + \beta_{16} lnQuantity + \mu_{ijt} \quad (3)$$

Whereby

$\beta_0$ is the constant,

$\beta_i$ ; where $i$ = 1, 2, …, n-1

and $\gamma_i$ ; $where\ i$ = 1, 2, …, n are model coefficients to be determined,

while $\mu ijt$ is the error term

The training and testing process leveraged machine learning models, specifically XGBoost and Random Forest regressors, to predict trade values effectively. The models were trained on a refined dataset where extreme outliers were excluded. Evaluation metrics such as $R^2$ and RMSE were employed to assess model accuracy, ensuring reliable trade flow predictions under varying economic and environmental conditions. The augmented gravity model's inclusion of additional trade determinants, coupled with advanced feature engineering, enhanced the model's predictive power and real-world applicability.

### 3.2.6 Model Evaluation

The model evaluation was performed using cross-validation, a robust method for assessing model performance on previously unseen data. By dividing the dataset into multiple folds, each acting as a test set only once, this approach ensured a more reliable estimate of the model's generalizability. Cross-validation helped reduce the risk of overfitting, ensuring that the model did not simply memorize the training data but performed well on new inputs. Although cross-validation prediction was not explicitly used to compute evaluation metrics, it was instrumental in forecasting observations and validating model performance.

To further evaluate the models effectiveness, structured trade datasets were used for training and testing. Hyperparameter tuning was applied to optimize each model for improved accuracy. The regression models were assessed using $R^2$ score, Mean Squared Error (MSE), and Mean Absolute Error (MAE), ensuring precise trade value predictions. Meanwhile, classification models were evaluated using Accuracy, Precision, Recall, and the F1 Score to measure their ability to correctly classify trade-related categories. These combined evaluation strategies provided a comprehensive understanding of model performance and reliability.

### 3.2.7 Deployment

The models will be consumed using API and Project Analysis presented using Power BI.

## 3.3 Ethical Considerations

Security, data privacy, and user-friendliness were prioritized in the deployment plan to protect sensitive data and offer a seamless user experience to all users. By using this approach, the study's results were maximized, and a data-driven culture is used in analyzing market trends in the horticultural subsector.

## 3.4 Conclusion

Conclusively, the methodology used to analyze data on horticultural goods trade took a multifaceted approach that integrated statistical and machine learning techniques. This included the application of the augmented gravity model, XGBoost, and Random Forest regressors to predict trade values while incorporating advanced feature engineering. Additionally, interactive visualizations and data applications enhanced accessibility and real-time analysis, ensuring practical insights for stakeholders in the horticultural trade sector.

# Chapter 4: Results

## 4.1 Introduction

Chapter 4 presents a comprehensive analysis and interpretation of the research findings on the horticultural goods subsector, focusing on key products such as tea, flowers, avocados, pineapples, vegetables, and other fruits. This study aimed at identifying the factors that influenced sectoral growth, as well as highlight the untapped market potentials, and determine which subsectors offer the greatest opportunities for expansion. Through a detailed examination of the data, this chapter provides valuable insights into the economic and trade dynamics of the horticultural industry, offering evidence-based recommendations for stakeholders. In addition, by assessing market trends, production patterns, and trade flows, this analysis contributes to a deeper understanding of the subsector's role in economic development, export diversification, and global competitiveness.

The following sections present the results using relevant data visualizations and statistical analyses, ensuring a clear and comprehensive representation of the study's outcomes. The methodology used included a thorough exploratory data analysis (EDA), which set the groundwork for our subsequent modelling efforts.

## 4.2 Data Analysis Results & Interpretation

### 4.2.1 Business Understanding

The analysis began with general horticultural goods exported and imported around the world by all countries. The horticultural goods internationally use the Standard International Trade Classification as listed below:

| HortiProductCode | ProductDescription |
|---|---|
| 810 | Other fruit, fresh. |
| 804 | Dates, figs, pineapples, avocados, |
| 902 | Tea, whether or not flavoured. |
| 709 | Other vegetables, fresh or chilled. |
| 603 | Cut flowers and flower buds of a ki |

*Figure 7: Horticultural Goods Codes*

The top importers of horticultural goods and top imported horticultural goods, by Trade Value in (‘000)USD are:

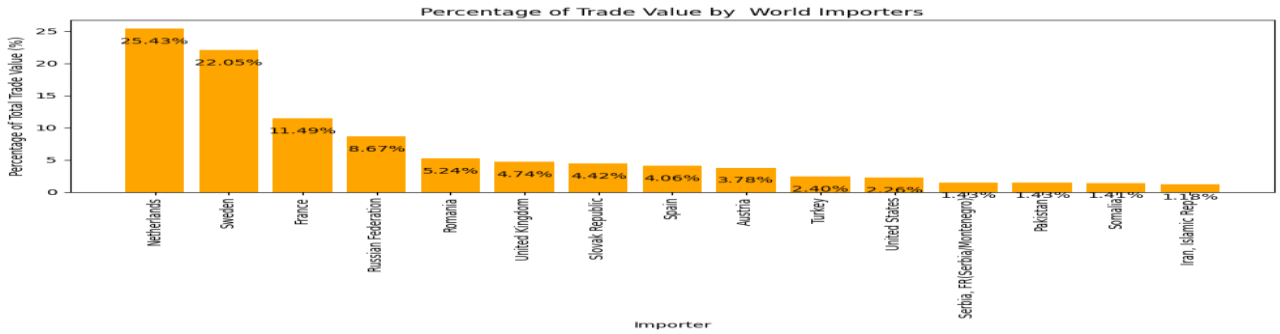



*Figure 8: Leading Importers of Horticultural Goods Worldwide*

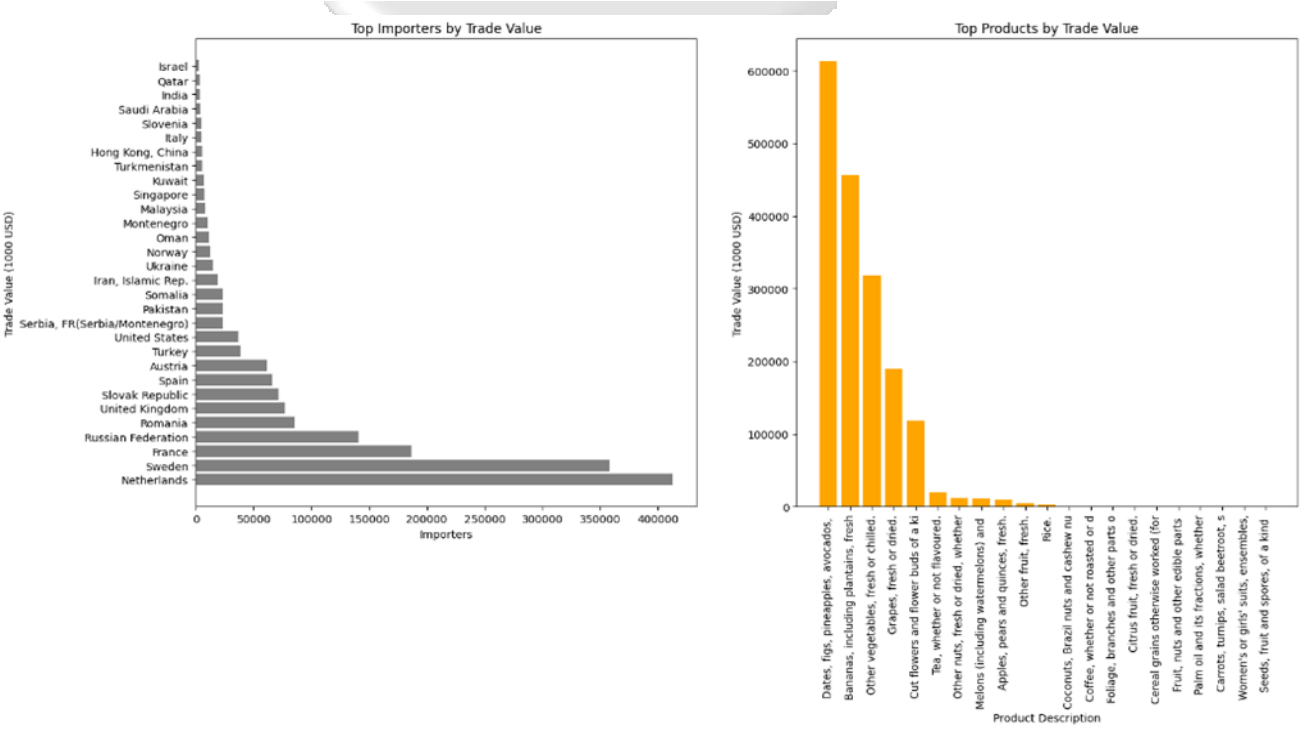



*Figure 9: Top Importers of Horticultural Goods*

However, considering that Values are at times based on exchange rates and other variables, in terms of count, the number of times each country has indicated it imported a horticultural good, these are the top imported goods:

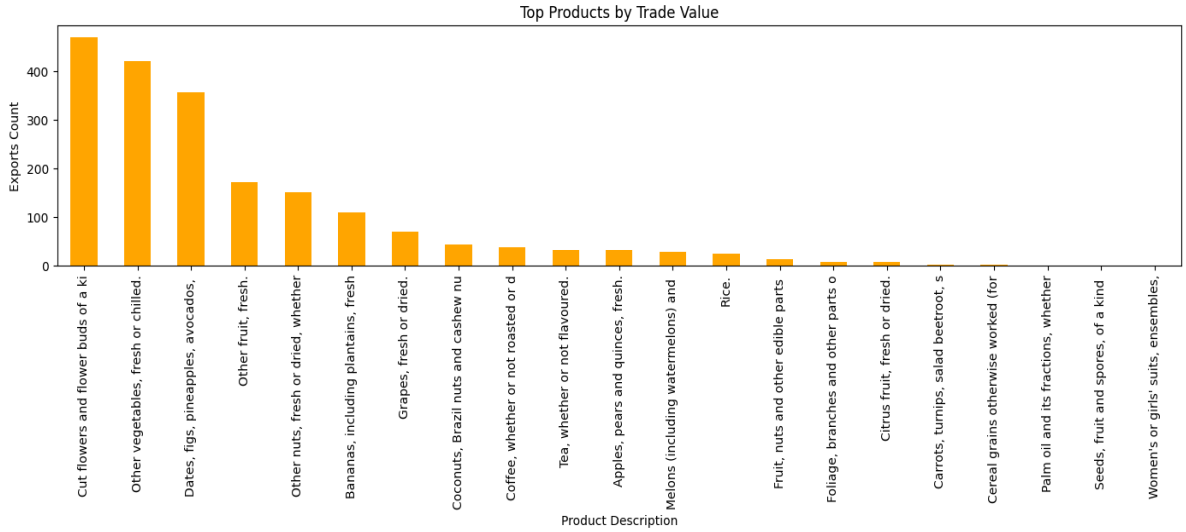



*Figure 8:Top Imported Horticultural Good In Terms of Count*

Top exporters of horticultural goods and top imported horticultural goods by Trade Value in (‘000)USD

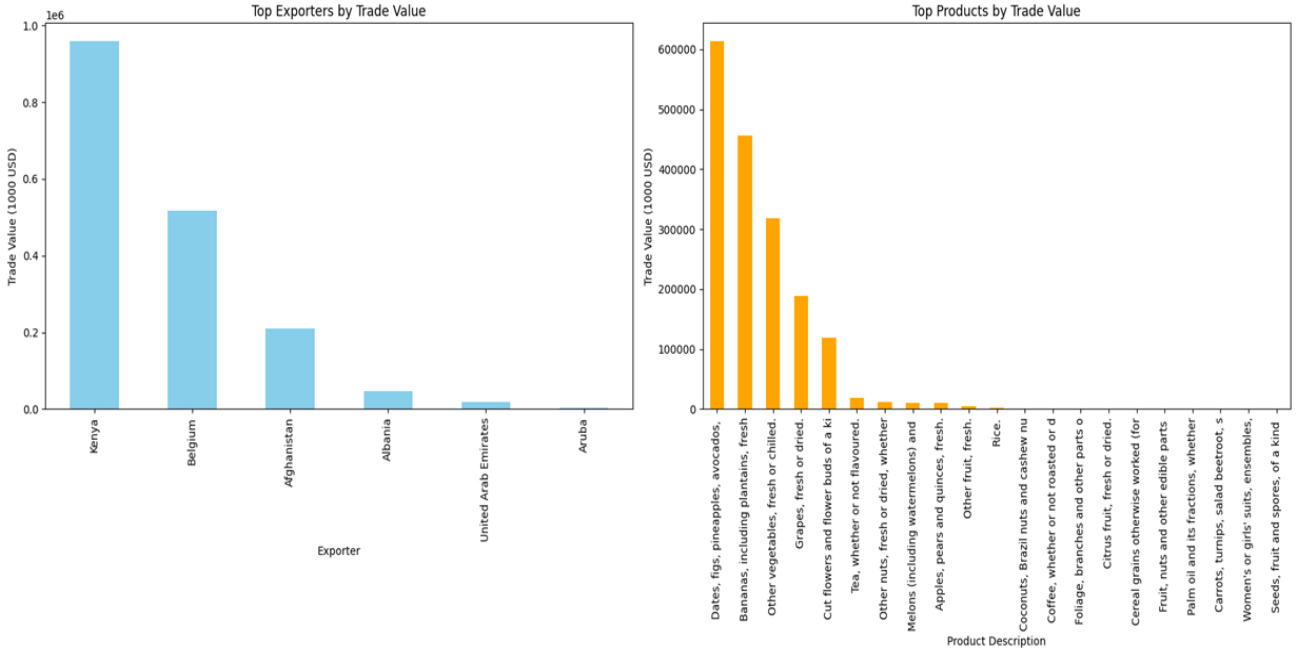



*Figure 9: Top Exporters of Horticultural goods*

The top importer of horticultural goods by Trade Value in (‘000)USD, exported from Kenya and their markets are:

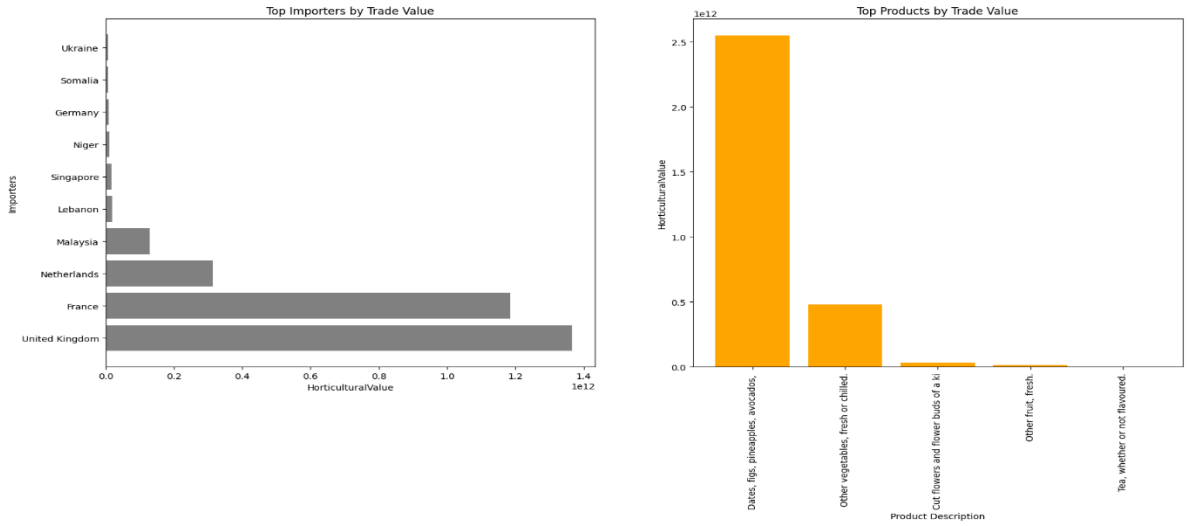



*Figure 10: Kenya's Export Market*

Based on the codes in Figure 7, The percentage value of the horticultural goods based on exports from Kenya are illustrated below:

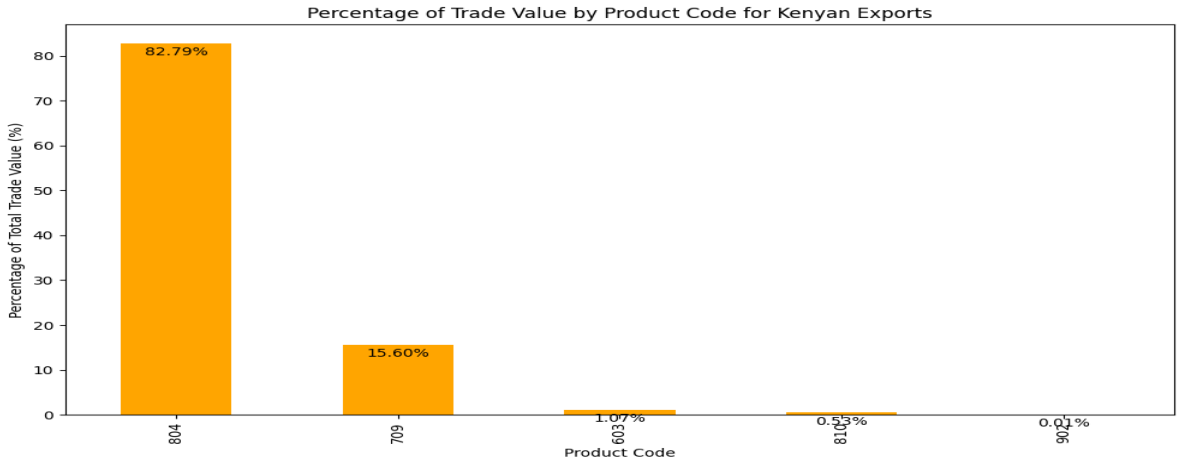



*Figure 11: Percentage Ratio of Exportation of Horticultural Types of Goods*

A review of the top markets Kenyan Horticultural goods based on values over the period 2001 to 2021 indicate that the strongest markets are the United Kingdom followed by France as per the correlation matrix below

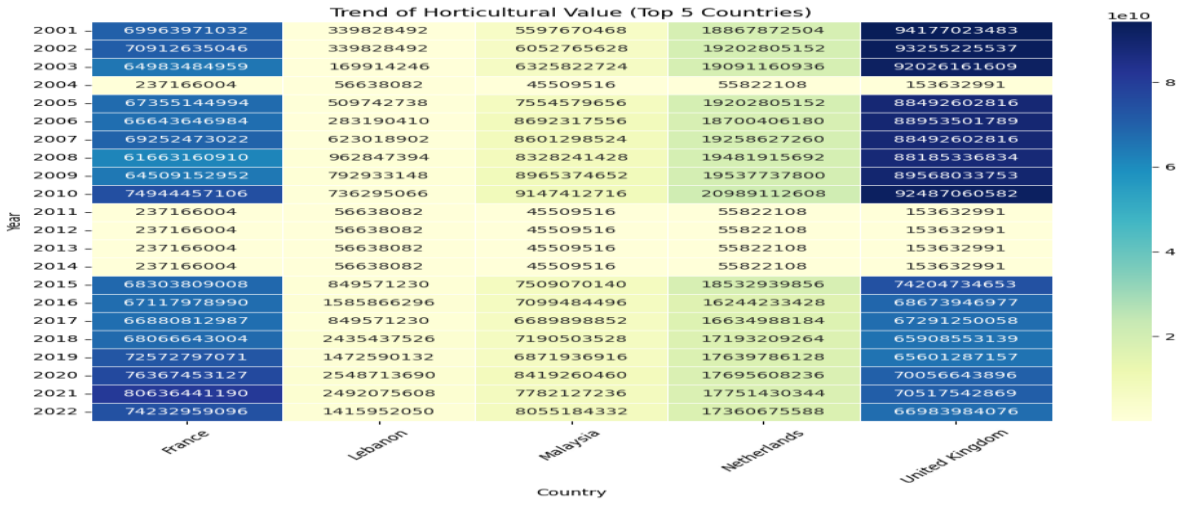



*Figure 12: Top Markets for Kenyan Horticultural Goods Based on Values in Kshs.*

### 4.2.2 Correlation Matrix on Features affecting the Horticultural Market

The augmented gravity model explained that trade flow between countries is affected by its economic size, distance, and additional influential factors such as GDP, exchange rates, trade openness, climate-related factors, and sector-specific influences. For this study various factors were included, such as GDP, greenhouse gas emissions, exchange rates, labor force size, and climate variables like temperature and humidity, which likely affect horticultural trade.

Using the Pearson Correlation Matrix, the following insights were established with regards to factors influencing the Kenyan Horticultural Industry generally.

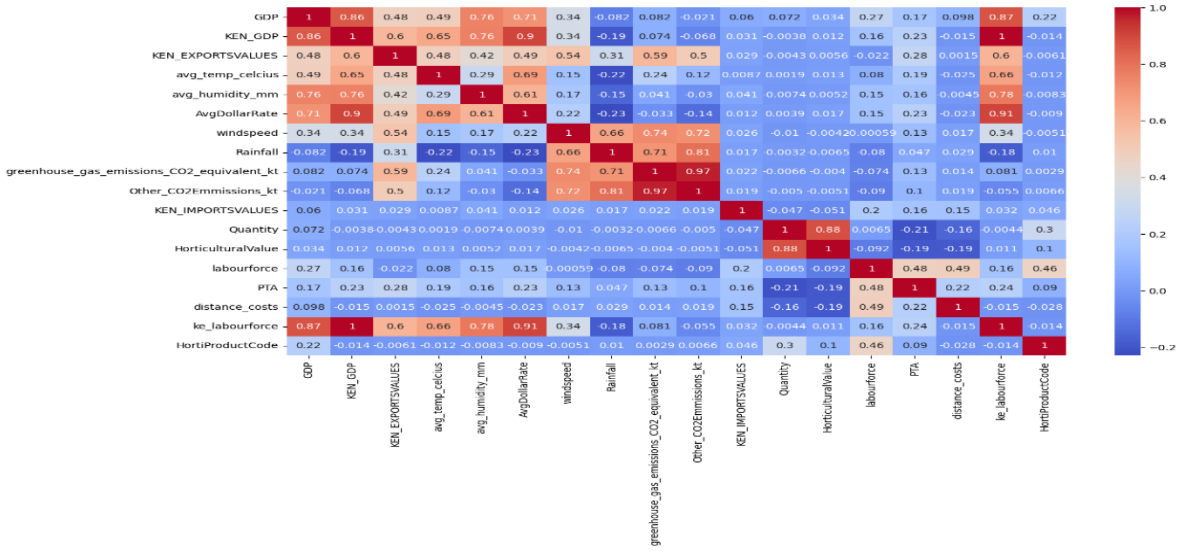


*Figure 13: Correlation Matrix of Factors affecting the Horticultural Industry*

Kenya's GDP is highly correlated (0.86) with overall GDP, meaning our economy moves in line with the global economy. Our Kenyan Labour force is affected by climatic conditions, dollar rates, and international trade because as noted, given this affects our economy. The values of our Horticultural goods are strongly affected by the quantity produced (0.88). The horticultural goods produced is also determined by the international/ global labour force(0.46) given the demand they make. Labourforce and PTA (0.48) have a relatively strong correlation; possibly indicating that larger labour forces are linked to regional trade agreements (PTA - Preferential Trade Agreements).

### 4.2.2.1 Factors Affecting Production of Horticultural Product 603: Flowers

The below results show that there is a strong correlation between humidity (0.78), temperature (0.66) awhen it comes to growing flowers. These also impact the quantities grown. Greenhouse emissions have a strong positive correlation with wind speed(0.77) and rainfall (0.69). This therefore has a huge indirect impact on this industry. PTA (0.46) also has an impact in this industry

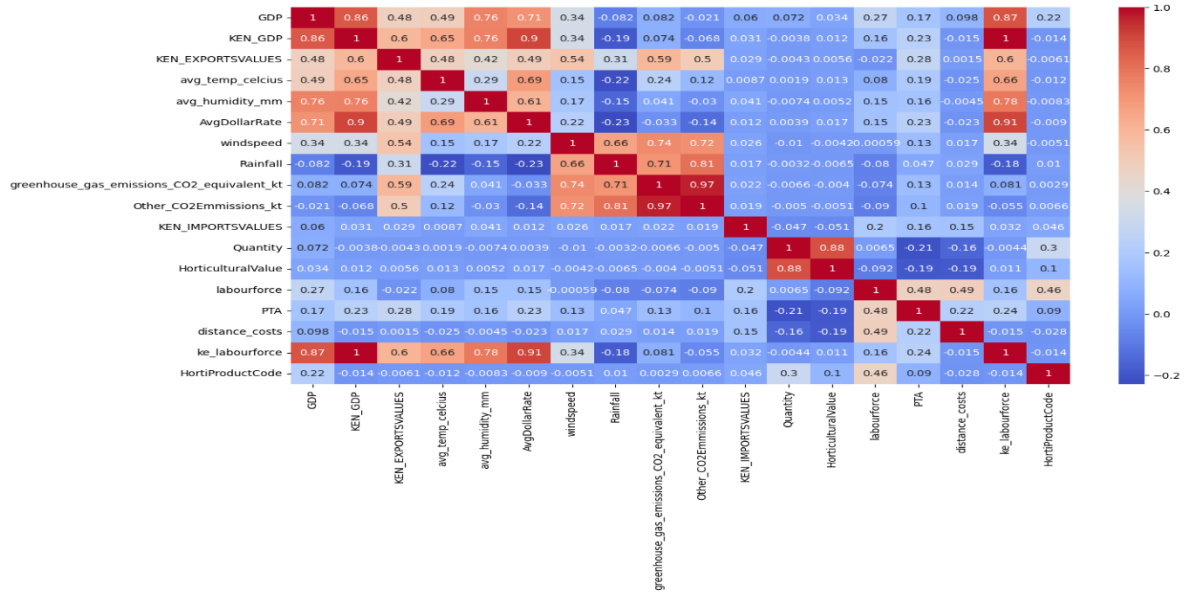



*Figure 14: Factors Affecting Production of Flowers*

### 4.2.2.2 Factors Affecting Production of Horticultural Product 709: Vegetables

The below results show that there is a strong correlation between humidity (0.78), temperature (0.66) and the Kenyan labourforce (0.77) when it comes to growing vegetables, These also impact the quantities grown. Greenhouse emissions have a strong positive correlation with wind speed(0.77) and rainfall(0.69). This therefore has a huge indirect impact on this industry.

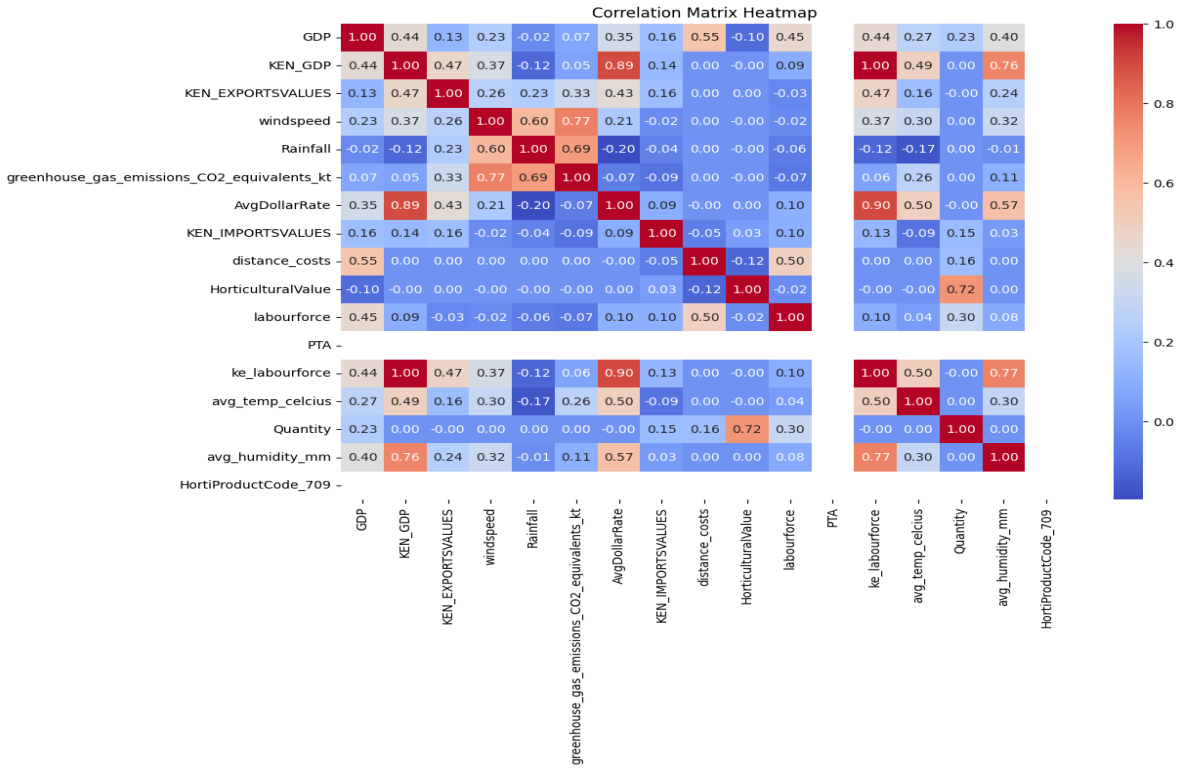



*Figure 15: Factors Affecting the Production of Vegetables*

### 4.2.2.3 Factors Affecting Production of Horticultural Product 804: Fruits

The figure below results shows that there is a strong correlation between humidity (0.77), temperature (0.50), export values (0.24) and the Kenyan labourforce (0.77) when it comes to growing vegetables. The industry is also affected by distance costs (0.25) which in turn are affected by the world economy GDP (0.55). Greenhouse gas emission again plays a role in this industry.

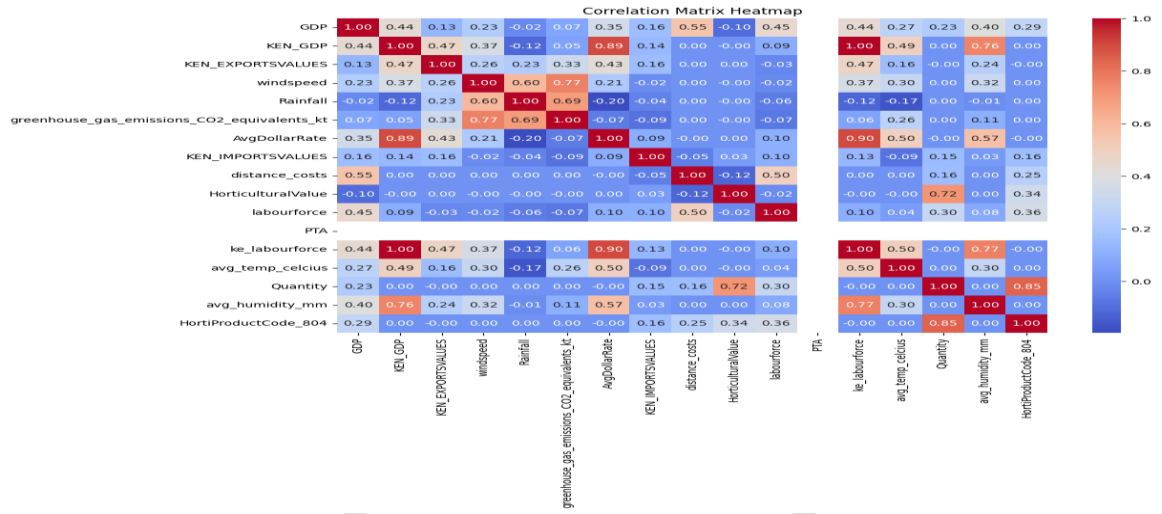



*Figure 16: Factors Affecting the Production of Fruits*

### 4.2.2.4 Factors Affecting Production of Horticultural Product 810: Avocadoes & Pineapples

The figure below results show that there is a strong correlation between humidity(0.77), temperature (0.50), export values (0.24) and the Kenyan labourforce (0.77) when it comes to growing avocadoes. The industry is also affected negatively by rainfall (-0.20). Greenhouse gas emission again plays a role in this industry.

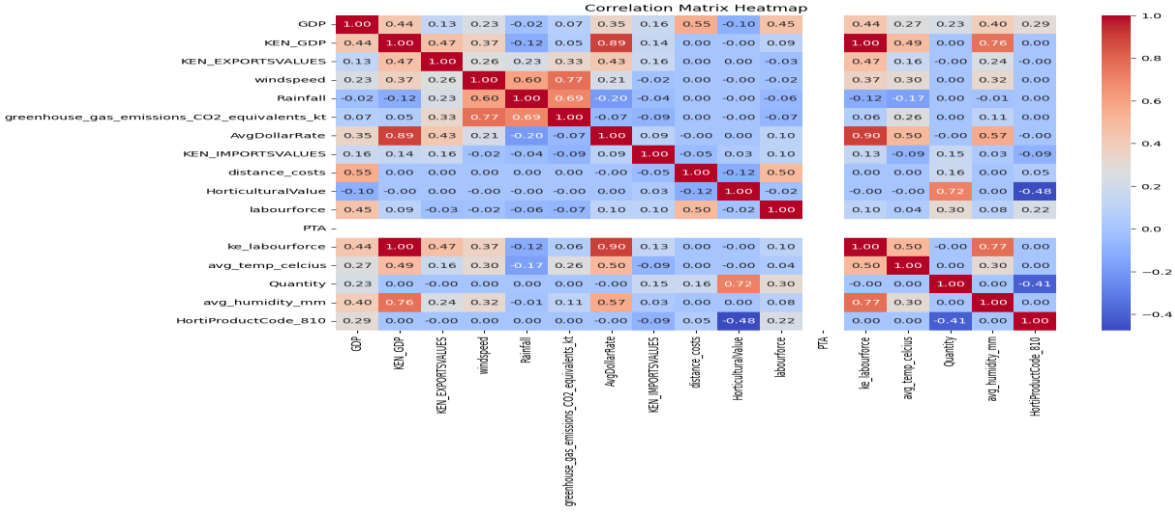



*Figure 17:Factors Affecting the Production of Avocadoes & Pineapples*

#### 4.2.2.5 Factors Affecting Production of Horticultural Product 902: Tea

The figure below results show that there is a strong correlation between humidity(0.77), temperature(0.50) and the Kenyan labourforce (0.77) when it comes to growing tea, These also impact the quantities grown. Green house emissions have a strong positive correlation with wind speed(0.77) but negative with rainfall. This therefore has a huge indirect impact in this industry.

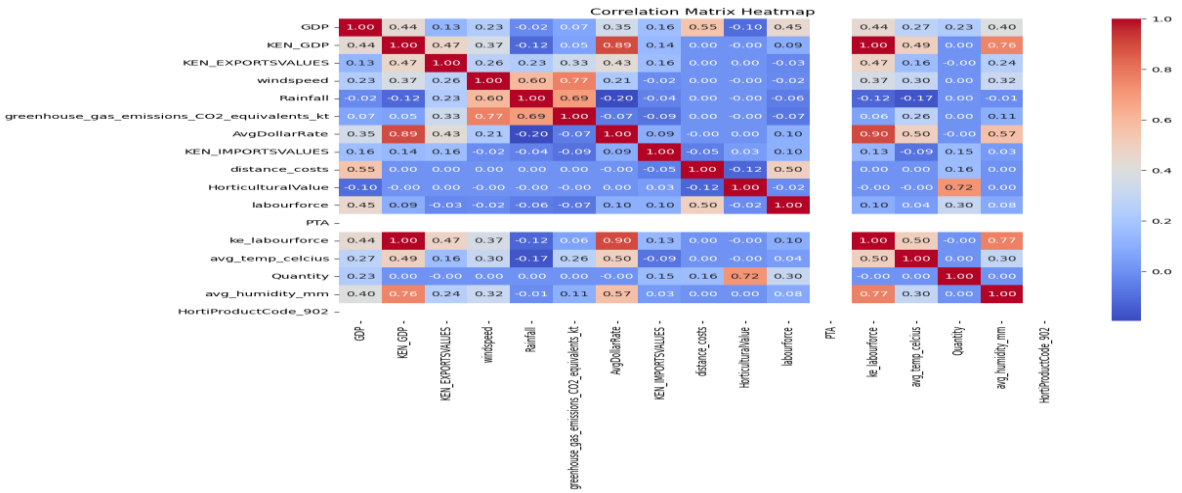



*Figure 18:Factors Affecting the Production of Tea*

### 4.2.3 Feature Selection

In developing a robust predictive model for trade and economic factors, feature selection was guided by an augmented gravity model framework. The gravity model explained trade flow between countries by incorporating economic size, distance, and additional influential factors. This model was extended by including variables such as GDP, exchange rates, trade openness, climate-related factors, and sector-specific influences like horticultural trade.

To optimize model performance, log transformation was applied to critical numerical features, ensuring a more normalized data distribution and improving model interpretability. Key transformed features included GDP, greenhouse gas emissions, exchange rates, labor force size, and climate variables like temperature and humidity. The transformation was performed

using the natural logarithm function to mitigate skewness and ensure consistent scale across variables.

The training and testing process leveraged machine learning models, specifically the XGBoost and Random Forest regressors, to predict trade values effectively. These models were trained on a refined dataset where extreme outliers were excluded. Evaluation metrics such as $R^2$ and RMSE were employed to assess model accuracy, ensuring reliable trade flow predictions under varying economic and environmental conditions. The augmented gravity model's inclusion of additional trade determinants, coupled with advanced feature engineering, enhanced the model's predictive power and real-world applicability.

Different selection techniques were applied such a correlation Analysis used to remove highly correlated variables (multicollinearity issues). Through this technique, features that had high correlation ($r > 0.85$) with each other were either combined or dropped. Secondly, features assigned as important through random forest were selected while those with low importance scores were dropped. For the XGBoost, SHAP values and gain-based importance scores was used. Through this, any features with low importance were eliminated. Another method was the Recursive Feature Elimination (RFE) used on the Random Forest models to iteratively remove least important features until an optimal subset was selected.

The final features selected are illustrated below:

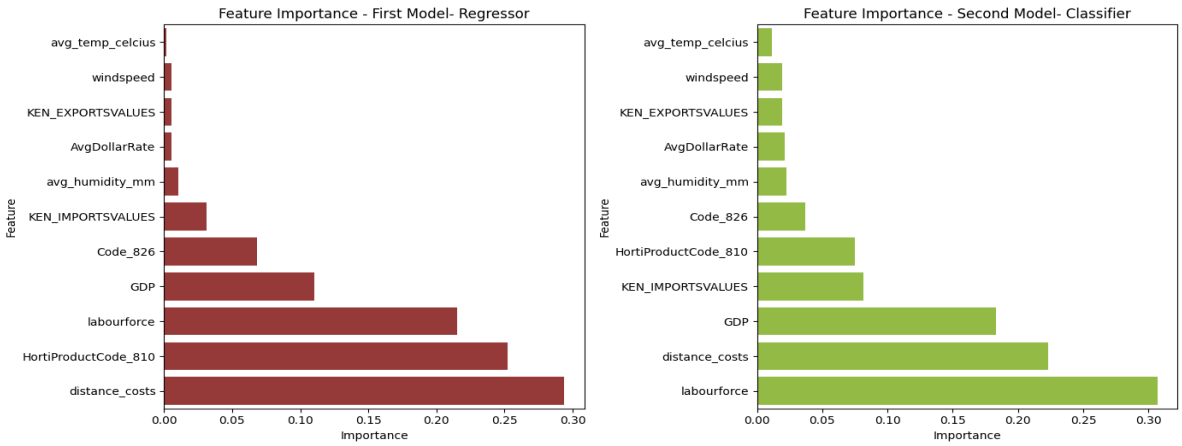



*Figure 19:Important Features based on Horticultural Values*

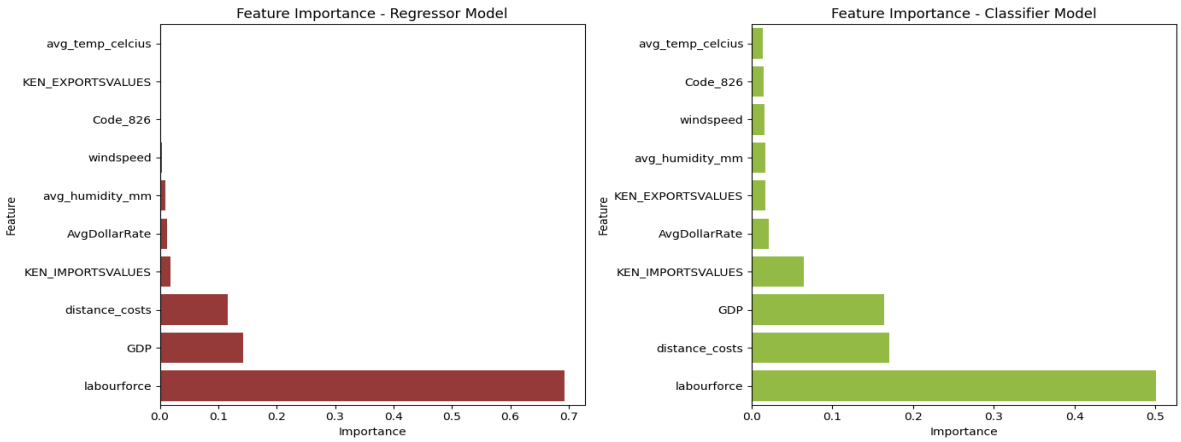



*Figure 20: Sample Important Features Based on Product Code*

### 4.2.4 Modelling

To predict trade values effectively, the machine learning models, captured the complex relationships between economic, environmental, and trade-related variables while handling non-linearity and feature interactions efficiently. This involved Feature Engineering where Log Transformation was carried out on key variables (e.g., GDP, $CO_2$ emissions, exchange rates, labor force size, and climate factors) to correct skewness and improve model interpretability. Next extreme outliers in trade values were removed using interquartile range (IQR) filtering and z-score methods to ensure the models were not influenced by rare but extreme fluctuations. Lastly, the augmented gravity model framework guided the feature selection process which included the inclusion of additional variables beyond GDP and distance, such as climate conditions, greenhouse gas emissions, and labor force size, to better capture trade flow.

#### 4.2.4.1 Random Forest (Multi- Class) Classifier

The Random Forest Classifier was used to categorize horticultural trade values into three bins (low, medium, and high) based on log-transformed and scaled features. This model is based on a robust ensemble of learning method that averages multiple decision trees to reduce overfitting while capturing non-linear relationships in trade data, through, training multiple

decision trees on different subsets of the data. It then averages their predictions to improve accuracy and generalizability. An advantage of this model is that it handled missing values and outliers effectively. This model achieved reasonable accuracy, but feature importance analysis helped refine feature selection.

#### 4.2.4.1.1 Model Training

Once trained, the following visualizations of the predictions and residuals enable us to understand the performance of the models with each horticultural good. For insights on providing insights on how well the models capture patterns in the data the prediction plots compare actual vs. predicted values, while the residual plots help identify any systematic errors by displaying the differences between actual and predicted values. These visualizations were the basis upon which we established model accuracy and detected potential biases.

The trend below indicates the points are not evenly distributed around the red dashed line (zero residuals). There are some outliers, especially at lower residual values (e.g., around -1.5). Possible model bias. Residuals appear to decrease in variance as predicted values decrease. This suggests the model underperforms for lower values.



##### 4.2.4.1.1.1 Horticultural Product 603: Flowers

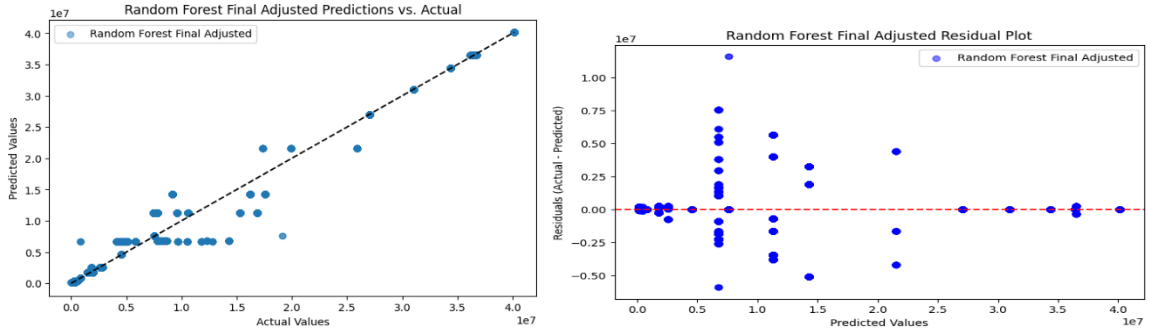



*Figure 21: Visualization of Predicted Values & Residuals on Flowers Model*

#### 4.2.4.1.1.2 Horticultural Product 709: Vegetables

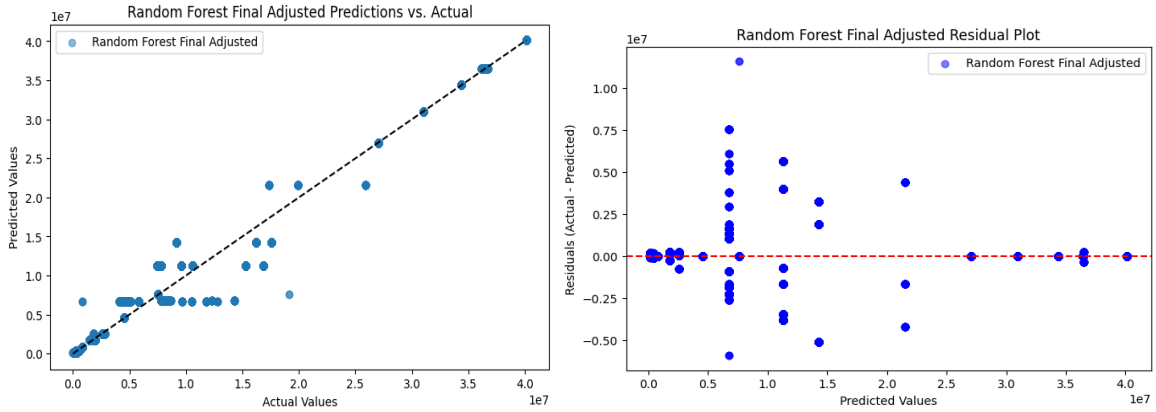



*Figure 22: Visualization of Predicted Values & Residuals on Vegetables Models*

#### 4.2.4.1.1.3 Horticultural Product 804: Fruits

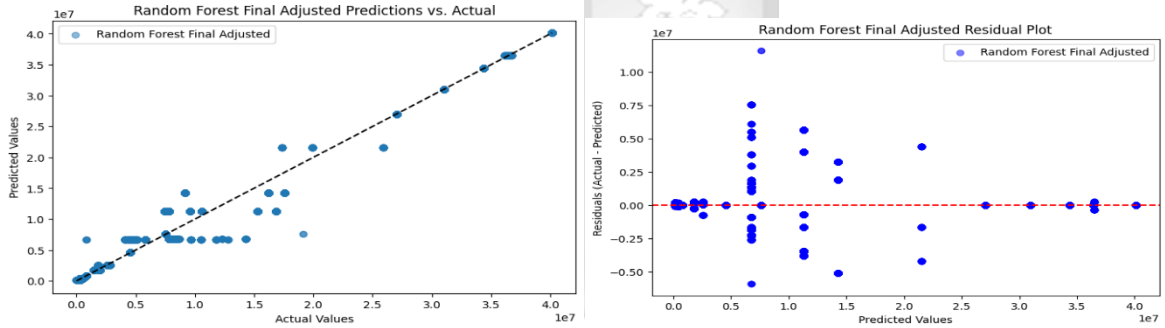



*Figure 23: Visualization of Predicted Values & Residuals on Fruits Models*

#### 4.2.4.1.1.4 Horticultural Product 810: Avocado & Pineapple

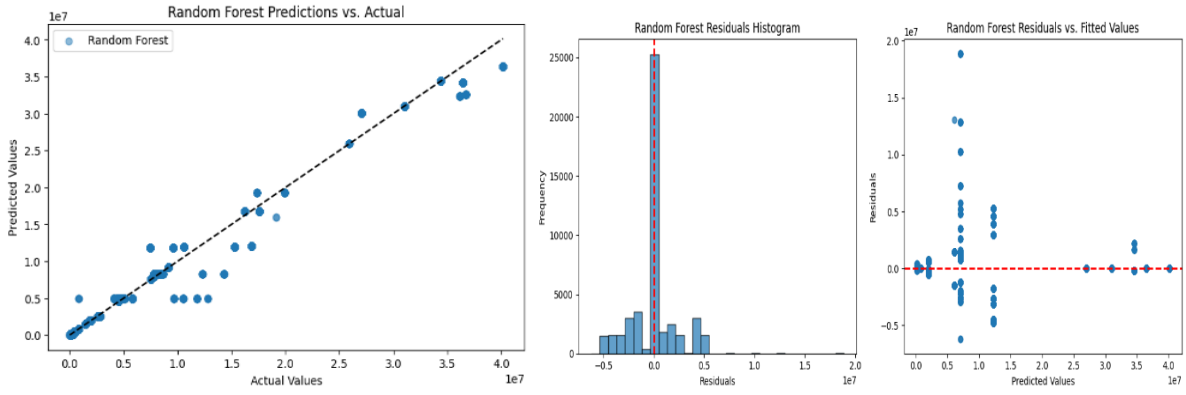



*Figure 24: Visualizations of Predicted values & Residuals on Avocado Model*

#### 4.2.4.1.1.5 Horticultural Product 902: Tea

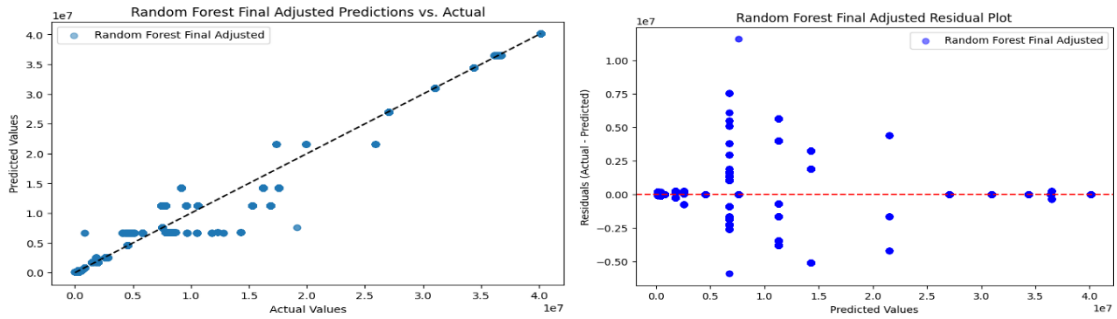



*Figure 25: Visualizations of Predicted values & Residuals on Avocado Model*

#### 4.2.4.1.2 Model Metrics Comparisons

```
Test Accuracy: 0.9662

Classification Report (Test Set):
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       120
           1       0.98      0.92      0.95       120
           2       0.92      0.98      0.95       115

    accuracy                           0.97       355
   macro avg       0.97      0.97      0.97       355
weighted avg       0.97      0.97      0.97       355
```

From the above the following can be inferred

For the metrics, Precision (how many of the predicted class labels were actually correct) or Recall (How many actual class labels were correctly identified) are the key determinants. The harmonic mean (F1-Score) is used as a balance between the two.

In terms of this the model had accuracy: 97%. This results indicate that the model correctly classified 97% of test samples. The Macro Avg (0.97) refers to the average performance across all classes, especially for imbalanced datasets. Lastly, the weighted Avg (0.97) takes class frequency into account implying the model shows high consistency.

#### 4.2.4.2 XGBoost (Multi-Class) Classification

The XGBoost Classification model was used for multi-class classification with optimized hyperparameters. After the Log transformation, scaling process and polynomial interaction terms, (humidity windspeed) were added. The adjusted hyperparameters such as adjusting learning rate and tree depth, improved regularization and reduced over fitting.

##### 4.2.4.2.1 Model Metrics Comparisons

This model was evaluated using a Confusion matrix, classification reports, and SHAP feature importance. It had an improved performance when compared to RF due to better handling of complex interactions by capturing non-linear dependencies more effectively.

The summary comparison metrics of both models indicates that;

```
COMPARISON

XGBoost Final R²: 0.9080
XGBoost Final RMSE: 3865392.82
Random Forest Optimized R²: 0.9680
Random Forest Optimized RMSE: 2280920.90
```

<u>$R^2$ Score Analysis</u>

- XGBoost (0.9080): Explains 90.80% of the variance,
- Random Forest Optimized (0.9680): Now explains 96.80% of the variance, meaning it captures more patterns in the data than XGBoost.

<u>RMSE Analysis</u>

- The deviation of the amount predicted from the actual for XGBoost is Ksh.3,865,392.82: The average prediction error is significant, suggesting it may not be as robust after adjustments.
- Random Forest Optimized is Kshs. 2,280,920.90: This Now has much lower prediction errors than XGBoost, making it the better performer in this case.

#### 4.2.4.3 Random Forest Regressor

The Random Forest model was trained as a baseline comparison using 100 decision trees, each with a max depth of 5. To enhance generalization, constraints like min_samples_split=5 and min_samples_leafs = 5 were added. The model reduced variance compared to individual decision trees but performed slightly worse than XGBoost in terms of accuracy and error metrics.

##### 4.2.4.3.1 Model Metrics

RMSE is a common metric used to evaluate the performance of a regression model. It measures the average error the model makes when predicting values. Since the target variable (Horticultural Value) is log-transformed, the RMSE values represent the average error in predicting the log-transformed target and not a percentage error and are absolute errors in the log-transformed space. The impact of these error is noted when we reverse the log transformation (by exponentiating), the error in the original scale might seem larger because the log transformation compresses large values. In the regressor models, RMSE penalizes large errors more than small ones (because of squaring), thus the Smaller the RMSE the better the model performance. Lastly, all RMSEs are in the same units as the target variable making them easy to interpret.

#### 4.2.4.4 XGBoost Regressor Random Forest

XGBoost, a powerful gradient boosting algorithm, was trained with 100 estimators, a max depth of 5, and a learning rate of 0.05. To prevent overfitting, subsampling (0.8), column sampling (0.8), and L1/L2 regularization (reg_alpha=5, reg_lambda=10) were applied. The model leveraged min_child_weight=10 to reduce small, unnecessary splits. Due to its ensemble nature, XGBoost provided strong predictive performance with high accuracy.

#### 4.2.4.4.1 Model Training

The following were the visualizations of the predictions and residuals of the XGBoost Regressor Random Forest models, that enabled us identify systematic errors of the individual models of each horticultural good by displaying the differences between actual and predicted values thus enabling improved performance of the models.

The trend below indicated Heteroscedasticity given the residuals are unevenly spread. Residuals appear to decrease in variance as predicted values decrease. This suggests the model underperforms for lower values. The points are not evenly distributed around the red dashed line (zero residuals). There are some outliers, especially at lower residual values (e.g., around -1.5) indicating a possible model bias. Residuals tend to be negative for lower predicted values and positive for higher predicted values. This could indicate that the model is overestimating small values and underestimating large values.

##### 4.2.4.4.1.1 Horticultural Product 603: Flowers

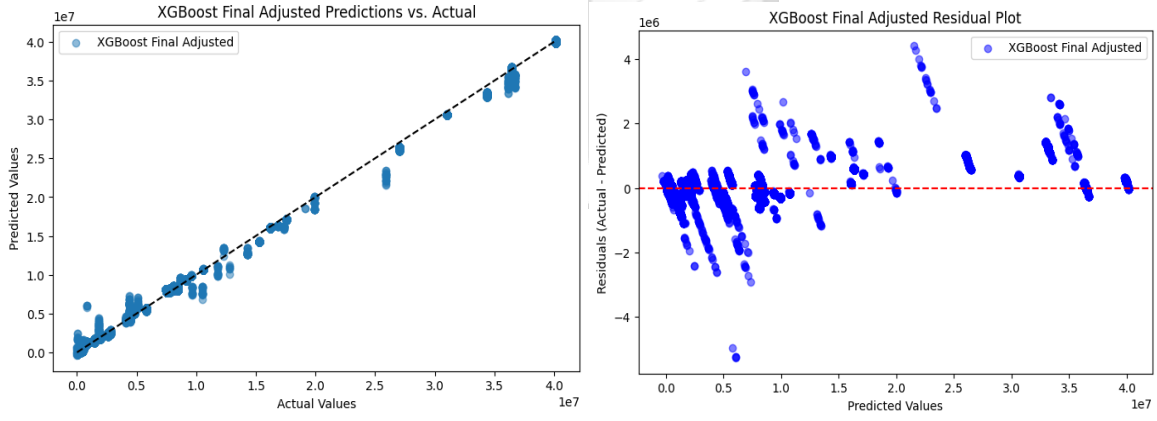



*Figure 26: XGBoost RF Regressor Predictions & Residuals of Flowers Model*

#### 4.2.4.4.1.2 Horticultural Product 709: Vegetables

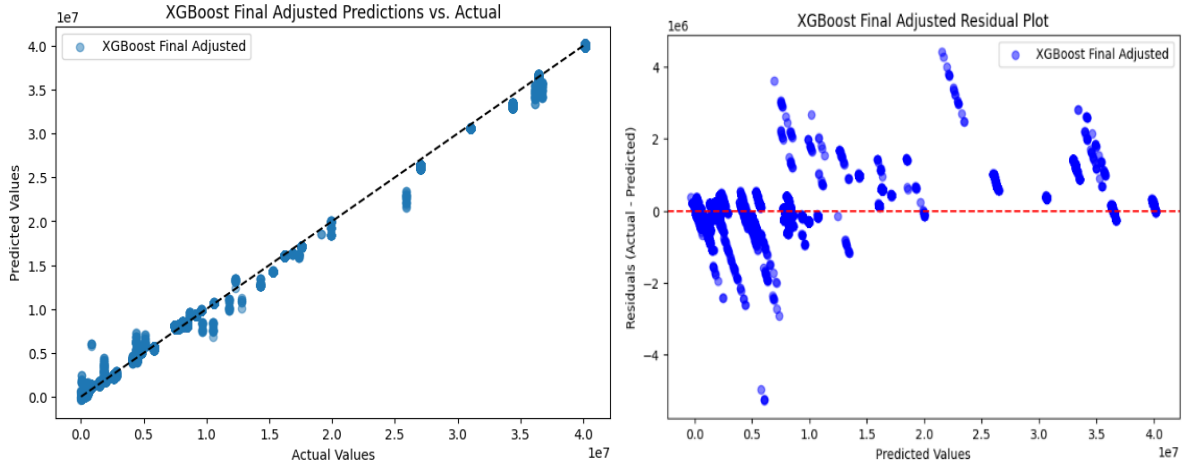



*Figure 27: XGBoost RF Regressor Predictions & Residuals of Vegetables Model*

#### 4.2.4.4.1.3 Horticultural Product 804: Fruits

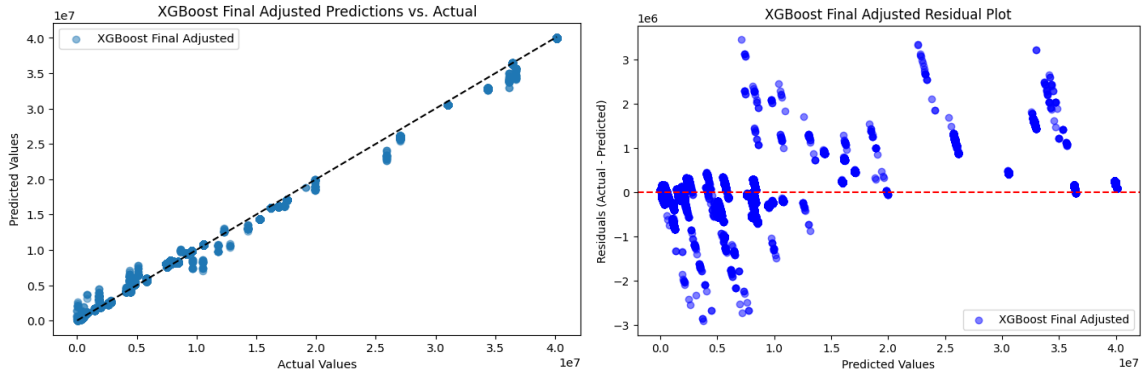



*Figure 28: XGBoost RF Regressor Predictions & Residuals of Fruits Model*

#### 4.2.4.4.1.4 Horticultural Product 810: Avocado & Pineapple

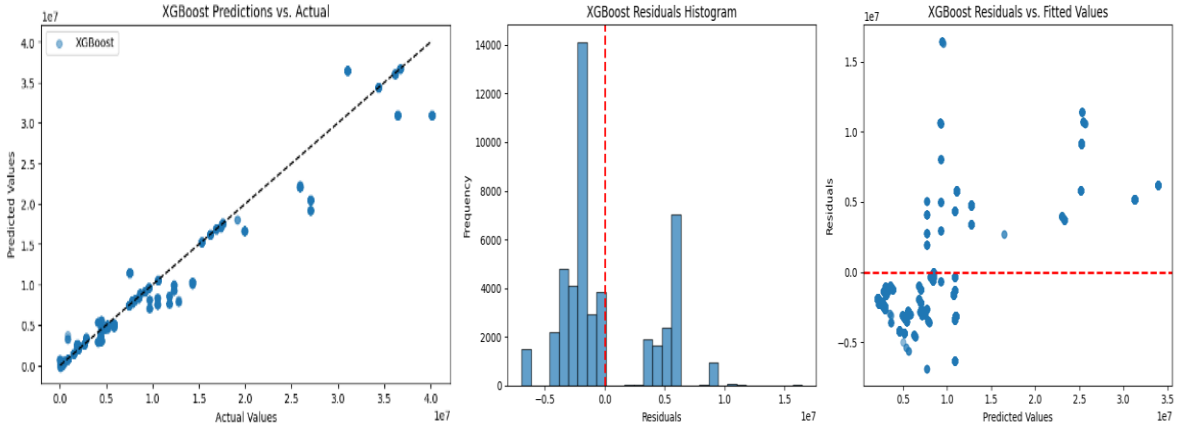



*Figure 29: XGBoost RF Regressor Predictions & Residuals of Avocado & Pineapple Model*

#### 4.2.4.4.1.5 Horticultural Product 902: Tea

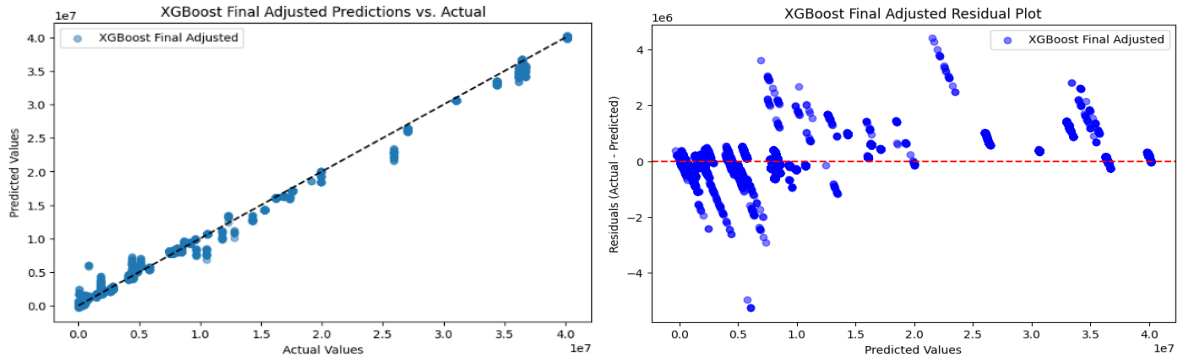



*Figure 30: XGBoost RF Regressor Predictions & Residuals of Tea Model*

XGBoost RF Regressor Predictions & Residuals of Tea Model

#### 4.2.4.4.2 Model Metrics Comparisons

For most of the horticultural product the following were the metrics

```
Model Comparison:
XGBoost R²: 0.9988
XGBoost RMSE: 436638.34
Random Forest R²: 0.9746
Random Forest RMSE: 2030656.77
XGBoost: R² = 0.9988, RMSE = 436638.34
Random Forest: R² = 0.9746, RMSE = 2030656.77
```

Based on the above we can infer the following

R² Score Analysis

- XGBoost (0.9988): The model explains 99.88% of the variance in the target variable, meaning it fits the data almost perfectly.
- Random Forest (0.9746): Still a strong model, explaining 97.46% of the variance, but slightly less accurate than XGBoost.

RMSE Analysis

- XGBoost (Kshs.436,638.34): The model's predictions, on average, deviate by around Kshs. 436,638 from actual values, indicating lower errors.

- Random Forest (Kshs. 2,030,656.77): The deviation is significantly higher, meaning it produces larger prediction errors compared to XGBoost.

### 4.2.5 Summary Model Comparisons

A generalized summary of the above models is indicated below:

**Classifier vs. Regressor Model Comparison**

| Metric | Classification Models | Regression Models |
|---|---|---|
| **Algorithms Used** | Random Forest, XGBoost, Gradient Boosting | Random Forest, XGBoost (RF Mode), Stacked Model |
| **Validation Accuracy / $R^2$** | RF: 96.24% | RF: 0.9854 |
| | XGB: 97.18% | XGB (RF Mode): 0.9856 |
| | | Stacked: 0.9846 |
| **Test Accuracy / $R^2$** | RF: 96.62% | RF: 0.9576 |
| | XGB: 96.34% | XGB (RF Mode): 0.9798 |
| | | Stacked: 0.9793 |
| **Cross-Validation Performance** | RF: 64.15% | RF: 0.6129 |
| | XGB: 77.61% | XGB (RF Mode): 0.6343 |
| | | Stacked: - |
| **Overfitting Risk?** | High (GB shows 100% accuracy) | Lower (slight generalization gap) |
| **Best Model?** | XGB (Validation: 97.18%, Test: 96.62%) | XGBoost (RF Mode) (Validation: 0.9856, Test: 0.9798, RMSE: 0.3128) |

For precise prediction of horticultural values, XGBoost RF Regressor is therefore our selected option.

### 4.2.6 Gradient XGBoosting Random Forest Regressor Predictions

#### 4.2.6.1 Model Metrics Comparisons : Horticultural Product 603: Flowers

Validation $R^2$: 0.9995, RMSE: 0.0499
Test $R^2$: 0.9993, RMSE: 0.0607

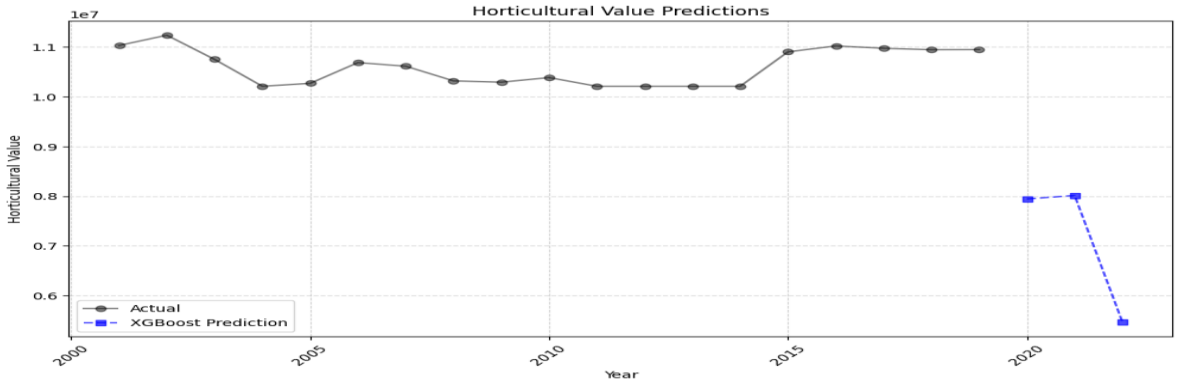



*Figure 31: Predicted Future Values of Flowers Between 2020-2025*

**Validation Set Performance**

- **$R^2$ (Coefficient of Determination): 0.9995** → The model explains **99.95%** of the variance in the validation data, meaning the predictions are extremely close to actual values.
- **RMSE (Root Mean Squared Error): 0.0499** → The average prediction error is very low, indicating high accuracy.

**Test Set Performance**

- **$R^2$: 0.9993** → The model maintains a **99.93%** accuracy on unseen test data, confirming that it generalizes well.
- **RMSE: 0.0607** → The slightly higher error compared to the validation set suggests minimal performance degradation, which is expected when applying the model to new data.

### 4.2.6.2 Model Metrics Comparisons : Horticultural Product 709: Vegetables

Validation $R^2$: 0.9995, RMSE: 0.0499

Test $R^2$: 0.9993, RMSE: 0.0607

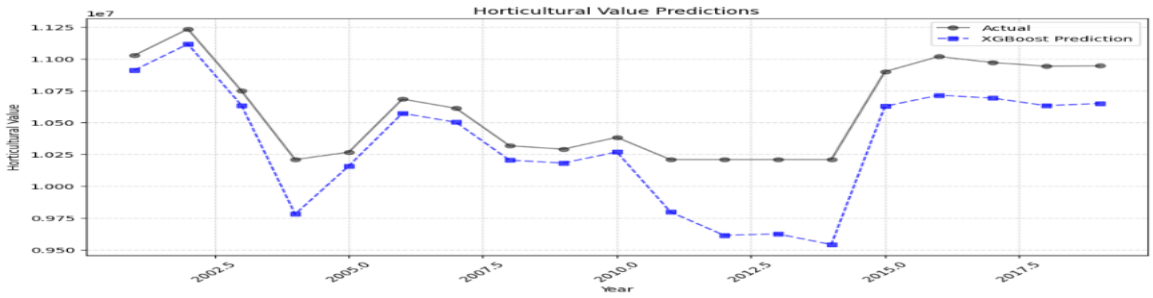



*Figure 32: Predicted Future Values of Vegetables Between 2020-2025*

**Validation Set Performance**

- **$R^2$ (Coefficient of Determination): 0.9995** → The model explains **99.95%** of the variance in the validation data, meaning the predictions are extremely close to actual values.
- **RMSE (Root Mean Squared Error): 0.0499** → The average prediction error is very low, indicating high accuracy.

**Test Set Performance**

- **$R^2$: 0.9993** → The model maintains a **99.93%** accuracy on unseen test data, confirming that it generalizes well.
- **RMSE: 0.0607** → The slightly higher error compared to the validation set suggests minimal performance degradation, which is expected when applying the model to new data.

### 4.2.6.3 Model Metrics Comparisons: Horticultural Product 804: Fruits

Validation $R^2$: 0.9997, RMSE: 0.0412

Test $R^2$: 0.9995, RMSE: 0.0530

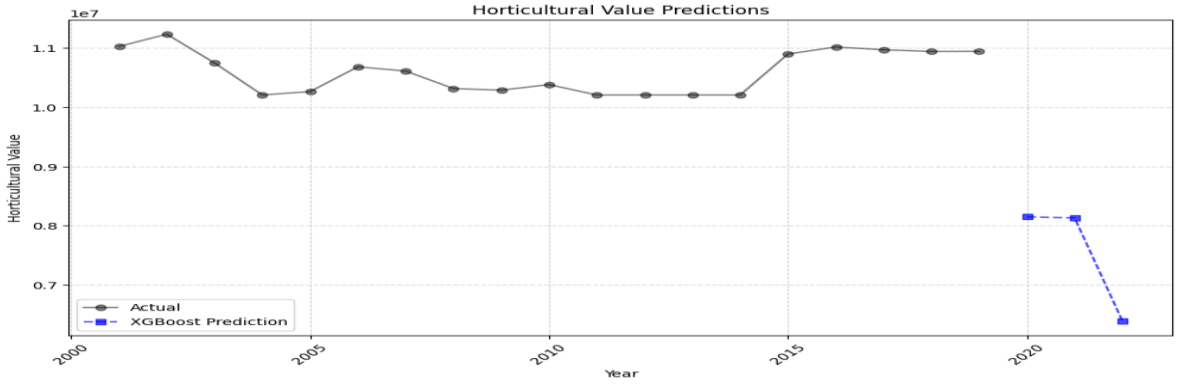



*Figure 33: Predicted Future Values of Fruits Between 2020-2025*

**Validation Set Performance**

- **$R^2$ (Coefficient of Determination): 0.9997** → The model explains **99.97%** of the variance in the validation data, meaning the predictions are extremely close to actual values.
- **RMSE (Root Mean Squared Error): 0.0499** → The average prediction error is very low, indicating high accuracy.

**Test Set Performance**

- **$R^2$: 0.9995** → The model maintains a **99.95%** accuracy on unseen test data, confirming that it generalizes well.
- **RMSE: 0.0530** → The slightly higher error compared to the validation set suggests minimal performance degradation, which is expected when applying the model to new data.

#### 4.2.6.4 Model Metrics Comparisons: Horticultural Product 810: Avocadoes & Pineapples

Validation $R^2$: 0.9996, RMSE: 0.0443

Test $R^2$: 0.9995, RMSE: 0.0533

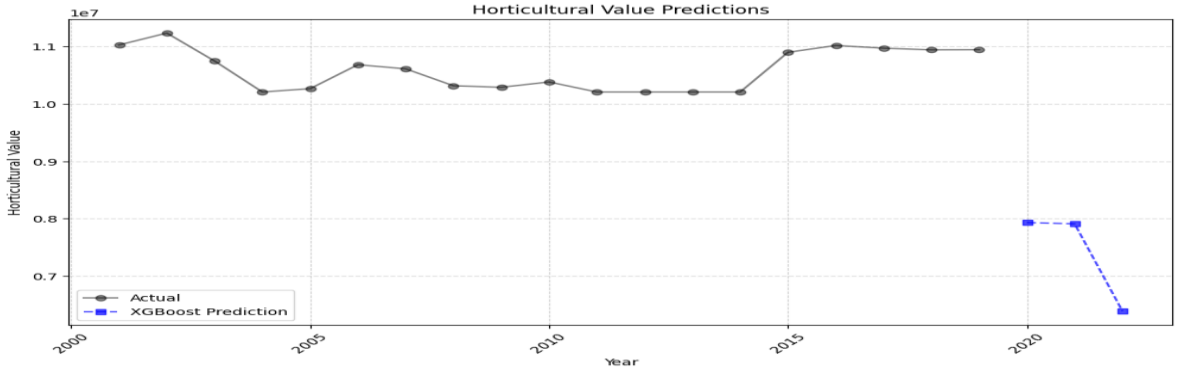



*Figure 34: Predicted Future Values of Avocadoes & Pineapples Between 2020-2025*

**Validation Set Performance**

- **$R^2$ (Coefficient of Determination): 0.9996** → The model explains **99.96%** of the variance in the validation data, meaning the predictions are extremely close to actual values.
- **RMSE (Root Mean Squared Error): 0.0443** → The average prediction error is very low, indicating high accuracy.

**Test Set Performance**

- **$R^2$: 0.9995** → The model maintains a **99.95%** accuracy on unseen test data, confirming that it generalizes well.
- **RMSE: 0.0533**→ The slightly higher error compared to the validation set suggests minimal performance degradation, which is expected when applying the model to new data.

### 4.2.6.5 Model Metrics Comparisons : Horticultural Product 902: Tea

Validation $R^2$: 0.9995, RMSE: 0.0499

Test $R^2$: 0.9993, RMSE: 0.0607

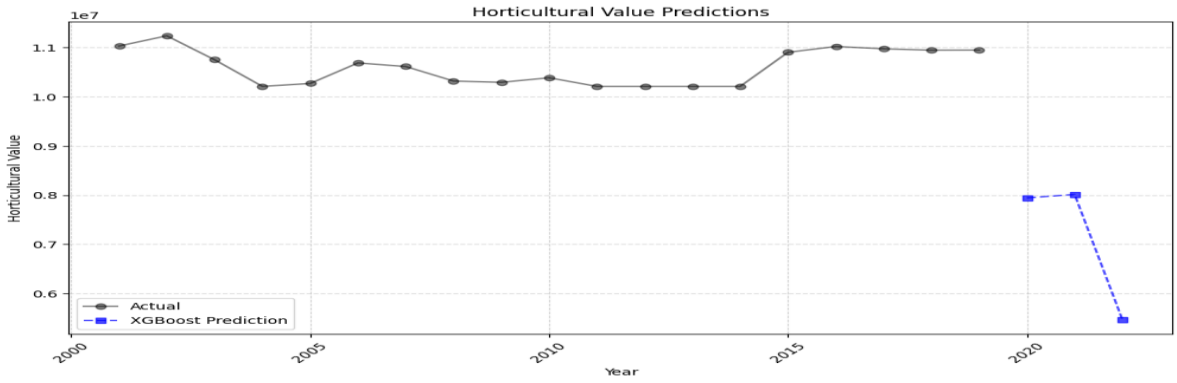



*Figure 35: Predicted Future Values of Tea Between 2020-2025*

**Validation Set Performance**

- **$R^2$ (Coefficient of Determination): 0.9995** → The model explains **99.95%** of the variance in the validation data, meaning the predictions are extremely close to actual values.
- **RMSE (Root Mean Squared Error): 0.0499** → The average prediction error is very low, indicating high accuracy.

**Test Set Performance**

- **$R^2$: 0.9993** → The model maintains a **99.93%** accuracy on unseen test data, confirming that it generalizes well.
- **RMSE: 0.0607** → The slightly higher error compared to the validation set suggests minimal performance degradation, which is expected when applying the model to new data.

# Chapter 5: Discussion

## 5.1 Interpretation of Findings

The key research objectives of this this study were to;-

i. Determine the factors that influence the horticultural sectoral growth
ii. Identify key horticultural subsectors that can be concentrated on based on their growth levels.
iii. Establish destinations to which Kenya has unrealized export potential.

### 5.1.1 Factors Influencing Horticultural Sub-Sectoral Growth

From both the Random Classifier and XGBoosting Random Regressor model, it was established that climatic features such as the temperature, the humidity and the rainfall played a key role in this industry. These indeed play a key role during the growth period of these goods as well as maintaining them once grown. The fresher they can be fresh they can be maintained before they can be delivered the higher their value. It would be therefore important for farmers, transporters and exporters to put in place measures that ensure, low heat levels are maintained during the transportation of these goods so as to increase their value despite the increase freight costs.

Another key contributor to this industry is greenhouse gas emissions which pollute the air depleting our natural resources such rain. Other factors that determine the market include the cost per unit to which we sell goods to the UK who seem to have a determining hand to the European Union which is our largest Market. Another huge determinant in this industry are the transportation costs. However, given that these goods are sensitive, little can be done to control this variable. We however can invest in self plug ports and railway systems which enable the reefer containers temperatures are maintained to the ports so as to ensure these products are kept fresh.

### 5.1.2 Horticultural subsectors with potential for growth

The type of horticultural good exported also plays a key role in determining its value. The current horticultural goods with the largest room for growth and revenue generation are vegetables followed by avocadoes and pineapples while tea has significantly dropped. These are the subsectors that as an economy we should now be fully concentrated on and resources input to enable them to thrive. Historically, Kenya has consistently relied on Coffee and Tea with many incentives being opened to those markets. However, the competition in that industry has grown.

However when it comes to vegetables and fruits, Kenya has the advantage of organically grown products whose demand is high with may foreigners who are health conscious. An advantage of this market is the fact that anyone grow these products even on small scale

### 5.1.3 Market expansion for the Kenya Horticultural Goods Subsectors

Our top export markets are the United Kingdom, France, Netherlands, Malaysia and Netherlands respectively. However, internationally, the United Kingdom is the 6th largest consumer of horticultural goods with the top five being Netherlands, Sweden, France, Russia and Romania. Other Countries include Serbia, Austria, Turkey and Spain. These are untapped markets that Kenya needs to explore to expand its market presence through appealing to the current labour force who are health conscious and thus increase the revenue generated. This will also reduce the control that the United Kingdom has over the Horticultural industry in Kenya.

## 5.2 Significance

The horticultural sector is a key driver of economic growth, particularly in agricultural economies. Export revenues from fruits, vegetables, and flowers contribute significantly to national GDP. Additionally, the sector provides employment opportunities across the value

chain, from farming and logistics to marketing and retail. Enhancing trade efficiency through predictive analytics ensures sustainable economic growth and job creation.

Horticultural trade affects the trade balance of exporting and importing nations. Accurate forecasting of trade flows helps governments formulate policies that maximize foreign exchange earnings while ensuring food security. For Kenya, horticultural goods generate a substantial revenue, stabilizing national economies and promoting investment in agricultural infrastructure.

Trade prediction models thus help stakeholders anticipate supply and demand trends, thereby mitigating market volatility. By identifying potential surpluses or shortages, governments and businesses can implement strategic interventions to stabilize prices, ensuring fair returns for farmers and affordability for consumers. Price stability is crucial for maintaining competitive markets and preventing inflationary pressures.

## 5.3 Implications

Governments can use predictive analytics to design evidence-based agricultural policies, optimize subsidies, and plan for trade negotiations. The ability to anticipate fluctuations in trade allows policymakers to mitigate economic risks and enhance global competitiveness in horticultural exports. Accurate forecasts encourage investment in critical infrastructure such as cold storage facilities, transportation networks, and processing plants. Investors and agribusinesses can make data-driven decisions on expanding operations, improving supply chains, and enhancing market accessibility, thereby increasing economic resilience.

Climate variability significantly impacts horticultural production and trade. Our models integrate climate-related factors such as rainfall, temperature, and $CO_2$ emissions, providing insights into how environmental changes affect trade patterns. This should enable stakeholders to adopt climate-smart agricultural practices, ensuring long-term sustainability.

Countries that leverage predictive modelling gain a competitive edge in international markets. By understanding trade dynamics, exporters can strategically position themselves, negotiate better trade agreements, and diversify their markets to reduce dependency on a single trading partner. Predicting agricultural trade patterns is critical to decision making in the public and private domains, especially in the current context of trade wars with tit-for-tat tariffs. For instance, farmers likely consider the potential demand from alternative foreign sources before deciding to plant crops, especially in large exporters.

Similarly, countries setting budgets for farm programs need better predictions of prices and trade flows for assessing domestic production and consumption needs and instruments employed to achieve those outcomes. This study demonstrates the high relevance of ML models to predicting trade patterns with a greater accuracy than traditional approaches for a range of time periods. Existing forecasts of trade such as those by WTO, OECD and USDA are a combination of model-based analyses and expert judgement and tend to have high variability.

The ML models, by relying on data and deep learning, allow for alternative and robust specifications of complex economic relationships. Moreover, the ML models are cross-validated and provide ways to simulate trade outcomes under alternative policy scenarios including their uncertainty in recent times. Future work focusing on data/matrix completeness (a major issue when dealing with zeros in trade and tariffs), multi-variate response variables and prescriptive ML techniques to compare with current causal models would greatly aid in public and private decision making.

## 5.4 Limitations

A major challenge in this study was on the data quality and availability. The data was limited and had quite a number of missing values. While we applied log transformations and scaling, some economic variables may have complex, non-linear relationships that were not fully captured, potentially limiting the models effectiveness. There were also external economic factors that the models does not account for such as sudden economic shocks, trade policy changes, or global supply chain disruptions, which can significantly affect horticultural trade. While the models performed well on the test dataset, their ability to generalize to different regions, product categories, or time periods remains uncertain. Advanced models like XGBoost which were used, offer strong predictive power but are less interpretable compared to simpler regression models, making it harder to explain key trade drivers.

# Chapter 6: Conclusion and Recommendation

## 6.1 Conclusion

Horticultural trade plays a critical role in global and regional economies, influencing GDP growth, employment, and food security. Accurate prediction models for horticultural goods trade enhance decision-making by governments, businesses, and policymakers. By leveraging advanced machine learning techniques such as XGBoost and Random Forest, this developed models forecast trade values, assess economic trends, and optimize resource allocation in the horticultural sector. The implementation of this machine learning-driven trade forecasts provides policymakers, businesses, and investors with the tools necessary to enhance decision-making, optimize resource allocation, and ensure sustainable economic growth. As the global economy continues to evolve, integrating data-driven strategies into horticultural trade will be crucial for maintaining resilience and competitiveness in the agricultural sector.

The economic significance of predictive modeling in horticultural trade is profound, because it influences GDP, employment, trade balance, and market stability. This study has demonstrated the need to monitor our climatic conditions because these are now having a direct impact in this industry and affecting not only our economy but also the labour force engaged in this industry.

### 6.2 Recommendation

Future work focusing on data/matrix completeness (a major issue when dealing with zeros in trade and tariffs), multi-variate response variables and ML techniques to compare with current causal models to aid in public and private decision making. In addition, studies can also involve complex, non-linear relationships that were not fully captured, potentially limiting the model's effectiveness. These include external economic factors that the models do not account for such as sudden economic shocks, trade policy changes, or global supply chain disruptions, which can significantly affect horticultural trade. Lastly, given the challenges in obtaining data, the Horticultural Crops Directorate should endeavor to maintain a database where varied data related towards this industry can be stored and be used for future predictive modelling and enhancement of this study.

**REFERENCES**


Albiman, M.M., Yussuf, H.A., Hemed, I.M. (2022). The Effect of Foreign Direct Investment and Trade Openness on the Firms' Export Competitiveness and Products Diversification Among East African Community Members. In: Demena, B.A., Van Bergeijk, P.A. (eds) Trade and Investment in East Africa. Frontiers in African Business Research. Springer, Singapore. https://doi.org.ezproxy.library.strathmore.edu/10.1007/978-981-19-4211-2_9

Anderson, J. E., & Wincoop, V. E. (2004). Trade Costs. Journal of Economic Literature, 691-751.

Barrie,A.S.I, Sillah, A., Bangura, M., (2021): An Empirical Investigation of the Export-Led Growth (ELG) and Import-Led Growth (ILG) Hypotheses in Sierra Leone. Research Department, Bank of Sierra Leone, Freetown, Sierra Leone.

Brenton, Paul, Cadot,m Oliver and Pierola, d, Martha: Pathways to African Export Sustainability, World Bank Publications, 2012. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/strathmore-ebooks/detail.action?docID=967093. Created from strathmore-ebooks on 2022-11-19 21:51:55.

Cali, Massimiliano; Ghose, Devaki; Montfaucon, Angella Faith; Ruta, Michele. 2022. Trade Policy and Exporters' Resilience : Evidence from Indonesia. Policy Research Working Paper;10068. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/37489 License: CC BY 3.0 IGO." URI http://hdl.handle.net/10986/37489. Pathways to African Export Sustainability

CHAPTER 236 Of The Laws Of Zambia :The Plant Variety And Seeds Act https://www.parliament.gov.zm/sites/default/files/documents/acts/Plant%20Variety%20and%20Seeds%20Act.pdf

Dabin, Z., Zhu Hou, Z., Jingguang, Z., (2009): Forecasting of Customs Export Based on Gray Theory. International Conference on Business Intelligence and Financial Engineering. Information Management Department Huazhong Normal University Wuhan, China.

East African community, Common External Tariff, 2017 Version

Eita, Joel Hinaunye & Jordaan, Andre. (2007). Export and economic growth in Namibia: A granger causality analysis. South African Journal of Economics. 75. 540-547. 10.1111/j.1813-6982.2007.00132. x.

Elam, T. E., & Uko, O. E. (1977). Palm Oil and the World Fats and Oils Economy. *Illinois Agricultural Economics*, *17*(2), 17–21. https://doi.org/10.2307/1348956

European Commission Food Alerts - Rapid Alert System for Food and Feed (RASFF): https://www.foodstandards.gov.scot/consumers/food-crime/report-a-problem/european-commission-food-alerts-rapid-alert-system-for-food-and-feed-rasff

Exports and Economic Growth: Some Additional EvidenceView article pageRati Ramformat_quoteCITECopyright 1985 The University of Chicagohttps://doi.org/10.1086/451468open_in_newPublisherUniversity of Chicago PressISSN0013-0079eISSN1539-2988PrintJan., 1985Pages415 - 425

Exports and Regional Economic GrowthView article pageCharles M. Tieboutformat_quoteCITECopyright 1956 The University of Chicagohttps://doi.org/10.1086/257771open_in_newPublisherThe University of Chicago PressISSN0022-3808eISSN1537-534XPrintApr., 1956Pages160 - 164

Frost, W. & Jones, C.L.: The Vegetable Oil Industry of France

Granger, C.W.J. (2001). 'Essays in Econometrics: Collected Papers of Clive W.J. Granger (Econometric Society Monographs Vol. 2)'. In E. Ghysels, N.R. Swanson, and M.W. Watson (eds), Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Cambridge: Cambridge University Press.

Grossman, G.M, and E. Helpman (1990). 'Trade, Knowledge Spillovers, and Growth.' NBER Working Paper 3485. Cambridge, MA: National Bureau of Economic Research.

Grossman, G.M, and E. Helpman: 'Quality Ladders in the Theory of Growth.' The Review of Economic Studies, 43–61.

H. Raza, I. Manarvi, J. Ahmed, K. Khan and K. ur Rehman (2009):, "A methodology of export sectors identification through data mining," 2009 International Conference on Computers & Industrial Engineering, 2009, pp. 1496-1499, doi: 10.1109/ICCIE.2009.5223645.

Hemal, M. (2017). Relationship between Export Diversification and Economic Growth: The Case for SADC countries.

Hodey, L.S. (2013): Export Diversification and Economic Growth In Sub–Saharan Africa. University Of Ghana

Chaney, T., (2011) :"The Gravity Equation in International Trade: An Explanation" https://www.tse-fr.eu/sites/default/files/medias/doc/by/chaney/distance.pdf

Hume: (1994)Political Essays , pp. 136 – 149, Publisher: Cambridge University Press, DOI: https://doi.org/10.1017/CBO9781139170765.023:

Ibrahim, I. (1996): "Exports and economic growth in developing countries" . ETD collection for University of Nebraska - Lincoln. AAI9715967. https://digitalcommons.unl.edu/dissertations/AAI9715967

Irandu, E.M. Factors influencing growth of horticultural exports in Kenya: a gravity model analysis. GeoJournal 84, 877–887 (2019). https://doi.org/10.1007/s10708-018-9888-x

Ireen C., (2008)"An empirical analysis of the determinants and growth of South African exports." Thesis, University of Fort Hare, 2008. http://hdl.handle.net/10353/198.

Joshi, A.(2012): Long Term Causality of Export Led Growth(ELG) using VECM model with reference to India. Asian Journal of Applied Science and Engineering. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

K. Kaushik, K., K. Klein, K., & Arbenser, L. (2006). An Econometric Analysis of Export-led Growth in A Multivariate Var Framework: Evidence from India. Foreign Trade Review, 41(3), 3–24. https://doi.org.ezproxy.library.strathmore.edu/10.1177/0015732515060301

Kalaitzi, A.S. (2015):The Causal Relationship Between Exports And Economic Growth: Time Series Analysis For UAE (1975-2012). Department of Accounting, Finance and Economics Manchester Metropolitan University.

Kanwal, U., & Sardar, M. A. (2009). Impact of International Trade on Sub Saharan Africa's Economic Growth (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-3522

Karamuriro, H.T., & Karukuza, W.N. (2015). Determinants of Uganda's Export Performance: A Gravity Model Analysis. International Journal of Business and Economics Research, 4, 45.

Kenya Association of Manufacturers: Kenya Business Guide, 2018

Kristjánsdóttir H. A gravity model for exports from Iceland. 2005:57.

Lien, L. T. B., Feng, L. X., & Fei, X. L. (2019). Determinants of China's Rice Export after WTO Accession: A Gravity Model Analysis. Asian Journal of Advances in Agricultural Research, 9(3), 1–12. https://doi.org/10.9734/ajaar/2019/v9i330008

Linnemann H.: (1966): An econometric study of international trade flows. North-Holland Publishing Company Amsterdam.

Lopez-Calix, J.R.,(2020): Leveraging Export Diversification in Fragile Countries : The Emerging Value Chains of Mali, Chad, Niger, and Guinea. International Development in Focus; World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/33259 License: CC BY 3.0 IGO."

M. L. Jhala. (1984). Restructuring Edible Oil and Oilseeds Economy of India. *Economic and Political Weekly*, *19*(39), A111–A128. http://www.jstor.org/stable/4373618

Maina, L.W (2012), Rapid Identification Of Edible Oils Manufactured In Kenya By Linda Wangeci Maina

Meme, S.M., (2013 ): Export Performance Of The Horticultural Subsector In Kenya-An Empirical Analysis http://erepository.uonbi.ac.ke/bitstream/handle/11295/93214/Meme_Export%20Performance%20of%20the%20Horticultural%20Sub-sector%20in%20Kenya-an%20Empirical%20Analysis.pdf?sequence=3

Moniruzzaman, M.D., Toy, M.M. & Hassan, A.B.M (2011): The Export Supply Model of Bangladesh: An Application of Cointegration and Vector Error Correction Approaches. International Journal In Economics and Financial Issues.

Muhoro, G.& Otieno, M. (2014): Export Led Growth Hypothesis: Evidence from Kenya. Journal of World Economic Research. https://www.academia.edu/31324964/Export_led_growth_hypothesis_Evidence_from_Kenya

Muhoro, G.W.(2012): Export Led Growth Hypothesis, Evidence from Kenyan Data. Unoversity of Nairobi

Muna Sulaiman and Norma Md. Saad, (2009): An Analysis of Export Performance and Economic Growth of Malaysia Using CoIntegraton and Error Correction Models. The Journal of Developing Areas , Fall, 2009, Vol. 43, No. 1 (Fall, 2009), pp. 217- 231 Published by: College of Business, Tennessee State University Stable URL: https://www.jstor.org/stable/40376281

Ndemo, M. (2020): Determinants of Kenyan Tea Exports: The Gravity Model Approach. Sato Agriculture Information Institute (AII), Chinese Academy of Agricultural Sciences, Haidian, Beijing 100081, China. Journal of Economics and Sustainable Development www.iiste.org ISSN 2222-1700 (Paper) ISSN 2222-2855 (Online) Vol.10, No.14, 2020 132

Ndemo, M. S., (2020): Determinants of Kenyan Tea Exports: The Gravity Model Approach. Agriculture Information Institute (AII), Chinese Academy of Agricultural Sciences, Haidian, Beijing 100081, China. Journal of Economics and Sustainable Development www.iiste.org ISSN 2222-1700 (Paper) ISSN 2222-2855 (Online) Vol.10, No.14, 2020

Ngumi, P. N. (2009): Exports& Economic Growth, The Case of Kenya. University Of Nairobi

Nyasulu, T. (2013): Assessing the Impact of Exports and Imports on Economic Growth: A Case study of Malawi from 1970 to 2010. University of Western Cape

Okumu, M.J. (2022): Determinant's in Kenyan Trade in Goods: A Gravity Model Approach. University of Nairobi. http://publication.aercafricalibrary.org/bitstream/handle/123456789/3446/MOMANYI_THESIS_22.09.2022.pdf?sequence=1&isAllowed=y

Orindi, M. (2011): Determinanats of Kenyan Exports: A Gravity Model Approach. http://erepository.uonbi.ac.ke/bitstream/handle/11295/4711/Orindi_Determinants%20of%20Kenyan%20Exports,%20A%20Gravity%20Model%20Approach.pdf?sequence=1

Pöyhönen P.(1963): A tentative model for the volume of trade between countries. Weltwirtsch Arch. 1963; Bd. 90:93-100.

Rondeau, F. and Roudaut, N., "What Diversification of Trade Matters for Economic Growth."

Rutto, R., Odhiambo, S., Obange N. & Enock N., (2019): Effects Of Foreign Direct Investments On Kenya's Manufacturing Exports To Regional Trade Blocs In Africa. International Journal of Economics, Business and Management Research Vol. 3, No. 02; 2019 ISSN: 2456-7760

Rwenyagila, G.A.(2013):Determinants Of Export Performance In Tanzania. Mzumbe University

Shafiullah, M. & Navaratnam, R. (2016): Do Bangladesh and Sri Lanka Enjoy Export-Led Growth? A Comparison of Two Small South Asian Economies. South Asia Economic Journal17(1) 114–132©2016 Research and Information System for Developing Countries & Institute of Policy Studies of Sri Lanka SAGE Publications sagepub.in/home.nav. DOI: 10.1177/1391561415621825http://sae.sagepub.com

Shetty, J., Challenges and Advantages of Exporting https://www.trademo.com/blog/2021/01/25/challenges-and-advantages-of-exporting/

Sikobi, A. P. (2021): The impact of international trade of commodities on the economic growth of South Africa. World Maritime University Dissertations

Sikobi, A.P. (2021): The Impact Of International Trade Of Commodities On The Economic Growth Of South Africa. World Maritime University.

Taslim, M. A., & Hossain, Md. A. (2015). Asymmetric Transmission of International Price of Edible Oil in Bangladesh. *The Bangladesh Development Studies*, *38*(1), 33–54. https://www.jstor.org/stable/26538647

Tenhoff, H. (2014) Distribution Of U.S. Beef Exports In The International Market. Kansas State University

The Grey Forecasting Model for the Medium-and Long-Term Load Forecasting Feng Song* , Junxu Liu, Tingting Zhang, Jing Guo, Shuran Tian and Dang Xiong , 2020, Corresponding author's e-mail: z16050421@s.upc.edu.cn

Thomas Chaney, "The Gravity Equation in International Trade: An Explanation", September 2011.

Tinbergen J. (1962): Shaping the world economy: Suggestions for an international economic policy. Twentieth Century Fund, New York: 1962.

Trade And Development Index: Developing Countries In International Trade 2005

USAID Kenya Horticultural Competitiveness Project: Global Competitiveness Study: Benchmarking Kenya's Horticulture Sector for Enhanced Export Competitiveness. https://www.tralac.org/images/docs/6933/usaid-khcp-global-competitiveness-study.pdf

Usman, Z., Landry, D., (2021): Economic Diversification in Africa: How and Why It Matters https://carnegieendowment.org/2021/04/30/economic-diversification-in-africa-how-and-why-it-matters-pub-84429

Varshini, N. M., & Manonmani, M. (2018). Causal Relationship between Trade & Economic Growth in India during Post WTO Period. Indian Journal of Industrial Relations, 54(1), 54–65. https://www.jstor.org/stable/26536512

Wamalwa, P. and Were, M., (2019): Is export-led growth a mirage? The case of Kenya. United Nations University World Institute for Development Economics Research,

X. Li, F. Kong, Y. Liu and Y. Qin, "Applying GM (1,1) Model in China's Apparel Export Forecasting," 2011 Fourth International Symposium on Computational Intelligence and Design, 2011, pp. 245-247, doi: 10.1109/ISCID.2011.163.

Yang, J. (2008). 'An Analysis of So-called Export-led Growth.' IMF Working Paper 08/220. Washington, DC: International Monetary Fund.

Yego, H. K. (2015):An Analysis Of Kenyan Livestock Exports: A Gravity Model Approach. Researchjournali's Journal of Economics Vol. 3 | No. 4 October | 2015 ISSN 2347-8233

Z. Dabin, Z. Hou and Z. Jingguang, "Forecasting of Customs Export Based on Gray Theory," 2009 International Conference on Business Intelligence and Financial Engineering, 2009, pp. 630-633, doi: 10.1109/BIFE.2009.148.

Munisamy Gopinath, Feras A. Batarseh, and Jayson Beckman, " Machine Learning In Gravity Models: An Application To Agricultural Trade," National Bureau Of Economic Research, 2020, NBER Working Paper No. 27151, http://www.nber.org/papers/w27151

# APPENDICES

## Appendix A: Turnitin Report



# Yvonne Julianadima Odera

## APPLICATION OF MACHINE LEARNING IN THE HORTICULTURAL EXPORT_145615_FINAL.docx

Strathmore University (Main Account)

### Document Details

Submission ID
trn:oid:::2945:275146205

Submission Date
Mar 28, 2025, 3:11 PM GMT+3

Download Date
Mar 28, 2025, 3:30 PM GMT+3

File Name
APPLICATION OF MACHINE LEARNING IN THE HORTICULTURAL EXPORT_145615_FINAL.docx

File Size
3.3 MB

82 Pages

15,069 Words

92,172 Characters

turnitin Page 1 of 92 - Cover Page Submission ID trn:oid:::2945:275146205

# 11% Overall Similarity

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

## Filtered from the Report

- Bibliography
- Quoted Text

## Match Groups

- 121 Not Cited or Quoted 10%
  Matches with neither in-text citation nor quotation marks
- 14 Missing Quotations 1%
  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

8% Internet sources
6% Publications
8% Submitted works (Student Papers)

## Integrity Flags

0 Integrity Flags for Review

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.

## Match Groups

- 121 Not Cited or Quoted 10%
  Matches with neither in-text citation nor quotation marks
- 14 Missing Quotations 1%
  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

8% Internet sources
6% Publications
8% Submitted works (Student Papers)

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

| | | |
|---|---|---|
| 1 | Internet | publication.aercafricalibrary.org 1% |
| 2 | Internet | de.overleaf.com 1% |
| 3 | Internet | www.econstor.eu <1% |
| 4 | Internet | www.nber.org <1% |
| 5 | Internet | www.coursehero.com <1% |
| 6 | Internet | wiredspace.wits.ac.za <1% |
| 7 | Submitted works | University of Liverpool on 2022-12-23 <1% |
| 8 | Internet | ir.knust.edu.gh <1% |
| 9 | Internet | su-plus.strathmore.edu <1% |
| 10 | Internet | erepository.uonbi.ac.ke <1% |

## Appendix B: Ethics Review Approval

Strathmore UNIVERSITY

**9th April 2024**

Ms Odera Yvonne,
yvonne.odera@strathmore.edu

Dear Ms Odera,

**RE: Application of Machine Learning in Establishing Determinants of Growth in the Horticultural Export Sub-Sector in Kenya**

This is to inform you that SU-ISERC has reviewed and **approved** your above SU-masters research proposal. Your application reference number is **SU-ISERC2094/24. The approval period is from 9th April 2024 to 8th April 2025.**

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**

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
09-Apr-2024
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA

VT OMNES
VNVM SINT