

# **A Machine Learning Approach for Accurate Valuation of imports in Uganda**

Thesis by

**Paul Sentongo**

Submitted in partial fulfillment of the requirements

for the degree of

Master of Science

in

Data Science and Analytics



**UGANDA CHRISTIAN  
UNIVERSITY**

*A Centre of Excellence in the Heart of Africa*

Uganda Christian University Mukono, Uganda

2024 (Defended 2025)

© 2024

Paul Sentongo

All Rights Reserved

*DEDICATION*

# DECLARATION

I, Paul Sentongo, declare that this thesis titled “**A Machine Learning Approach for Accurate Valuation of Imports in Uganda**” is my original work.

It has not been submitted for any degree or examination in any other university or institution. All sources used in the research have been properly acknowledged through citations and references.

**Paul Sentongo**

Candidate

Sign:.....

**Mr. Ian Raymond Osolo**

Supervisor

**School of Computing, Faculty of Engineering Design and Technology**

Uganda Christian University



# ACKNOWLEDGMENTS

# ABSTRACT

Accurate import valuation is crucial for revenue generation and promoting fair trade practices among countries. Traditional valuation methods are often hindered by inconsistency and vulnerability to fraud, resulting in revenue losses estimated at \$4.9 billion between 2006 and 2015.

This research presents a machine learning approach designed to enhance the precision and efficiency of import item valuation within Uganda's customs framework. Using Uganda's historical trade data from 2005 to 2023, we developed and validated predictive models that included Random Forests, XGBoost, and Artificial Neural Networks (ANN). Key variables, such as the country of origin and unit price, were analyzed to train the predictive models. The Random Forest model achieved excellent performance, registering 95% accuracy and outperforming conventional methods. This demonstrates the transformative potential of machine learning.

The research findings showcase the potential of machine learning to mitigate revenue leakage, reduce valuation fraud, minimize valuation disputes, and ensure compliance with international trade regulations.

Looking at the bigger picture, this study offers a framework for developing countries facing similar challenges. Recommendations include institutionalizing machine learning-powered valuation systems, upgrading data infrastructure, and integrating AI expertise within customs operations to ensure sustainable improvements.

In conclusion, this research combines technological innovation with policy action, introducing artificial intelligence as a strong pillar for economic resilience in the international trade network.

# Contents

<b>Declaration</b>	<b>iv</b>
<b>Acknowledgments</b>	<b>vi</b>
<b>Abstract</b>	<b>vii</b>
<b>1 INTRODUCTION</b>	<b>2</b>
1.1 BACKGROUND TO THE STUDY . . . . .	2
1.1.0.1 Contextual Background . . . . .	2
1.1.1 Theoretical background . . . . .	3
1.2 PROBLEM STATEMENT . . . . .	3
1.3 OBJECTIVES OF THE STUDY . . . . .	4
1.4 RESEARCH QUESTIONS . . . . .	5
1.5 JUSTIFICATION OF THE STUDY . . . . .	5
1.6 SIGNIFICANCE OF THE STUDY . . . . .	5
1.7 HYPOTHESIS . . . . .	6
1.7.1 Null Hypothesis (H0) . . . . .	6
1.7.2 Alternative Hypothesis (H1) . . . . .	6
1.8 SCOPE OF THE STUDY . . . . .	6
1.9 THEORETICAL FRAMEWORK . . . . .	6
1.10 Chapter Arrangement . . . . .	7
<b>2 LITERATURE REVIEW</b>	<b>8</b>
2.1 Customs valuation: Concepts . . . . .	8
2.1.1 Machine learning in trade . . . . .	9
2.1.1.1 <b>Predictive Analytics in Customs</b> . . . . .	10
2.1.1.2 HS code classification with Naive Bayes Algorithm . . . . .	11
2.1.2 ML-Powered Customs Operations . . . . .	12
2.1.2.1 Brazil . . . . .	12



2.1.2.2	China . . . . .	12
2.1.2.2.0.1	AI-based NII (Non-intrusive detection devices) image recognition system . . . . .	12
2.1.2.2.0.2	Intelligent Passenger Face Recognition System . . . . .	13
2.1.2.3	Belgium: BCTC Behavioral consequences of tariff changes . . . . .	13
2.1.2.4	The RECTS project: Uganda . . . . .	14
2.1.2.5	Ongoing enhancement initiatives . . . . .	15
2.1.2.6	The DATE model . . . . .	15
2.1.3	Importance of Machine Learning in Trade Facilitation. . . . .	15
2.1.4	Barriers to successful implementation of Machine learning in Customs Trade Practices . . . . .	16
2.1.5	Gaps in Existing Literature . . . . .	17
2.1.6	Conceptual Framework for Machine Learning Powered Customs Valuation . . . . .	18
2.1.7	Limitations of the study . . . . .	19
2.1.7.1	Limitations summary . . . . .	20
2.2	Conclusion . . . . .	20
<b>3</b>	<b>METHODOLOGY</b>	<b>21</b>
3.1	Introduction and Scope . . . . .	21
3.2	Data Description . . . . .	21
3.3	Data Pre-processing . . . . .	22
3.3.1	Handling Missing Data . . . . .	22
3.3.2	Outlier Detection . . . . .	23
3.3.3	Exploratory Data Analysis: (EDA) . . . . .	24
3.3.3.1	Data Visualization . . . . .	24
3.3.3.2	Descriptive Statistics . . . . .	25
3.3.4	Feature Engineering . . . . .	26
3.3.5	Data Normalization . . . . .	27
3.4	Model Development . . . . .	27
3.4.1	Algorithm Selection . . . . .	27
3.4.1.1	Linear Regression Model . . . . .	27
3.5	Model Selection . . . . .	27
3.5.1	Linear Regression Model . . . . .	27
3.5.1.1	Graphical Representation . . . . .	28
3.5.2	Random forest model . . . . .	29
3.5.3	Training and Validation . . . . .	29
3.5.4	Model performance comparison . . . . .	29
3.5.5	Model deployment . . . . .	30



# List of Figures

2.1	A flowchart illustrating the physical movement of goods in international trade. Goods start from a supplier in the exporting country (Factory/Farm/Store) and move through domestic transport (Truck/Train/Airplane) to the customs frontier (Wharf/Airport). They are then loaded onto international transport (Container) and shipped via sea or air. Upon arrival in the importing country, they pass through the customs frontier before reaching their final destination. Additionally, goods transported via gas pipelines are valued at the point of entry into the pipeline (FOB) in the exporting country and upon arrival in the importing country (CIF). . . . .	9
2.2	A customs officer inspecting a cargo container equipped with the RECTS tracking system, ensuring secure and transparent transit monitoring. Picture Source: URA Vol. 1, Issue 1 FY 2015/16 . . . . .	14
3.1	Boxplots showing the distribution of key import valuation variables while confirming the treatment of outliers . . . . .	24
3.2	A bar plot showing countries of origin. The country receives most of its imports from the USA . . . . .	25
3.3	In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable (y) and an independent variable (x). <b>Source:</b> Wikipedia - Linear Regression Image . . . . .	28

# List of Tables

1.1	Statistics on Import Volumes and Values in Uganda . . . . .	3
2.1	Key limitations in the adoption of AI-driven customs valuation systems . . . . .	20
3.1	Top Records of the Dataset (Transposed) . . . . .	22
3.2	Variables Description . . . . .	23
3.3	Descriptive Statistics of the Dataset (Transposed) . . . . .	26
3.4	Model Performance Metrics for Linear Regression. The performance on the training data appeared satisfactory, with reasonable MAE and RMSE values and a moderate $R^2$ . However, the model's performance on the test data indicates significant overfitting, as seen in the much poorer MAE, RMSE, and $R^2$ values. . . . .	28
3.5	Random Forest Model Performance. The model exhibited strong predictive capability with low MAE and RMSE values and a high $R^2$ on both training and test data. While there is a slight drop in performance on test data, the generalization remains significantly better compared to the linear regression model. . . . .	29
3.6	Model Performance Metrics (MAE, RMSE, and $R^2$ Scores) . . . . .	29



## CHAPTER 1

# INTRODUCTION

This chapter introduces research on the application of machine learning methods to enhance import items valuation accuracy within the Uganda's customs framework. It serves as the foundation for the research providing context, stating the problem, the research objectives and the questions that guide the research as well as justification for the research.

## 1.1 BACKGROUND TO THE STUDY

"Custom duties are a backbone of Uganda's economy with a contribution to tax revenue estimated at 30% " (URA, 2023). However, the customs framework faces significant and persistent challenges such as under-declaration of goods, misclassification of goods based on their Harmonized system codes (HS codes), inaccurate and inconsistent application of transaction value methods of valuation. These challenges have greatly impacted the country's revenue where an estimated \$ 200 million revenue is lost annually (World Bank, 2022).

The traditional imports valuation methods such as the rule-based checks and the manual audits are often too slow, subjective, labor intensive and prone to human error and manipulation. With the digital era, comes the machine learning technologies that offer transformative capacity to automate tasks and enhance accuracy in customs valuation process, streamline customs operations as demonstrated in countries that have adopted these technologies such as India, China and Kenya (UNCTAD, 2021)

The increasing volume of transactions and the complexity of the matter of international trade further exacerbate the problem calling for the adoption of automated valuation methods.

The statistics in Table 1.1 highlight the need for improvement in valuation. The volumes of imports have been on the rise year on year.

### 1.1.0.1 Contextual Background

The WTO Valuation Agreement, officially known as the Agreement on Implementation of Article VII of the General Agreement on Tariffs and Trade (GATT) 1994 (The Agreement on Customs

Year	TIV (M Tons)	TIV (USD B)	Top Imports	GDP (%)
2020	8.5	7.3	Machinery, Petroleum, Vehicles	18.2
2021	9.1	8.1	Petroleum, Electronics, Pharma	19.5
2022	9.8	9.0	Machinery, Petroleum, Food	21.0
2023	10.2	9.8	Vehicles, Chemicals, Consumer Gds	22.5

**TABLE 1.1** Statistics on Import Volumes and Values in Uganda

**Key:** TIV = Total Import Volume, TIV (USD B) = Total Import Value in Billion USD, GDP (%) = Percentage of GDP, Pharma = Pharmaceuticals, Consumer Gds = Consumer Goods.

Valuation, n.d.), replaced the GATT Valuation Code during the Uruguay Round discussions that founded the WTO in 1994. This Agreement establishes a Customs valuation system based principally on the transaction value of imported goods (the price paid or due when sold for export), subject to certain adjustments. When estimating Customs value based on transaction value is not feasible, alternative approaches must be used in the following hierarchical order:

- The transaction value of identical goods
- The transaction value of similar goods
- The deductive value method
- The computed value method
- The fallback method

These methods seek to establish a fair, standard system for valuing imported commodities that is consistent with market reality while banning arbitrary or fraudulent Customs values.

### 1.1.1 Theoretical background

The theoretical foundation of this research is anchored in artificial intelligence (AI), particularly its subset known as machine learning (ML). By adopting machine learning techniques, this study illustrates how AI can enhance accuracy and efficiency within customs processes while contributing valuable insights into its application across various sectors of public administration and policy formulation.

## 1.2 PROBLEM STATEMENT

Uganda relies heavily on customs revenue which accounts for an estimated one-third of the national tax base. This funds key public service projects and infrastructures. However, the Uganda Revenue Authority continues to face significant and persistent problems arising out of failure to appropriately address inconsistent and inaccurate valuation of imported goods resulting in an estimated \$200 million in lost revenue annually. The key identified causes of this include misclassification of goods, under declaration of goods by the importers yet is aided by the subjective valuation

methods, poor audit mechanisms and fragmented data systems. The current valuation methods in Uganda, also often referred to as the traditional valuation methods in this study mainly rely on the declarant disclosed transaction values as guided by the World Trade Organization (WTO's) transaction value means. The approach is meant to provide simplicity, but it lacks a standard mechanism to verify the declared values against the global actual benchmarks, resulting in vulnerabilities that are exploited by fraudulent schemes such as the manipulation of invoices and the misclassification of HS codes. The operational inefficiencies worsen the situation where manual system audits are reported to only cover 5% of the imports due to resource constraints. (Okello, 2022) and since Uganda's Customs databases are often isolated from external price references such as the UNcomtrade, it restricts the ability to identify price irregularities on time.

The neighbouring countries have already shown the transformative power of machine learning methods in their customs, whereas Kenya and Tanzania have recorded success in reducing discrepancies (Kiprop, 2023).

The same innovations have registered success elsewhere in developed economies such as Brazil with SISAM and Nigeria with the DATE model where the valuation of items is automated valuation giving a success rate (of 75% error reduction in Brazil) and 28% fraud detection improvement in Nigeria.

On the other hand, Uganda has been stuck on traditional rule-based valuation methods that lack adaptability to ever changing fraud schemes. This not only causes loss in revenue, but also hinders fair trade practices. In this context, the study aims to answer a very critical question: How can machine learning models be effectively applied to improve the accuracy of the valuation of imported items in Uganda's customs framework?

The study aims to provide the URA with a scalable system to minimize revenue losses and improve compliance with trade regulations by bridging technological adoption and context-specific policy implementation.

### 1.3 OBJECTIVES OF THE STUDY

- **Develop Machine Learning Models:** Tailor machine learning models specifically for Uganda's context to predict the accurate value of imported items.
- **Performance Evaluation:** Compare the developed models against traditional valuation methods using metrics such as MAE and  $R^2$ .
- **Feasibility & Impact Assessment:** Analyze the feasibility and potential impact of integrating machine learning valuation methods into Uganda's customs framework.
- **Web-Based Deployment:** Develop and deploy a web-based application to implement the machine learning model in a real-world setting.



## 1.4 RESEARCH QUESTIONS

The study seeks to answer the questions as follows:

- **Application of Machine Learning:** How can machine learning models be effectively applied to improve the accuracy of the valuation of imported items in Uganda's customs framework?
- **Performance Comparison:** What is the performance of the developed models as compared to the traditional import items valuation methods?
- **Implementation Challenges:** What are the implications and challenges of implementing machine learning valuation methods in Uganda's customs context?
- **Success Factors:** What are the success factors for the implementation of machine learning import valuation methods in Uganda's customs framework?

## 1.5 JUSTIFICATION OF THE STUDY

The study is justified by the urgent need to address the efficiency and accuracy of import valuation in Uganda's customs framework. By leveraging the power of machine learning, the research aims to contribute to the transformation of the framework into a more robust and transparent one.

The findings of the study inform policy decision-making, improve revenue collection and promote fair trade practices. The research contributes to the growing body of knowledge in the application of machine learning methods to streamline customs operations for developing economies.

## 1.6 SIGNIFICANCE OF THE STUDY

Successful implementation of machine learning mechanisms in customs valuation has notable economic, technological and policy implications for Uganda. Economically, improving the accuracy of import items valuations contributes to improved revenue collection which aligns well with the country's National development plan III (NDP III)'s objective of strengthening domestic revenue mobilization. By minimizing undervaluation, the machine learning methods help to reduce revenue leakages, hence supporting the plan's initiatives. From the technological point of view, the research introduces the first machine learning specific mechanism for Uganda's customs valuation framework. This improves efficiency and transparency in the framework hence laying a foundation for further research in digital transformation in trade and tax administration.

The study also has significant policy implications where the findings help inform the URA's digital transformation strategy of 2025. Using data-driven insights, policymakers can improve existing customs policies, strengthen fraud detection means and streamline customs operations.

Overall, the research contributes to the goal of promoting sustainable growth and development as outlined in Vision 2040 Uganda.

## **1.7 HYPOTHESIS**

The hypotheses were formulated based on the premise that the introduction of advanced technologies, particularly machine learning approaches, could significantly improve the accuracy of import valuations and streamline customs operations.

### **1.7.1 Null Hypothesis (H0)**

There is no significant difference in accuracy between the application of machine learning methods and the traditional valuation methods of imports.

### **1.7.2 Alternative Hypothesis (H1)**

Machine learning methods perform better than the traditional valuation methods in determining the accurate value of import items.

## **1.8 SCOPE OF THE STUDY**

The research focuses on import valuation in Uganda's customs framework. It considers several categories of imported items from the Uganda Revenue Authority trade data. the applicability of machine learning methods for Uganda while leveraging the country's historical import transaction data from the Uganda Revenue Authority. We explore a range of machine learning models that are best suited for the dataset while primarily focusing on predicting the accurate value of imports and evaluating the performance of these models against the traditional methods while considering the development of a web-based application to represent the models in a real-world setting.

## **1.9 THEORETICAL FRAMEWORK**

The research grounds on theories of international trade, economic regulation, and machine learning. The theoretical framework combined concepts from the Heckscher-Ohlin model of global trade, which explains how countries benefit from trading goods in which they have a comparative advantage, and regulatory compliance theory, which focuses on how effective regulatory enforcement can improve compliance and reduce fraud.

The conceptual framework emphasizes the application of machine learning and predictive analytics to process large volumes of data and identify discrepancies in values. The framework suggests that accurate and data-driven valuation models significantly improve the customs valuation process by reducing under and over-valuation of imports, thereby improving revenue collection and trade policy effectiveness.

## **1.10 Chapter Arrangement**

The rest of the thesis is organized as follows:

Chapter 2 of this thesis provides a comprehensive literature review, especially on machine learning applications in customs, and international trade with a focus on valuation and pricing.

Chapter 3 of the thesis details the methodology used to complete the study. It explains the data collection, preprocessing and statistical techniques used in the process.

Chapter 4 shows the results of the study, explains the performance of the machine learning models used and discusses the implications of the results of the models while comparing them to the performance of the traditional valuation methods.

Chapter 5 gives a summary of findings and recommendations.

Chapter 6 gives an overall conclusion and a discussion of the results.

float

## CHAPTER 2

# LITERATURE REVIEW

This section explores the theories, findings and gaps identified in the existing research that is related to the applicability of machine learning and artificial intelligence in customs operations. The analysis highlights contributions from various scholars and identifies the approach used in the previous studies while highlighting the gaps that need to be addressed.

Overall, the section serves as the foundation for our research methodology and guides the formulation of the research questions, and the different hypotheses tested to address the gaps identified.

## 2.1 Customs valuation: Concepts

Customs valuation refers to the process of determining and assigning a monetary value of imported items to assess duties and taxes. This is crucial for revenue collection and trade facilitation. The practice of valuation dates to the 1947 General Agreement on Tariffs and Trade (GATT), which established the first globally accepted principles of customs valuation (Wulf & Sokol, 2019).

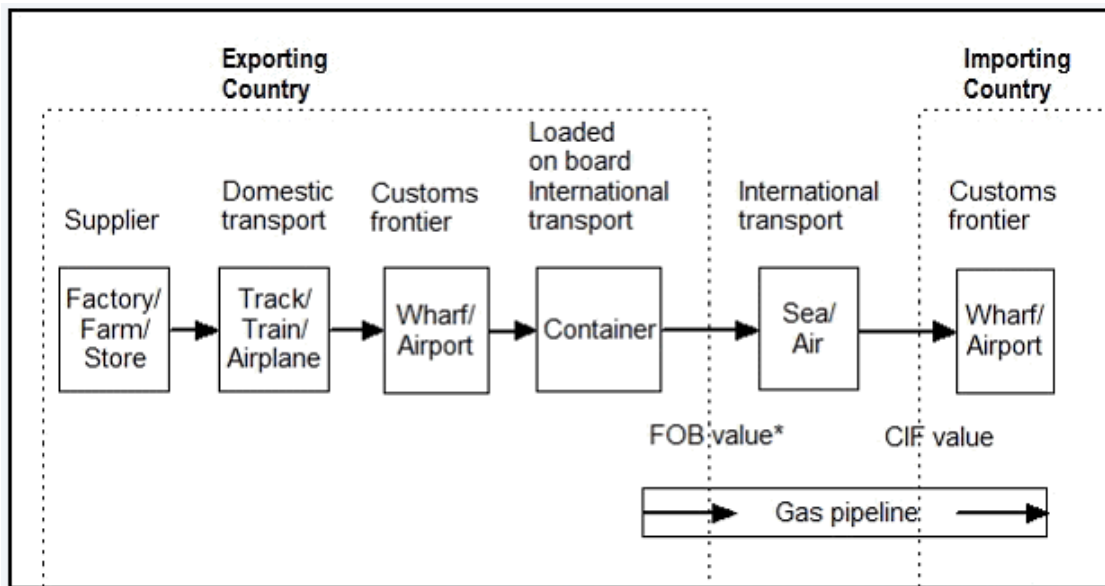
The World Trade Organization's Agreement on Customs Valuation and Actual Cash Value (ACV) later standardized these practices by providing six traditional methods for determining the customs value in a hierarchical order (WTO, 2020). These include: The transaction value method, The transaction value of identical goods, The transaction value of similar goods, The deductive method, The computed method, and the Fall-back method. The Uganda Revenue Authority (URA) as the regulator, notes that the primary method of determining the customs value of imported goods is the transaction value method as per Article 1 of the agreement on customs valuation (GATT, 1994).

The World Trade Organization (WTO) further clarified that the basic principle is the transaction value, whereby the agreement stipulated that customs valuation shall, except in specified circumstances, be based on the actual price of the goods to be valued, which is generally shown on the invoice. This price, plus adjustments for certain elements listed in Article 8, equals the transaction value, which constitutes the first and most important method of valuation referred to in the Agreement.

The methods highlighted above are often referred to as the traditional valuation methods throughout this thesis. These have been inconsistent, leading to disputes and revenue losses.

Imports valuation processes in Uganda are slow and cumbersome primarily due to the reliance on these traditional valuation methods. In contrast, adopting machine learning has revolutionized these processes by automating tasks such as document classification and risk assessment while significantly reducing processing times in countries where machine learning methods have been implemented.

Amid evolving trade policies, machine learning's capacity for data analysis allows it to predict potential disruptions, ensuring compliance with regulations while facilitating efficient trade practices.



**FIGURE 2.1** A flowchart illustrating the physical movement of goods in international trade. Goods start from a supplier in the exporting country (Factory/Farm/Store) and move through domestic transport (Truck/Train/Airplane) to the customs frontier (Wharf/Airport). They are then loaded onto international transport (Container) and shipped via sea or air. Upon arrival in the importing country, they pass through the customs frontier before reaching their final destination. Additionally, goods transported via gas pipelines are valued at the point of entry into the pipeline (FOB) in the exporting country and upon arrival in the importing country (CIF).

### 2.1.1 Machine learning in trade

Machine learning methods have become a force in transforming trade operations worldwide particularly in addressing undervaluation and detecting fraudulent schemes where these systems have been successfully installed. There has been notable success, especially in anomaly detection and improving accuracy in the valuation and classification of goods. For reference, in India,

Sharma et al. (2021) developed and deployed a random forest model to identify undervalued shipments while comparing declarant values with the trade database values and their model achieved a remarkable accuracy of 89% in identifying price discrepancies enabling the tax authorities to recover an estimated \$47 million in underpaid duties within 6 months of the model's implementation.

Similarly, Tanzania's Implementation of the XGBoost model to analyze import data registered success where valuation errors were reduced by a notable 36% with the model identifying irregularities in declarations and misclassification of goods by their HS codes (Kiprop, 2023) which highlights the potential of machine learning to handle non-linear relationships and high dimensional data that is inherent in customs trade records. However, supervised learning methods have a notable limitation in low compliance environments such as for Uganda where labeled datasets of recorded fraud cases are scarce, Ferreira et al. (2020) clearly warn against models trained on such incomplete or biased datasets which was evidenced with the Nigeria's customs union where a Neural Network flagged 20% of legitimate shipments due to over-fitting on outdated fraud patterns.

This emphasized the need to employ methods that incorporate both unsupervised and supervised learning methods such as clustering to reduce over reliance on labeled data while adapting to ever changing fraud tactics.

#### **2.1.1.1 Predictive Analytics in Customs**

Predictive analytics is the application of statistical techniques and machine learning (ML) models to analyze past data and forecast future trends and results. Predictive analytics is critical in trade facilitation because it improves decision-making, lowers risks, and optimizes the efficiency of international trade operations. Predictive algorithms can forecast market trends, identify potential logistics bottlenecks, and improve customs risk assessments by using massive amounts of trade-related data.

Predictive analytics is critical in trade facilitation because it improves decision-making, lowers risks, and optimizes the efficiency of international trade operations. Predictive algorithms can forecast market trends, identify potential logistics bottlenecks, and improve customs risk assessments by using massive amounts of trade-related data.

Time series analysis, regression models, and classification algorithms are the most common predictive models utilized in trading processes today. Time series forecasting, for example, is often used to forecast shipment arrivals and delays using historical patterns as well as external factors such as weather and geopolitical events. Regression models are used to assess the relationships between numerous trade parameters and their influence on logistics performance. Furthermore, categorization algorithms assist in identifying high-risk transactions by analyzing trends in trade documents, shipment data, and compliance histories.

Despite these advancements, the growing complexity and volume of international trade necessitates the development of more sophisticated predictive models capable of managing large and

diverse datasets. This has sparked a rising interest in using sophisticated ML approaches to improve the effectiveness of predictive analytics in streamlining customs trade operations.

The application of machine learning techniques in customs gained traction in recent years and a study by Jentsch et al., (2019) noted the potential of machine learning in predicting accurate customs values based on historical customs records. Their study showed promising results in figuring out cases of undervaluation. Similarly, Vetter et al., (2021) explored the application of ensemble learning algorithms for detecting customs fraud, whilst combining more than one algorithm to enhance accuracy and robustness over the traditional rule-based techniques.

Applications of these ML algorithms have shown considerable gains in predicting trade patterns and risk assessment. However, integrating these models into existing trade systems presents several obstacles, which will be described in the following section.

#### **2.1.1.2 HS code classification with Naive Bayes Algorithm**

The HS code is a 6-digit international numerical code that is used to designate and identify goods in international trade. In addition to the internationally recognized 6-digit number, each country may add further digits to the code to make it 8, 10, or 12-digit for tariff and statistics purposes. HS Classification is the process of determining the most exact description in the harmonized system (HS) for the commodities to be classified.

The suggested research now focuses on unsupervised learning techniques, notably Naive Bayes classification, for categorizing and predicting labeled data. The Naive Bayes strategy is a simple and effective method for multivariate classification. The model's objective is to predict the HS Code based on the dataset and variables that have been created using the Naive Bayes algorithm. It is envisaged that by limiting these risks, the state's revenues will be maximized through the determination of suitable tariffs and/or customs values on imported commodities.

(Muslim, 2022) highlighted the power of the Naive Bayes algorithm when used in the classification of HS codes for optimizing customs revenue and mitigation of potential restitution achieved a remarkable accuracy of 99.97% with a classification error of only 0.03% which demonstrated how data mining techniques optimize customs revenue and therefore reducing the risk of unpaid duties when applied.

Overall, the issue that usually arises is the return of unpaid import duty and/or administrative punishments in the form of fines based on the objection decision. The application of data mining techniques is intended to give useful information regarding the HS Code categorization technique, which can help Customs officials determine tariffs and/or customs values.

## 2.1.2 ML-Powered Customs Operations

CHEN, Z. (2024) highlighted the revolutionary impact of machine learning when it was applied to a customs dataset, discovering trends and anomalies. He goes on to say that by automating the valuation process, machine learning methods were able to improve risk assessment and detect discrepancies in declared values, allowing customs authorities to make more informed decisions, reduce errors, and ultimately lead to a more accurate and efficient valuation of items.

### 2.1.2.1 Brazil

Since 1997, the country's import declarations have been logged in Siscomex, an integrated commerce system. If errors are discovered during inspection by a customs officer, a corrected copy of the declarations is saved, and both copies are retained indefinitely. The AI system utilized is SISAM, which learns from both versions of the dataset to enhance mistake detection. To handle the dataset's many properties, Bayesian approaches with smoothing hierarchies are used. It uses both supervised and unsupervised learning approaches to adjust to legislation changes without the need for retraining, allowing the system to maintain high accuracy in its classifications and predictions.

If more than 75% errors are identified in an import declaration, SISAM advises a physical examination by customs officials. The system's handling of complex errors currently outperforms random selection. (Artificial Intelligence in the Customs Selection System via Machine Learning (Sisam) 1, n.d.).

### 2.1.2.2 China

In recent years, China Customs has continued to use technology and innovation to address the contradiction between an ever-increasing Customs control burden and insufficient regulatory resources.

**AI-based NII (Non-intrusive detection devices) image recognition system.** Based on the expertise of Customs officers who conduct Customs inspections using NII devices, this system employs artificial intelligence technology to learn information about goods and articles from massive historical H986 (Large-scale container X-ray scanner) and CT (Computed Tomography) inspection images and creates automatic recognition algorithms. With a huge number of information on commodities, articles, and modes of transportation, the system can automatically recognize photos and alert Customs officials to perform image reviews or physical inspections. "Through continuous optimization, the ultimate goal of this system is to replace human beings with machines in the field of NII inspection". (ANNEX-The Case Studies 109 Study Report on Disruptive Technologies, n.d.).

**Impact:** One of the initial effects of implementing the Autonomous Selection of Algorithms model was to free up some capacity on local IT servers, allowing for faster algorithm computation times.



”Furthermore, statistics reveal that once the model was deployed, the accuracy of automated image analysis on large-scale NII devices increased by around 5%, while the false alarm rate was reduced by about 8%. CT scanners’ accuracy increased by approximately 6%, while false alert rates were reduced by almost 5%”. ( WCO News 104 - Issue 2 / 2024 )

***Intelligent Passenger Face Recognition System.*** This system uses face recognition technology and is linked to the low-temperature detection system for quarantine and inspection. It has been implemented in various Customs by placing facial recognition cameras in control areas classified into three categories: Customs alerting area, Customs processing area, and Customs re-exam area.

”Key passengers (including blacklist passengers, multiple cross-border passengers, and high-risk passengers for inspection and quarantine) walking through these three operational areas will be spotted and Customs officers who are equipped with hand-hold mobile devices and face recognition devices will stop them for further investigation”(Annex-The Case Studies 109: Study Report on Disruptive Technologies, n.d.). A passenger information database has been created and gradually developed, allowing for the filtering and analysis of relevant images and videos. Customs can consequently conduct risk analysis, profiling, and query statistics.

At present, the alarm accuracy rate of the system is over 99%. It plays a vital role in fighting against “high-risk traffickers”, and several smuggling gangs have been apprehended. At the same time, due to the characteristic of being “non-intrusive”, the efficiency of Customs clearance for passengers has been greatly improved. In the future, China Customs will explore more possibilities to make passenger inspection smarter and provide better services for inbound and outbound passengers.

### **2.1.2.3 Belgium: BCTC Behavioral consequences of tariff changes**

This research examines the impact of EU Customs tariff measures on commodity trade flows.

The primary purpose is to detect fraudulent activity by economic operators following the implementation or rise of tariff measures. These protectionist policies aim to safeguard the European Union’s internal market by sheltering domestic producers and protecting industries from international competition. Attempts to escape imposed taxes are frequently made using various fraud tactics, resulting in revenue losses for the Union and damage to involved European industries.

Based on historical data, two potential fraud strategies are being investigated: the declaration of a fraudulent country of origin, a false product code, or a combination of the two. More precisely, the initiative seeks to detect sudden behavioural changes in an operator’s import profile that deviate significantly from the “normal” trends recorded before the tariff measure is applied.

### 2.1.2.4 The RECTS project: Uganda

”Three revenue authorities (Kenya, Uganda, and Rwanda) agreed in 2017 at a tripartite head of state meeting to establish a Regional Electronic Cargo Tracking System (RECTS) hosted by the Revenue Authorities to ensure data security, provide end-to-end tracking across partner states’ borders, and provide tailored cargo tracking and monitoring solutions. Bsmart Technologies was contracted as the sole provider of the new regional system. Julius and Christabel (n.d.).

The system consists of four major components: dry cargo seals and wet cargo fuels, arming personnel at the release locations, the Centralized Monitoring Center at the head, and twelve Rapid Response Units throughout the transit route. All teams operate around the clock to ensure real-time cargo monitoring while in transit.

In addition, a reconciliation team checks all transit cargo movement documentation to guarantee compliance with legislation governing their transportation and that any malpractices discovered are corrected.



**FIGURE 2.2** A customs officer inspecting a cargo container equipped with the RECTS tracking system, ensuring secure and transparent transit monitoring. Picture Source: URA Vol. 1, Issue 1 FY 2015/16

Benefits realized after the implementation of the project include:

- 1) Transit duration decreased to three to four days on average, resulting in shorter transit times.
- 2) Increased revenue due to Rapid Response Unit interceptions.
- 3) Improved data control to protect its integrity.

- 4) Improved regional coordination and integration of joint technical working groups.
- 5) Real-time cargo monitoring and faster incident reaction times of 60 minutes.
- 6) A decrease in cargo diversion cases.
- 7) Reduced business costs.

-6

#### **2.1.2.5 Ongoing enhancement initiatives**

The combined technical working groups of the revenue authorities are constantly examining the operational modules of the RECTS system in light of the operating environment and suggesting changes aimed at improving operational efficiency. There are plans to build more Rapid Response Units throughout all transit corridors to strengthen the presence and monitoring of items under Customs control. This requires allocating resources to all RECTS components to ensure that they are well-equipped and that personnel are sufficiently trained to perform cargo monitoring duties.

#### **2.1.2.6 The DATE model**

Chen and Liu (2022) developed this, which showed a significant improvement in customs data processing. "DATE (Dual Attentive Tree Embeddings) demonstrated that tree-structured attention models can capture relationships in customs declarations, improving accuracy by 28% compared to traditional machine learning methods."

The World Customs Organization's BACUDA project introduced the model, which represents a significant shift in customs fraud detection with notable findings including superior performance as compared to the XG Boost model, especially in classification tasks. This model was also successfully used in Nigeria (FSI, 2023). Additionally, a simpler web-based user interface was used to effectively prevent item misclassification.

FSI (2023) conducted a study on the use of artificial intelligence in developing Nigeria's customs system, and the DATE model demonstrated superior performance to existing traditional methods. To avoid misclassification of imported items, relevant users are given a web link where they can enter the item name and another unique identifier, and the model returns the correct class to which those items belong. This study fills gaps in the existing literature by creating a machine-learning model to address valuation issues inside Uganda's customs regime.

### **2.1.3 Importance of Machine Learning in Trade Facilitation.**

Machine learning (ML) has emerged as a strong technology that has the potential to transform trade facilitation. ML identifies patterns, trends, and anomalies in large datasets by employing

---

<sup>-6</sup>A machine learning approach for accurate import valuation

data-driven algorithms. This functionality is especially useful in trade environments that create massive amounts of data daily.

**Automated risk assessment,** real-time customs clearance time prediction and supply chain operation optimization are all examples of machine learning applications in trade facilitation. Predictive models, for example, can analyze past customs data to identify shipments that are likely to have delays or violations, allowing customs officials to better deploy resources. Similarly, machine learning-based risk assessment models can aid in the identification of high-risk consignments, boosting inspection accuracy and efficiency. These examples highlight machine learning's transformational potential in improving trade processes and reducing inefficiencies.

**Enhanced efficiency:** The implementation of machine learning systems in customs trade streamlines operations by reducing costs and improving efficiency. While analyzing large trade datasets, these systems reduce human error, automate customs clearance, and reduce delays in document processing hence enhancing efficiency.

**Anomaly detection:** Machine learning systems often support classifying trade transactions through outlier detection mechanisms.

#### 2.1.4 Barriers to successful implementation of Machine learning in Customs Trade Practices

**Data quality and its availability:** The use of predictive analytics and machine learning models in trade facilitation confronts several challenges, the most significant of which are data quality issues and system integration. Machine learning systems often rely on high-quality data to operate smoothly. Biased data leads to incorrect results and conclusions which undermines the reliability of the results of the ML systems.

Customs trade uses a variety of data sources, such as customs records, shipping manifests, and regulatory documents. These datasets' discrepancies, missing values, and inconsistencies can have a major impact on predictive model accuracy.

**Ethical and Regulatory Considerations:** The adoption of machine learning models sometimes inadvertently perpetuates biases that are often present in the training data which leads to unfair results. Furthermore, the sensitivity of trade data poses privacy and security concerns, restricting the availability of large datasets for model training and analysis.

**Technological Constraints:** Successful implementation of machine learning systems requires special skills and knowledge. The lack of enough skilled professionals in Artificial Intelligence hinders the adoption and effective adoption of these technologies. This calls for substantial investment in infrastructure especially investing in high-performance computing resources and data storage mechanisms.

**Public Perception:** Building public trust in these AI systems is critical. There is a constant

fear of loss of jobs due to many tasks being automated by AI. Therefore addressing the societal impact of AI systems requires the development of programs to up-skill the workforce and to provide opportunities for employees to acquire new skills.

**AI Regulation:** The lack of an organized legal framework for data collection and use of electronic data. The absence of standardization in data formats and communication protocols makes ML model integration problematic. In addition, traditional trade systems frequently lack the technical infrastructure needed to support real-time ML predictions.

These problems underscore the importance of improved data management techniques and strong integration frameworks to properly employ ML in trade facilitation.

### 2.1.5 Gaps in Existing Literature

Despite the demonstrated success of machine learning in neighboring economies, Uganda's customs ecosystem has been noticeably underserved by such innovations. Even though Uganda shares structural challenges such as porous borders, informal trade networks, and limited audit capacity with Kenya and Tanzania, where ML adoption has yielded measurable gains, no peer-reviewed studies or policy frameworks have explored the application of ML to Ugandan import valuation. Existing research in the region is primarily concerned with fraud detection as a standalone goal, frequently ignoring integration with the WTO's transaction value methodology or Uganda's legal valuation frameworks. Models developed in Kenya, for example, prioritize identifying outright smuggling while failing to address more nuanced forms of undervaluation, such as manipulating invoices for high-risk goods such as used vehicles and textiles (Nattuthurai, 2021). This creates a critical gap: while anomaly detection models can identify suspicious shipments, they lack the methodological accuracy to establish contextually accurate values that are consistent with global price benchmarks or Uganda's tariff laws. Furthermore, Uganda's unique data landscape, which includes sparse digitization of paper-based declarations, and inconsistent HS code granularity, has received little attention in the ML literature.

Bridging this gap necessitates not only adapting proven algorithms to Uganda's infrastructural realities, but also reviewing model design to align with the URA's operational priorities, such as minimizing clearance times while increasing revenue recovery. By addressing these issues, this research seeks to advance a comprehensive ML framework that goes beyond fraud detection to provide actionable, policy-compliant valuation findings

Despite the rising volume of research on predictive analytics and machine learning applications in trade, there are still several gaps in the present literature. One notable gap is the limited investigation into the applicability of machine learning to the context of developing economies that face challenges with data integrity and transparency.

While some studies have looked at individual ML algorithms, there has been little research into merging these approaches with decentralized data storage options to increase forecast accuracy and

trust.

Another undiscovered field is the use of reinforcement learning in complex trading scenarios with several stakeholders and dynamic conditions. While preliminary findings are promising, additional research is needed to construct scalable reinforcement learning models capable of optimizing whole trade networks in real-time.

The scope of the research noted insufficient study of data approaches, especially for Uganda's context. While some countries have successfully implemented data-driven approaches, Uganda's special trade patterns, border practices and existing technological infrastructure presented distinct challenges that are yet to be addressed adequately in the current research. For instance, the integration of data from the Automated System for Customs Data (ASYCUDA) world system and extraction of regional trade databases data requires special approaches that consider local data quality and accessibility constraints.

There exists a gap in understanding the real impact of AI-based valuation systems on revenue collection. While the theoretical analysis suggests potential improvement, there has been limited empirical evidence that quantifies the real-world impact of such systems on customs revenue especially in developing countries. Research needs to not only focus on the revenue impact but also on other effects such as reduced clearing times, decrease in valuation disputes and cases and improved compliance of traders.

There also exists challenges in implementing the technology advancements as highlighted in this research. The current literature predominantly focuses on the implementation of these innovations in well-developed economies, but a gap remains regarding the adoption of these innovations for developing economies like Uganda that are associated with challenges like intermittent internet connectivity and specialized manpower. For instance, the implementation of some algorithms like BERT models in environments with inadequate computing resources needs further exploration.

There is a crucial need for research on the institutional resistance to change regarding customs modernization, staff adaptation and organizational change at large needs to be studied more.

Finally, there is a need for comprehensive research into the long-term influence of predictive analytics on customs trade practices. The existing literature focuses mostly on short-term improvements in efficiency and risk management, leaving the broader economic, social, and regulatory ramifications unexplored.

These gaps point to potential future study topics, emphasizing the necessity of multidisciplinary techniques that combine machine learning, blockchain, and trade research to progress the subject of customs trade and valuation efficiency.

### **2.1.6 Conceptual Framework for Machine Learning Powered Customs Valuation**

Based on the literature reviewed, the conceptual framework for the installation of machine learning methods can be developed while considering important steps such as data acquisition,

data preprocessing, feature engineering, model selection and training, performance evaluation and finally integration with existing customs systems used in operations. The framework also considers legal and regulatory aspects of customs valuation, ensuring full compliance with both the international and national set regulations.

### **2.1.7 Limitations of the study**

Although the purpose of the study was to provide comprehensive solutions for import assessment, several obstacles must be acknowledged. While the study aimed to develop models suitable for all import categories, the unique characteristics of certain goods required further adaptation and refinement of the models.

External factors such as changes in international trade policies, economic fluctuations and global supply chain disruptions can affect the effectiveness of proposed solutions, requiring constant monitoring and adaptation to meet these dynamic conditions. Recognizing these limitations, research can strategically address potential challenges and provide robust, adaptive solutions to improve the value of Uganda's imports.

The accuracy and efficiency of machine learning methods are extraordinarily dependent on the volume and completeness of available information. Incomplete or incorrect information may also influence the reliability of the results. Deploying these models requires tremendous technical infrastructure and expertise, and URA's resource constraints can present situations in deploying and maintaining those systems.

Institutional resistance to exchange and the need for considerable education of customs officials can prevent the adoption of new innovative methods, which require careful planning and stakeholder engagement.



### 2.1.7.1 Limitations summary

Limitation	Description	Reference(s)
<b>Data Quality</b>	Issues such as incomplete, inaccurate, and inconsistent data affect model reliability and decision-making.	Grundy (2015)
<b>Data Availability</b>	Limited access to comprehensive datasets hinders effective model training and evaluation.	KRA (2020)
<b>Risk Management</b>	Challenges in accounting for complex and dynamic risk factors in trade and customs operations.	eClear (2020)
<b>Implementation</b>	Constraints related to limited resources, inadequate infrastructure, and lack of technical expertise.	Szabo (2017)
<b>Explainability</b>	Difficulty in interpreting and explaining AI models reduces trust and regulatory acceptance.	WCO (2020)

**TABLE 2.1** Key limitations in the adoption of AI-driven customs valuation systems

## 2.2 Conclusion

The literature review section highlighted the potential of machine learning to enhance customs valuation accuracy. By addressing the challenges noted and leveraging the opportunities for Uganda's context, machine learning innovations can contribute to a more efficient customs framework, promoting fair trade practices and improved revenue collection. The chapters below present the research's methodology, Results and further analysis.



## CHAPTER 3

# METHODOLOGY

### 3.1 Introduction and Scope

This study employed a comprehensive analysis of Uganda's import dataset, which contains a wide range of import transactions and other related variables. Statistical and data analysis techniques were applied to uncover associations and trends.

We acknowledge that the findings may have broader applicability globally in customs operations concentrating on import data from 2013 to 2023, allowing for a great analysis of valuation trends patterns and discrepancies in the valuation processes over a longer period. Overall, the study is multidisciplinary as it combines data science, trade operations, customs operations and statistical analysis to ensure that the customs valuation issue is comprehensively covered.

The dataset was accessed through the Uganda Revenue Authority's research team. The primary variables of interest include Harmonized system commodity code (HS code), Country of origin (code), Unit price in local currency, Valuation method, CIF (Cost, Insurance, and Freight) value in local currency, and Item price.

### 3.2 Data Description

The dataset comprises 10 columns, each providing valuable information related to imported goods. Dr Yufeng Guo's seven steps of machine learning were applied in this study. These included data gathering, data preparation, model selection, model training, model evaluation, parameter tuning and prediction. Different data exploration techniques were applied to address any missing values, outliers and duplicates in the data. Visualization techniques included bar charts, and scatter plots among others to depict the distribution of various variables further.

Feature	Record 1	Record 2	Record 3	Record 4	Record 5
HS Code	91338102.5	91338102.5	85181000.0	91338102.5	85163100.0
Country of Origin	CN	CN	CN	CN	CN
Gross Mass	194.54	194.54	77.82	972.72	155.64
Net Mass	150.0	150.0	50.0	900.0	124.0
Item Price	1285.39	1344.86	102.91	1543.66	164.66
CIF Value (Local)	1285.39	1344.86	102.91	1543.66	164.66
Duties	447.56	468.26	35.81	537.49	92.55
Unit Price (Local)	257.08	268.97	51.46	61.75	41.17
Invoice Amount (Local)	1285.39	1344.86	102.91	1543.66	164.66
Currency	USD	USD	USD	USD	USD
Added Costs	0.0	0.0	0.0	0.0	0.0
Internal Freight	0.0	0.0	0.0	0.0	0.0

**TABLE 3.1** Top Records of the Dataset (Transposed)

### 3.3 Data Pre-processing

Data preprocessing is an important stage in Machine Learning because it improves data quality and promotes the extraction of relevant insights from it. Data preparation in Machine Learning is the process of preparing (cleaning and organizing) raw data to create and train Machine Learning models. In layman's terms, data preprocessing in Machine Learning is a data mining approach that converts raw data into a readable format. This involved renaming columns to give them meaningful names as well.

#### 3.3.1 Handling Missing Data

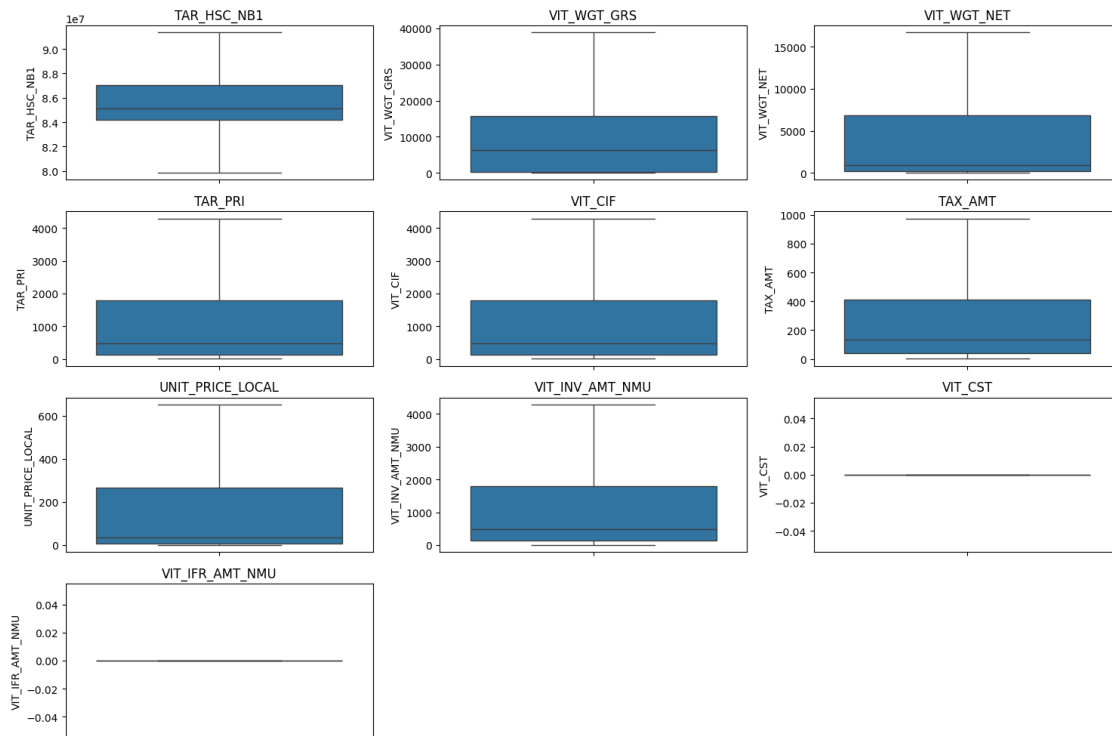
The dataset went through a cleaning phase where missing data was detected in different columns and addressed by dropping where they were over 60% apart from the Target variable (Unit price). The percentage of missing values in a column that should have been dropped was determined by the dataset and analytic objectives. However, as a starting point, a column with a significant percentage of missing values (more than 50%) was dropped entirely, especially if it is not critical to the analysis. In contrast, in a column with a low percentage of missing data (less than 10%), imputation techniques were employed instead of removing entries. In circumstances when a column contained a moderate number of missing values (between 10% and 50%), the impact on the analysis was carefully considered before determining whether to eliminate missing values or employ imputation methods.

**TABLE 3.2** Variables Description

Variable Name	Description
HS_Code	The Harmonized System code, a standardized classification of imported goods used globally.
Country_of_Origin	The country from which the goods originate, which may influence valuation due to trade agreements or tariffs.
Gross_Mass and Net_Mass	These variables provide insights into the weight of the imported goods, affecting transportation and duty costs.
Item_Price and CIF_Value	The declared price of items and their Cost, Insurance, and Freight (CIF) value, which represents the total landed cost.
Duties_Taxes	The taxes and duties levied on the imported goods based on their category and origin.
Unit_Price_Local	The price per unit of the goods in the local currency, crucial for understanding pricing variations.
Valuation_Method	The method used by customs officials to determine import values, though this field may have inconsistencies or missing values.
Invoice_Amount_-NMU	The total invoice amount in the Non-Monetary Unit (NMU), which helps in determining valuation accuracy.
Invoice_Currency_-Code	The currency used in invoicing, requiring conversion to a standardized currency for comparative analysis.
Additional Costs and Freight Charges	These include extra expenses such as handling fees, internal freight, and storage costs.

### 3.3.2 Outlier Detection

An outlier is an observation that deviates from other values in a random sampling of a population. These were detected in the dataset using the Interquartile Range (IQR) which is a statistical dispersion metric. It denotes the range in which the middle 50% of the data lies. The IQR is calculated by subtracting the 75th percentile (Q3) from the 25th percentile (Q1). The approach was chosen due to its ability to handle skewed data distributions. It detects outliers based on percentiles, making it less susceptible to extreme numbers. **Data trimming** technique was employed and any values below the lower bound or above the upper bound were replaced with the nearest valid bound. The outliers were checked again after the trimming technique to ensure data quality and integrity as seen below.



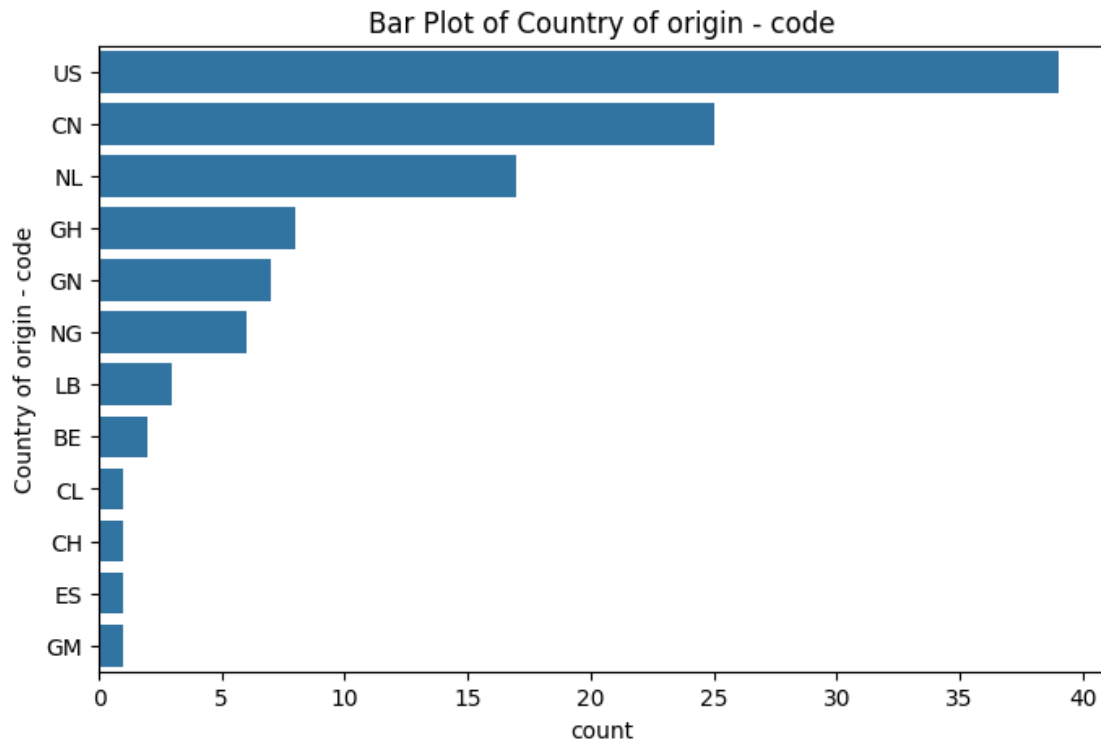
**FIGURE 3.1** Boxplots showing the distribution of key import valuation variables while confirming the treatment of outliers

### 3.3.3 Exploratory Data Analysis: (EDA)

This is a key process as it enables Data scientists to analyze and study data sets, as well as summarize their key properties, frequently using data visualization techniques. The process determines how to best modify data sources to obtain the answers required, making it easier to detect patterns, identify anomalies, test a hypothesis, or confirm assumptions.

#### 3.3.3.1 Data Visualization

Data visualization is the process of displaying data or information using graphs, charts, or other visual representations. Using visuals, we can better comprehend how data is related. Data visualization is another type of visual art that captures our attention and keeps us focused on the message. When compared to scanning rows of data on a spreadsheet, converting information into images makes it easier to see patterns, trends, and outliers. Since the goal of data is to reveal insights, visualized data is far more beneficial. Below is an extract on the count of the country of origin variable. With a bar plot, it is easy to identify the country with the highest number of imports to Uganda.



**FIGURE 3.2** A bar plot showing countries of origin. The country receives most of its imports from the USA

### 3.3.3.2 Descriptive Statistics

Mean, median and standard deviation were used to analyze the import valuation data. On the other hand, inferential statistics included testing the hypotheses and regression analysis. Python Programming language version 3.11.9 was used for the processing and analysis tasks. Descriptive statistics techniques assisted in defining the data by measuring its frequency, positioning, variance, and central tendency, among other things.

Statistic	Count	Mean	Std Dev	Min	25%	50%	75%
TAR_HSC_NB1	111	80,177,240	17,805,080	2,064,900	84,163,450	85,131,000	87,033,
VIT_WGT_GRS	111	9,737.87	20,139.52	1.00	200.00	6,352.00	15,664.
VIT_WGT_NET	111	5,638.93	7,616.08	1.00	200.00	920.00	6,804.0
TAR_PRI	111	3,361.90	8,487.76	6.81	140.17	471.88	1,800.2
VIT_CIF	111	3,387.01	8,490.17	6.81	140.17	471.88	1,800.2
TAX_AMT	111	784.88	2,300.13	1.45	38.83	136.13	413.30
UNIT_PRICE_LOCAL	111	1,871.22	6,908.06	0.02	5.00	33.85	264.23
VIT_INV_AMT_NMU	111	3,361.90	8,487.76	6.81	140.17	471.88	1,800.2
VIT_CST	111	25.12	264.63	0.00	0.00	0.00	0.00
VIT_IFR_AMT_NMU	111	25.12	264.63	0.00	0.00	0.00	0.00
VIT_OTC_AMT_NMU	111	0.00	0.00	0.00	0.00	0.00	0.00

**TABLE 3.3** Descriptive Statistics of the Dataset (Transposed)

### 3.3.4 Feature Engineering

Feature engineering transforms raw data into a machine-readable structure. It improves ML model performance by converting and selecting relevant features.

Feature engineering is the process of transforming raw data into useful information for machine learning models. In other terms, feature engineering refers to the process of developing predictive model features. A feature, sometimes known as a dimension, is an input variable that generates model predictions. Because model performance is heavily dependent on the quality of data used during training, feature engineering is an important preprocessing strategy that entails identifying the most relevant parts of raw training data for both the prediction job and the model type under consideration.

- **Price Deviation Ratio (PDR):** Calculated as  $\frac{\text{Declared Value}}{\text{Comtrade Benchmark}}$ , values exceeding  $\pm 20\%$  were treated as high-risk.
- **Importer Risk Index:** A weighted score based on historical discrepancies and shipment frequency.

### 3.3.5 Data Normalization

All numerical variables (e.g., Unit\_Price\_Local, CIF\_Value) were standardized using Min-Max scaling to ensure equal contribution during model training.

## 3.4 Model Development

### 3.4.1 Algorithm Selection

This study applied various machine learning algorithms since the task was a regression task in nature and these included: The linear regression model, The random forest algorithm, XG Boost and the Neural Networks.

#### 3.4.1.1 Linear Regression Model

This model determines the coefficients of a linear equation using one or more independent variables that best predict the value of the dependent variable. Linear regression finds a straight line or surface that minimizes the difference between expected and actual output values. Linear regression models are generally straightforward, providing an easy-to-understand mathematical formula for generating predictions.

## 3.5 Model Selection

### 3.5.1 Linear Regression Model

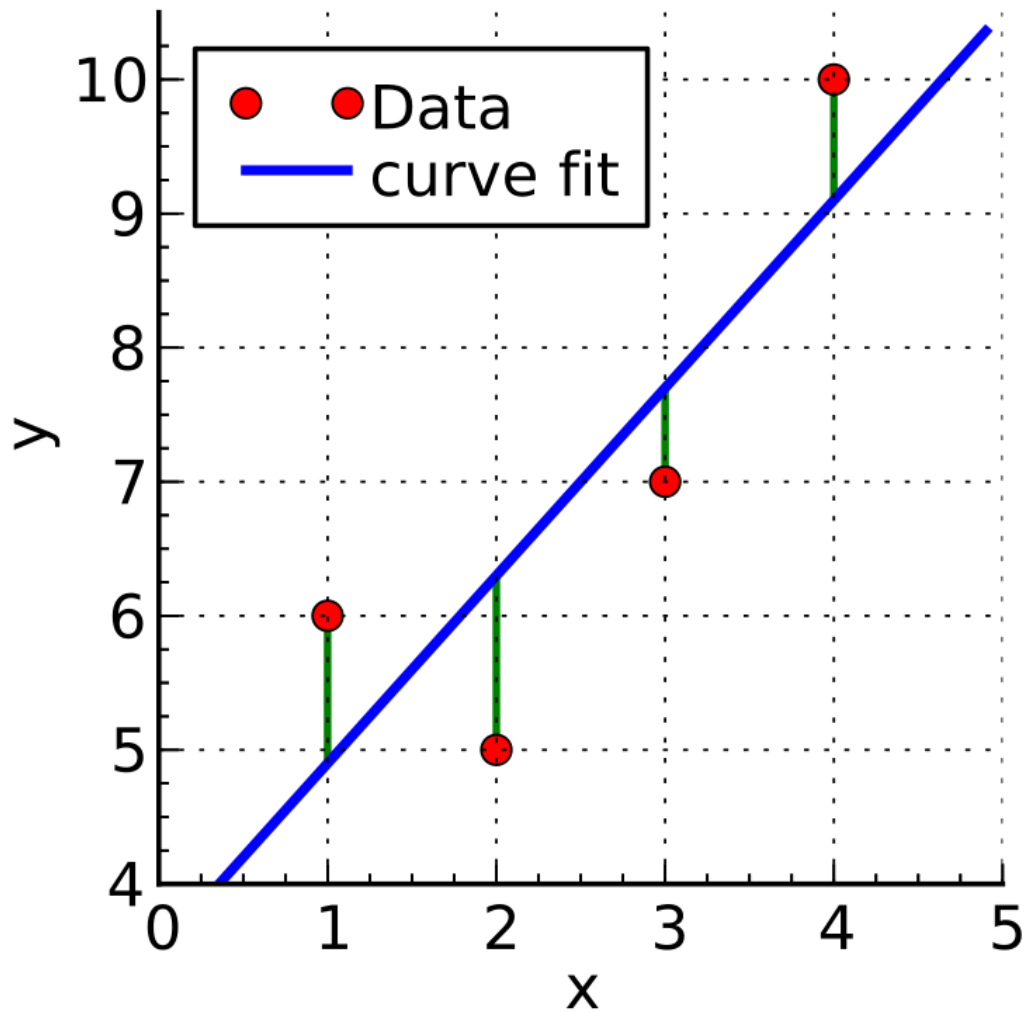
A simple linear regression model can be expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon \quad (3.1)$$

where:

- $y$  is the dependent variable (e.g., predicted import valuation).
- $x$  is the independent variable (e.g., weight, unit price, or tax amount).
- $\beta_0$  is the intercept (baseline value when  $x = 0$ ).
- $\beta_1$  is the slope coefficient (effect of  $x$  on  $y$ ).
- $\epsilon$  is the error term (unexplained variance).

### 3.5.1.1 Graphical Representation



**FIGURE 3.3** In linear regression, the observations (red) are assumed to be the result of random deviations (green) from an underlying relationship (blue) between a dependent variable ( $y$ ) and an independent variable ( $x$ ).

**Source:** Wikipedia - Linear Regression Image

Dataset	MAE	RMSE	$R^2$
Training Data	0.5154	0.6882	0.5508
Test Data	0.6248	0.8318	0.0471

**TABLE 3.4** Model Performance Metrics for Linear Regression. The performance on the training data appeared satisfactory, with reasonable MAE and RMSE values and a moderate  $R^2$ . However, the model's performance on the test data indicates significant overfitting, as seen in the much poorer MAE, RMSE, and  $R^2$  values.



### 3.5.2 Random forest model

Random forest is a popular machine learning technique developed by Leo Breiman and Adele Cutler that combines the outputs of numerous decision trees to produce a single outcome. Its ease of use and adaptability fuelled its popularity and confirmed why it was chosen for the task, as it can handle both classification and regression problems.

Dataset	MAE	RMSE	R <sup>2</sup>
Training Data	0.1696	0.2572	0.9373
Test Data	0.2404	0.3497	0.8316

**TABLE 3.5** Random Forest Model Performance. The model exhibited strong predictive capability with low MAE and RMSE values and a high R<sup>2</sup> on both training and test data. While there is a slight drop in performance on test data, the generalization remains significantly better compared to the linear regression model.

### 3.5.3 Training and Validation

The dataset was split into an 80% training set and a 20% test set. Stratified sampling ensured a proportional representation of HS chapters. Model performance was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

### 3.5.4 Model performance comparison

**TABLE 3.6** Model Performance Metrics (MAE, RMSE, and R<sup>2</sup> Scores)

Model	MAE		RMSE		R <sup>2</sup> Score	
	Train	Test	Train	Test	Train	Test
Linear Regression	0.5154	0.6248	0.6882	0.8318	0.5508	0.0471
Ridge Regression	0.5412	0.5245	0.7103	0.6761	0.5216	0.3705
Lasso Regression	0.8937	0.7940	1.0269	0.8992	0.0000	-0.1136
Random Forest	0.1696	<b>0.2404</b>	0.2572	0.3497	0.9373	<b>0.8316</b>
XGBoost	0.0052	0.2423	0.0076	0.5066	0.9999	0.6465
Neural Network	0.3941	0.3037	0.6045	0.3848	0.6534	0.7961

Note: Bold values indicate the best performance in the test set. MAE = Mean Absolute Error, RMSE = Root Mean Squared Error.

### 3.5.5 Model deployment

The following were used as the platforms to deploy the model: Streamlit application was built due to its simplicity and it offers great options for sliders, text input forms and being user friendly. Render was utilized to host the Streamlit application and this was chosen due to its ability to offer more services for free.

# Bibliography