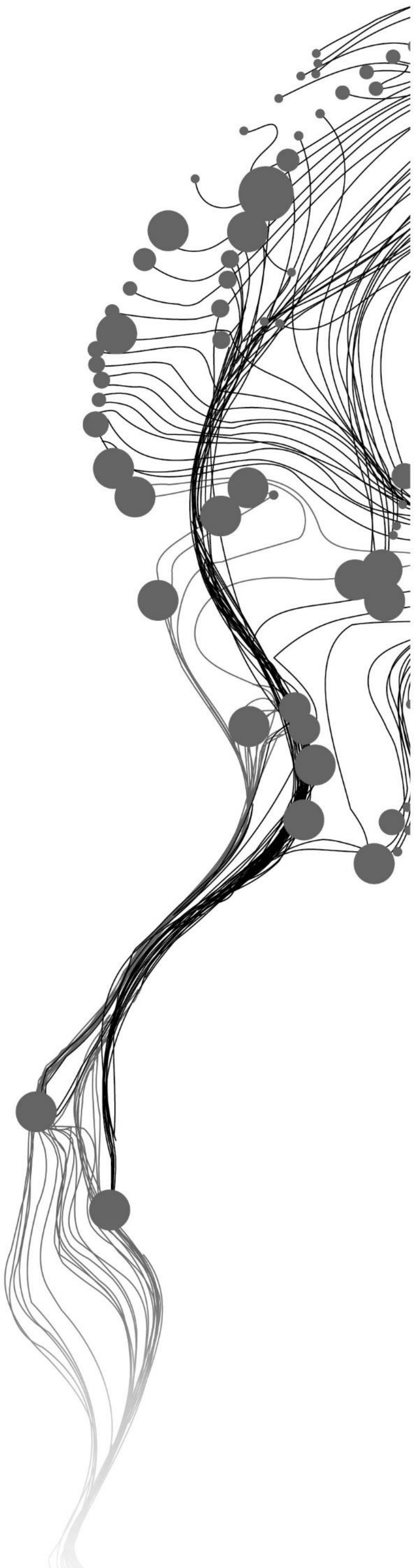
FOODGRAVITY: UNDERSTAND FOOD FLOWS USING CLASSIC GRAVITY MODEL AND EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES

BELISE DUSABE
July, 2024

SUPERVISORS:
Dr. Dou Yue
Dr. Claudia Paris



FOODGRAVITY: UNDERSTAND FOOD FLOWS USING CLASSIC GRAVITY MODEL AND EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES

BELISE DUSABE
Enschede, The Netherlands, JULY, 2024

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geoinformatics

SUPERVISORS:
Dr. Dou Yue
Dr. Claudia Paris

THESIS ASSESSMENT BOARD:
Dr. Ir.T.A. Groen
Dr. Ziga Malek

ABSTRACT

Mapping food flows from production areas to consumption areas is essential and often challenging, especially at local scales (Moschitz & Frick, 2021). Knowing how food moves over space and time is crucial for policy-making to maintain food and nutrition security across scales. Nevertheless, there is a tendency to prioritize flows between countries at the global level i.e., Food and Agriculture Organization trade data (FAO, 2023a) whereas the internal food flows within a country are often neglected. This oversight can lead to less efficient spatial planning and agricultural interventions, particularly in less-developed areas where food and nutrition security remains a critical challenge.

However, the food flow is a complex issue resulted from socio-ecological characteristics of both origin and destination areas, as well as the linkages in between. To untangle this complexity, this research combined concepts of classic gravity model with machine learning techniques, relying on Explainable Artificial Intelligence techniques (xAI) to enhance the transparency of the predictive models. The Irish potato was chosen as the focus crop to study its flow distribution among 30 districts of Rwanda. Objectives included compiling a comprehensive database of socio-economic and environmental factors along with district pair food flows, and leveraging machine learning methods to predict whether a particular district pair presents Irish potato food flows or not. Specifically, Random Forest (RF) and Support Vector Machine (SVM) were trained to predict Irish potato food flows, while the Local Interpretable Model-agnostic Explanations (LIME) xAI technique was used to further investigate particular district pair instance prediction and its most influencing features.

Both RF and SVM models demonstrated high overall accuracy (both above 90%) in predicting district level Irish potato flow. However, it is important to note that the dataset presented imbalanced classes where district pairs that contained Irish potato flows were about 7% of the total data samples, while the remaining dataset comprised the absence of Irish potato flow. F1 score, which is the harmonic mean of precision and recall, was used to evaluate the class prediction accuracy of the models. On the both RF and SVM models, F1 score of class 0 (absence of flow) was 0.96 whereas on class 1 (Presence of flow) was 0.61 for RF and 059 for SVM. These F1 scores shows that both models were accurate at predicting the absence of Irish potato flows (class 0) than the presence of flow (class1), reflecting an imbalance in the dataset where instances of Irish potato flows were less frequent. Using Local Interpretable Model-agnostic Explanations (LIME) xAI technique it was observed that environmental factors notably at the origin district were the most flow influencers compared to socio-economic features.
The study recommends the integration of market level flow data, the scope and temporal expansion for a more granular analysis.

Keywords: Food security; Food flows; Machine Learning; gravity model; Explainable Artificial Intelligence techniques(xAI); Local interpretable Model-agnostic Explanations(LIME)

ACKNOWLEDGEMENTS

I want to express my appreciation to my thesis supervisors: Dr Yue Dou, Dr Claudia Paris, for their advice and mentorship during this research. Thank you for the continued guidance and support. I am appreciative of the endless opportunities you've given me. Thank you for the insights full comments and most importantly, your will to teach and share knowledge.

Extended appreciation goes to the IITA (International Institute of Tropical Agriculture) organisation for providing us the used data and facilitating the fieldwork. Your contributions to agriculture related research is indeed noticed. Special thanks also goes to "ITC BlueSky" for facilitating this research and providing valuable insights and contributions.

I would like to express my gratitude to Maria del Carmen, thank you for ridding the GFM journey with me, an sticking to our goals of ending this research on time. I also extend my gratitude to all my fellow GFM students, You made the journey worth going through. To the food flow partners, Marco ad Manuka, thank you for sharing your insights and providing feedback and support; this research wouldn't have been the same without you.

Special thanks to the GFM specialization teachers, of course we developed a thick skin and got through this research journey. You helped us develop a sense of independence with our work especially when introducing us to the Geocomputing course. This forever changed our perspective on learning.

To my family, my lovely father, Munyangaju Aloys, and my supporting mother, Uwambaye Odette , this educational millstone achievement is due to the continued love, care, patience, encouragement, support, and believe me; this success is entirely for you, thank you. To my beloved sisters . Thank you for cheering along and to many more achievement. Thank you all

Finally, I want to express my gratitude to the Student Association Board (SAB) now Orbit at ITC; It was an honour to lead this association. To the Faculty of ITC, Thank you for providing a learning space for growth and skill sharpening.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML: Machine Learning

xAI: Explainable Artificial Intelligence

FAO: Food and Agriculture Organization

RF: Random Forest

SVM: Support Vector Machine

UN: United Nations

WFP: World Food Program

NISR: National Institute of Statistics of Rwanda

IITA: International Institute of Tropical Agriculture

WFP: World Food Program

OSM: Open Street Map

LIME: Local interpretable Model-agnostic Explanations

1. INTRODUCTION

1.1. Background

Ensuring accessibility to enough, safe, healthy and nutritious food is a challenge that requires a holistic approach (Bala, 2023). Food security is defined by the United Nations as the availability, accessibility, and consumption of sufficient, healthy, and nutritious food (Sage, 2019). Achieving food security requires addressing four pillars: availability, access, usage, and stability.
(Capone et al., 2014). Urbanization is closely related to food security and nutrition, as cities grow and populations concentrate in urban areas; ensuring that food reaches these densely populated regions efficiently and sustainably becomes a challenging problem (Zou et al., 2023). The UN projections indicate that by 2030, over 4.9 billion people accounting for sixty percent of the world's population, will be living in urban areas. This rapid growth in urban populations intensifies the need for food, highlighting the importance of developing resilient food distribution systems capable of meeting the heightened demand.

The production of food is undoubtedly vital; it serves as a bedrock on which food security is built (Swaminathan & Bhavani, 2013). Producing more food would directly address the need to feed the growing population as it would ensure that the escalating demands of people are met. However, growing sufficient and safe food to meet the global demand is only one part of the solution. Equally important are efficient and resilient food distribution systems (James & Friel, 2015). These systems allows for the flow of food from high-producing regions to serve the low-producing areas, thereby balancing the availability of food resources. This flow also facilitates the movement of essential nutrients, ensuring that people from different areas have access to a varied and balance diet.

Food flow describes the movement of agricultural products from their origin of production to their final consumption location, covering the movement from urban and rural areas as well as across different regions (D. Zhou et al., 2012). This process involves a wide range of activities including food production, processing, distribution and retailing (Condratchi, 2014). These are interconnected steps that allow the delivery of food to consumers across different places, transcending the national or regional borders. Even though this process has enhanced connectivity, efficiency and accessibility , it has also introduced vulnerabilities, as any disruption at any point in the process can result in widespread consequences across the entire system (Puma et al., 2015).

Disruptions like pandemics, wars, or even natural disasters can considerably disturb the flow of food (Mishra et al., 2021). The covid-19 pandemic served as an example and revealed the vulnerabilities of global food systems, resulting to disruptions in supply chains and food availability affecting more long distance supply chains and causing losses to farmers and traders(Benton, 2020). For instance, global food trade dropped by roughly 3.5 percent in 2020, with some regions observing more severe disruptions (Engemann & Jafari, 2022). The pandemic lockdowns and restrictions resulted in labour shortages in major sectors like agriculture, transportation, and retail. Delivering perishable food items also became challenging, escalating food shortages. According to the World Food Programme (WFP), the number of people that experienced heightened food insecurity rose by 82 percent, from 149 million in 2019 to 270 million in 2020 (FSIN, n.d.). Moreover, the pandemic highlighted the reliance of urban areas on complex global supply systems which are vulnerable to shocks. Farmers had difficulty accessing markets and supplies, while consumers encountered scarcity and higher prices for basic food items, with global food prices reaching a three-year high in 2020 (Engemann & Jafari, 2022).

A variety of factors can govern the food flow patterns, which are environmental and socio-economic. Where and how food cultivated is determined by environmental factors like climate, water availability, soil fertility and agricultural practices and so forth (Bricas et al., 2019). For instance, regions with favourable weather patterns and fertile lands often become major agricultural hubs, producing excess food that can serve regions that rely on food imports as local production cannot meet demand. Additionally, social economic factors like population density, income levels, education, cultural traditions all influence the type and quantities of food consumed in a certain area as well as imports and exports (Mor & Sethia, 2014). In this context understanding why and how food moves through space and factors driving the flow is essential in achieving food security and availability. This also helps make progress as well as advance achievements of Sustainable Development Goal 2, which aims to end hunger, improve nutrition, and promote sustainable agriculture.

Prior research, particularly on urban food flow, has mainly concentrated on the movement of food within a city or between a city and its surrounding rural areas, often on a national or international level (Schreiber et al., 2021). These studies tend to utilize broader data sets that are readily available, which may overlook the finer details of local food flows. This challenge was addressed in a research by (Mkondiwa & Apland, 2022) where they developed a spatially detailed mathematical programming model to fine-tune district food flows in Malawi. The researchers pointed out that data on food flows between districts is often not collected, presenting a major obstacle in understanding and analyzing food flows within the region. A study examining actual food distribution in China specifically within the confines of Tianjin's administrative boundaries, fails to fully capture the complexities of food flows among the city's districts, as noted by (D. Zhou et al., 2012). Conversely, research by Yang et al. (2021) highlights the difficulties in accurately simulating actual rice flows, citing factors like individual taste preferences and policy regulations that are hard to quantify in simulation models. They propose employing market samples as a more practical approach to represent authentic rice flows. Lastly, Karg et al., (2023) provide an extensive dataset on food flows in four west African cities, covering multiple season and transportation modes but fails to cover intra-city distribution patterns. This reflects a tendency to prioritize food flow research between countries or big cities at the global level whereas internal flows within a country are often neglected potentially due to the challenges in collecting and analyzing data at a more localized level.

1.2. Problem Statement

Mapping food flows from production areas to consumption areas is essential and often challenging, especially at local scales (Moschitz & Frick, 2021). Knowing how food moves over space and time is crucial for policy-making to maintain food and nutrition security across scales. Nevertheless, there is a tendency to prioritize flows between countries at the global level i.e., Food and Agriculture Organization trade data (FAO, 2023a) whereas the internal food flows within a country is often neglected. This oversight can lead to less efficient spatial planning and agricultural interventions, particularly in less-developed areas where food and nutrition security remains a critical challenge. There is an urgent necessity to better understand how food is moving from production to consumption sites. Through food distribution systems lens this linkage can be comprehended by focusing on midstream system segments such as district markets. These markets aggregate small volumes of food from various locations and supply the large urban population, influenced by various environmental and socio economic factors. Few studies report on how food is moved within a country, and often relies on downscaled national or regional data. Since collecting ground truth data is expensive and time consuming, using open access geospatial data, and socio-economic factors coupled with machine learning methods presents potential in giving detailed insights into the most governing factors affecting food flows at the local level.
Machine learning models are often criticized for being black boxes, often lacking transparency in their internal mechanisms and decision-making process (Carrillo et al., 2021). Explainable Artificial Intelligence techniques (xAI) help address this issue by shedding light on why models make certain decisions. In food flow context, xAI transparency can help explain global and local importance of flow driving factors helping in the representation how people interact with the environment in

terms of food production, transportation, and consumption, thus enhancing our understanding of food flows and informing policy-making in the face of increasing natural and social shocks.

This research seeks to integrate machine learning methods coupled with the gravity model to predict district level food flows, using driving factors like environmental and socio-economic. xAI techniques will also be used to explain predictions made by the models. The aim is to better understand key factors influencing food flows between districts within a country. The expected results have the potential to revolutionize the study of food flows, ultimately assisting policymakers and researchers in fostering resilient agrifood systems, contributing significantly to sustainable food security.

1.3. Objectives, research questions and hypotheses

1.1.1. Main Objective

This research seeks to predict district level food flows by constructing machine-learning models with spatial-explicit social-economic and environmental features, based on ground-truth data collected at markets and districts across Rwanda. Using the gravity model as the modeling framework, these models are enhanced with Explainable Artificial Intelligence techniques(xAI) to effectively identify and explain the most influencing factors driving food flows.

1.1.2. Specific Objectives

1. To identify and integrate potential socio economic and environmental features that may influence food flows using multiple data sources.
2. To predict district-level food flows using identified socio-economic and environmental features by leveraging machine learning models.
3. To identify the most influencing features of food flows and explain the model's predictive decisions using Explainable Artificial intelligence techniques (xAI) for greater transparency.

1.1.3. Research questions and hypothesis

Questions1: What is the spatial pattern of the district-level food flow? Which socio-economic and environmental features can be used to describe the production, transportation, and consumption that shape and influence district level food flows?

Hypothesis: Related features collectively shape and determine food flow patterns.

Question 2: Which socio-economic and environmental features that can be used to accurately predict district level food flows using machine learning models?

Hypothesis: Food flow influencing features have different weights of contribution in district level food flow predictions.

Question 3: What are the most influencing features to the food flow, and how can explainable AI (xAI) techniques be used to enhance transparency in the Machine Learning model's predictive decisions?

Hypothesis: Environmental features that determine food production and movement are most influential factors in food flows.

2. METHODOLOGY

2.1. Study Area and Dataset

This research focuses on Rwanda, a small landlocked country located in East Africa with a population exceeding 13 million, a land area of approximately 26,338 square kilometres. Rwanda is geographically located between latitude of 1° 2' 50.82", - 2° 50' 21.3792" and longitude of 28° 51' 42.1488", 30° 53' 45.456". It is composed of 30 districts and 5 provinces and mostly characterized by hilly and mountainous landscapes. Agriculture has a vital role in the nation's economy, accounting for 31% of its overall GDP. This industry engages more than 70% of the workforce and is typically marked by limited productivity. Within this sector, nearly 90% of the families involved engage in traditional subsistence farming methods. (Republic of Rwanda, 2018).

Roughly 51% of the country's land area is designated for agriculture, with about three-quarters of this land dedicated to cultivating diverse crops including food crops, cash crops, and forages. The Eastern Province boasts the largest area of agricultural land, spanning 439,000 hectares, whereas the Northern Province has the smallest at approximately 212,000 hectares (Giertz et al., 2015). Due to the hilly nature of Rwanda's landscape (see Figure 1), 70% of its land is located on hillsides Conversely, just 30% of agricultural areas are found on flat land. This topography poses challenges, as the hilly regions are vulnerable to drought, soil erosion, and landslides, while the marshlands are at risk of flooding during periods of intense rain. (The World Bank Group, 2021). Rwanda is also among the most densely populated countries in the world, with 535 people per square kilometer, the country has a population growth rate of approximately 2.3% as of the 2022 recent census (National Institute of Statistics of Rwanda (NISR), 2023). This increases the necessity for food security and adds weight on the country's agricultural sector requiring efficient food flows and sustainable agricultural practices. Figure 1 illustrates the considered study area, false color representation of the Digital Elevation Terrain Model (DTM).

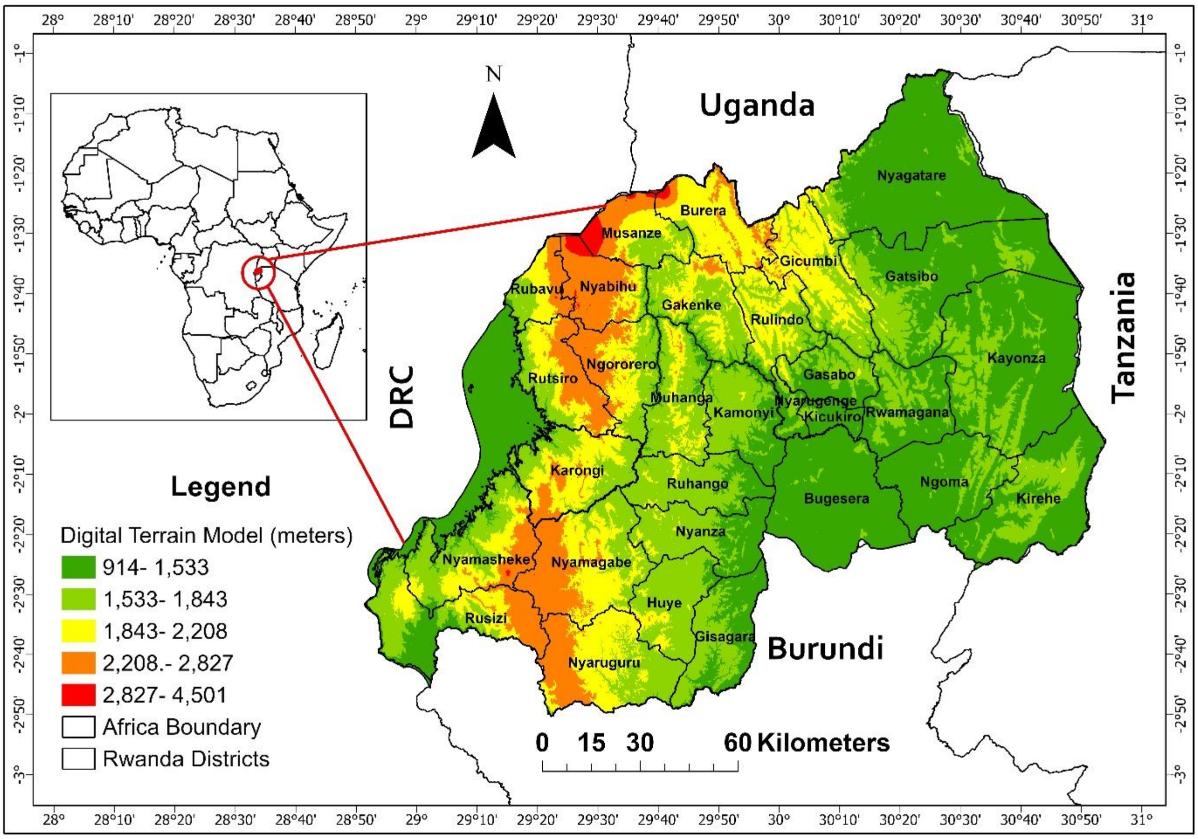


Figure 1 : Considered study area, false colour representation of the Digital Elevation Terrain Model (DTM)

2.1.1. Dataset

❖ **Field market data**

To carry out this research a comprehensive dataset that includes ground truth data was essential. The ground truth primary data was acquired through a research collaboration with International Institute of Tropical Agriculture (IITA) whose vision is to facilitate agricultural solutions to overcome hunger and poverty in the tropics. This dataset was part of an ongoing one year data collection (April 2023-April 2024). The dataset encompasses vendors market survey and market characterization data. This dataset contains a wide range of variables pertinent to the study. to highlight a few details, various vendors in different markets across the country were interviewed about the crops that they usually sell, their location was recorded as well as the production source district. Used transportation means of sold food items was also recorded , market days and the food source channel. Market characteristics were recorded as part of the survey. Figure 2 shows the locations of the surveyed markets across the country.

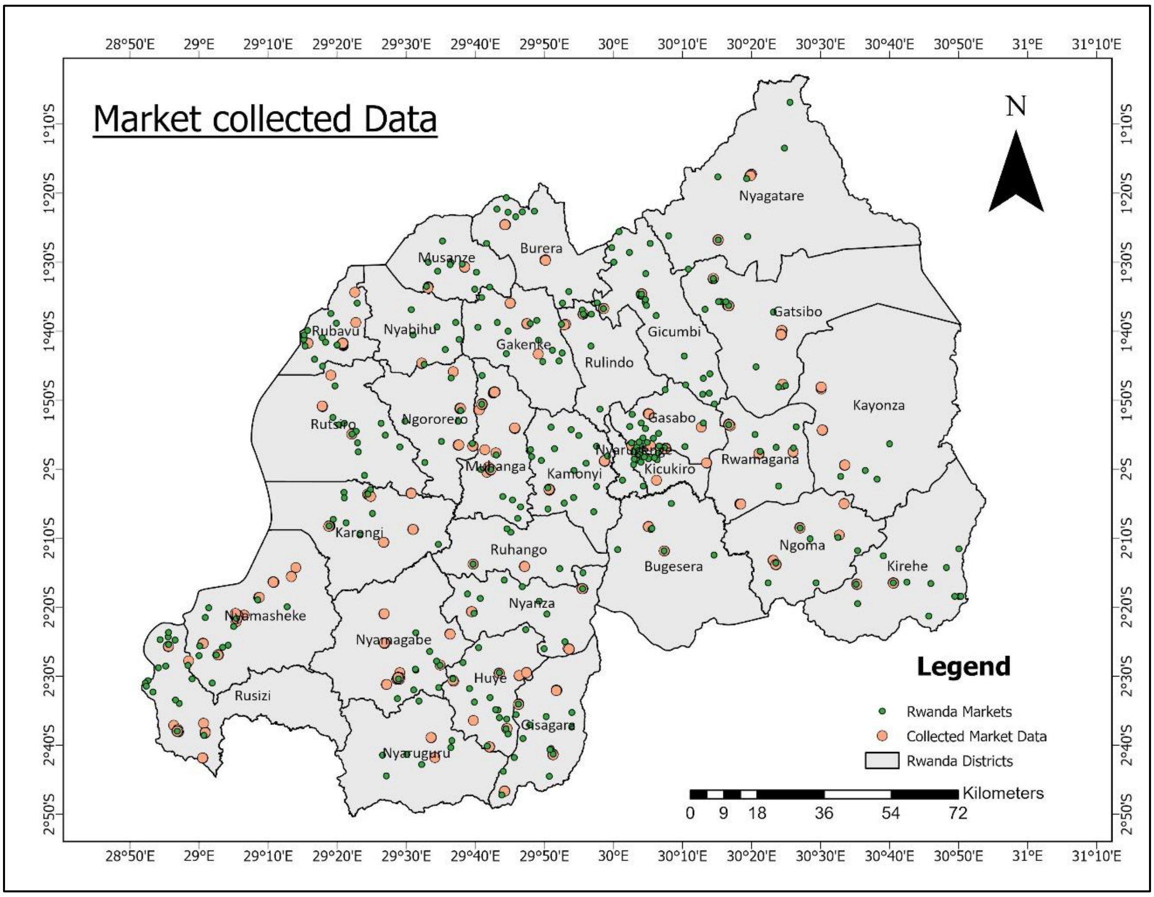


Figure 2: Spatial distribution of collected market data across the study area

In the acquired dataset, only food item source district is recorded not the specific production location. Additionally, the acquired data showed that 60% of vendors reported wholesalers/ aggregators as their food source channels which doesn't directly link food flow to the source production area. Since only districts are recorded as producing locations, ensuring that the collected information is robust and reliable a qualitative market data survey which includes a survey of open-ended questions and interviews to confirm observations of the overall market survey was conducted in February to March 2024. One specific crop was selected by considering the local agriculture season at that period and Irish potatoes were of interest to this study.
Irish potatoes are a significant part of Rwanda's agricultural output and are a staple food in the Rwandan diet. Falling under the government six priority crops within the crop intensification program (CIP). Rwanda ranks as the 6th largest potato producer in Africa, which is of interest

considering the country's relative land size (FAO, 2023). Potatoes grow well in several parts of country mainly above elevations of 1800 m. Four districts in the north-west (Rubavu, Musanze, Nyabihu and Burera) are responsible for most of the production. Currently, Rwanda has over 70,000 potato farmers grouped in 30 cooperatives that produce over 19, 000 tones monthly during the harvesting season.

To link production to consumption, the potato flow channel was investigated during the qualitative data survey. Different market vendors were surveyed as well as wholesalers/aggregators. Reported source district from the initial dataset was confirmed by the surveyed wholesalers/aggregators. Figure 3 depicts the field observed potatoes flow channel from production areas to the last retailer and consumer. Raw images illustrating the process are also attached in the appendix.

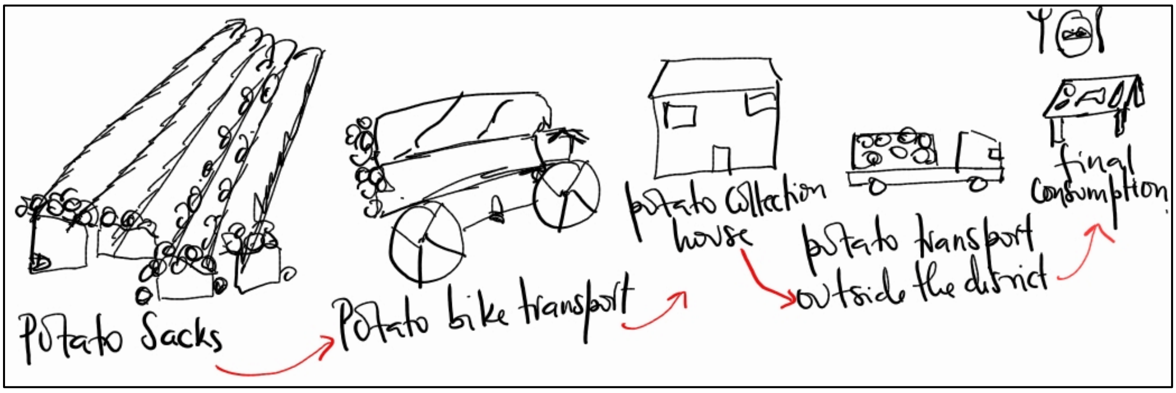


Figure 3: Potato flow channel

Potatoes are harvested from the field and are directly packed into big sacks of about 100kg. They are then transported by bicycles to the nearby collection centre house. These potato collection centres are established by the government and serve as important points where farmers can deliver their harvests for aggregation and distribution to larger markets. From these collection centres, wholesalers/aggregators can come to collect the harvests for distribution across the country. This flow depicts how potatoes get to the final consumer from the farm field.

❖ Environmental data

A 2022 land use/land cover (LULC) map was accessed from ESA Sentinel-2 imagery of 10m resolution. This data is generated with the impact Observatory's deep learning AI land classification model and is produced by Impact Observatory, Microsoft, and Esri. The map was downloaded through the following link: Esri | Sentinel-2 Land Cover Explorer (arcgis.com). Seven landcover classes were selected based on characteristics of the study. Bareground, Built Are, Rangeland, crop, water, trees and flooded vegetation were of importance.

Topographical features used in this study such as elevation was freely available and accessed. A 10m resolution Digital Terrain Model (DTM) prepared by Rwanda's National Land Authority (NLA) was downloaded through the United Nations Development Programme's GeoHub. Minimum and Maximum elevation per district was extracted to be used in the analysis. Land use patterns data like agricultural land area and crop type yield data were available at the district level from the recent seasonal agricultural survey 2023 and can be accessed through The National Institute of Statistics of Rwanda via the link:
https://www.statistics.gov.rw/publication/2022

Climate data and weather data like annual temperatures and precipitation, were used in this study. A dataset was downloaded from "World Clim", with a spatial resolution of 30seconds (~1km^2). Yearly mean temperature and precipitation was calculated on 12 GeoTiff files of representing a whole year. https://www.worldclim.org/data/worldclim21.html

Open Street Map (OSM) data was used to extract points of interest (POIs) relevant to our study area using the "QuickOSM" QGIS plugin. The POIs, which include restaurants, distinctive buildings, churches, and touristic attractions were selected because they serve as proxies for physical landmarks that provide insights into the characteristics of each district. Using QuickOSM, a specific query was formulated to target these categories, relevant POI data was downloaded. A spatial join with district boundary shapefile was done to ensure accurate assignment of POIs. This process allowed for a clear representation of the distribution and density of POIs within each district.

❖ Socio-Economic Data

Population demographics like population density, district population count and gender distribution percentage was used in this study. This data was extracted from the recent 2022 Population and Housing Census provide by the national institute of statistics of Rwanda and was accessed through the link: https://www.statistics.gov.rw/publication/main_indicators_2022. These demographic factors were selected because they provide critical insights into social structure and resource needs of each district which directly influence food flows. Population density offers a measure of how crowded or sparse a district is, affecting demand for food and the efficiency of food distribution systems. The district population count helps in understanding the overall size and scale of each district. Gender distribution percentages was also selected due to the impact it may have on food consumption patterns and nutritional needs.

Infrastructures like roads, acquired from the Ministry of Infrastructure (MININFRA). For Roads, only “road surface length” attribute which specified whether a road is paved or not paved was chosen. This was because this attribute provided more information than the road category attribute which included details like primary, secondary and tertiary road. Road data was downloaded through the following link : https://datacatalog.worldbank.org/search/dataset/0040262/Rwanda-Roads. District school and health centres count was used and was used to approximate the level of education and availability of health services in a district. Market infrastructure was also acquired from the vendor survey data.

Economic indicators like the Gross Domestic Product, can be found through recent GDP national accounts report of the first, second, third and fourth quarter of 2023. The report is published by the national institute of statistics of Rwanda and can be accessed through this link: https://www.statistics.gov.rw/publication/2016. The GDP per district was extracted to be used in this study. Table 1, below summarizes the acquired and used data.

Table 1: Data Description

Data	Data description	Data owner
DTM, 10 m resolution	Maximum elevation	Rwanda National Land Authority (NLA)
DTM, 10 m resolution	Minimum elevation	Rwanda National Land Authority (NLA)
LULC, 10m resolution	Bareground	Sentinel-2 Land Use/Landcover_Esri
LULC, 10m resolution	Built Area	Sentinel-2 Land Use/Landcover_Esri
LULC, 10m resolution	Rangeland	Sentinel-2 Land Use/Landcover_Esri
LULC, 10m resolution	Crop	Sentinel-2 Land Use/Landcover_Esri
LULC, 10m resolution	Trees	Sentinel-2 Land Use/Landcover_Esri
LULC, 10m resolution	Water,	Sentinel-2 Land Use/Landcover_Esri
LULC, 10m resolution	Flooded vegetation	Sentinel-2 Land Use/Landcover_Esri
Agri-survey (2023)	Potato Yield quantity	National Institute of Statistics of Rwanda(NISR)
Agri-survey (2023)	Agricultural land area	National Institute of Statistics of Rwanda(NISR)
Climate data, 30seconds (~1km2).	Max & Min temperature	WorldClim
Climate data, 30seconds (~1km2).	Average precipitation	WorldClim
Open Street Map	Points of interest	OSM
Administrative boundary	District pair distance	Global Administrative Areas (GADM)
Administrative boundary	District boundaries	Global Administrative Areas (GADM)
Population demographics (2022)	Population count	National Institute of Statistics of Rwanda(NISR)
Population demographics (2022)	Population density	National Institute of Statistics of Rwanda(NISR)
Gross domestic product (2023)	District GDP	National Institute of Statistics of Rwanda(NISR)
Public infrastructure	Roads	Ministry of Infrastructure (MININFRA)
Public infrastructure	District school count	Ministry of Infrastructure (MININFRA)
Public infrastructure	District heath centres count	Ministry of Infrastructure (MININFRA)
Market characteristics (2023)	Market type	International Institute of Tropical Agriculture (IITA)
Market characteristics (2023)	Market infrastructure	International Institute of Tropical Agriculture (IITA)

2.2. Methodological Flow Chart

To predict district level food flows, various steps were undertaken, figure 4 shows the flow chart of undertaken steps. The major steps include data preprocessing that lead to a database compilation of all flow data. District level food flows were classified and the accuracy assessed using different models. The Local interpretable Model-agnostic Explanations (LIME) xAI technique was also applied to gain insights on the model's output. The chart also illustrates the study's objective achievement along the way.

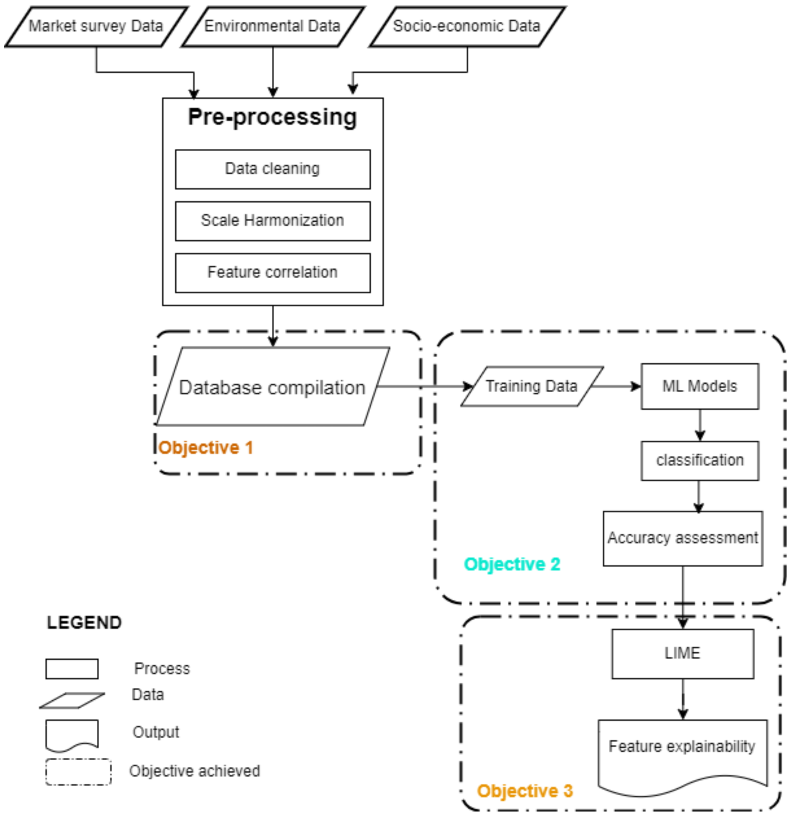


Figure 4: Methodological flow chart

2.1.2. Data pre-processing

District flow data was extracted from the acquired market data by filtering potatoes as food items of interest. Inconsistencies were removed and district flow pairs of origin and destination were made. All feature data were brought to the district level to match the flow data extracted from the vendor survey. Percentage aggregation and feature categorization was done to achieve feature representation. Additionally a correlation matrix was used to check and eliminate highly correlated features before the data analysis. In the correlation matrix, it was observed that the District GDP and the population count were highly correlated. The Female and male percentage at the district level was used to represent the population count. Figure 5 depicts a correlation matrix of the used features. Additionally, district centroids were used to calculate the inter district distance. This choice was motivated by the need to simply and standardize the travel distance.

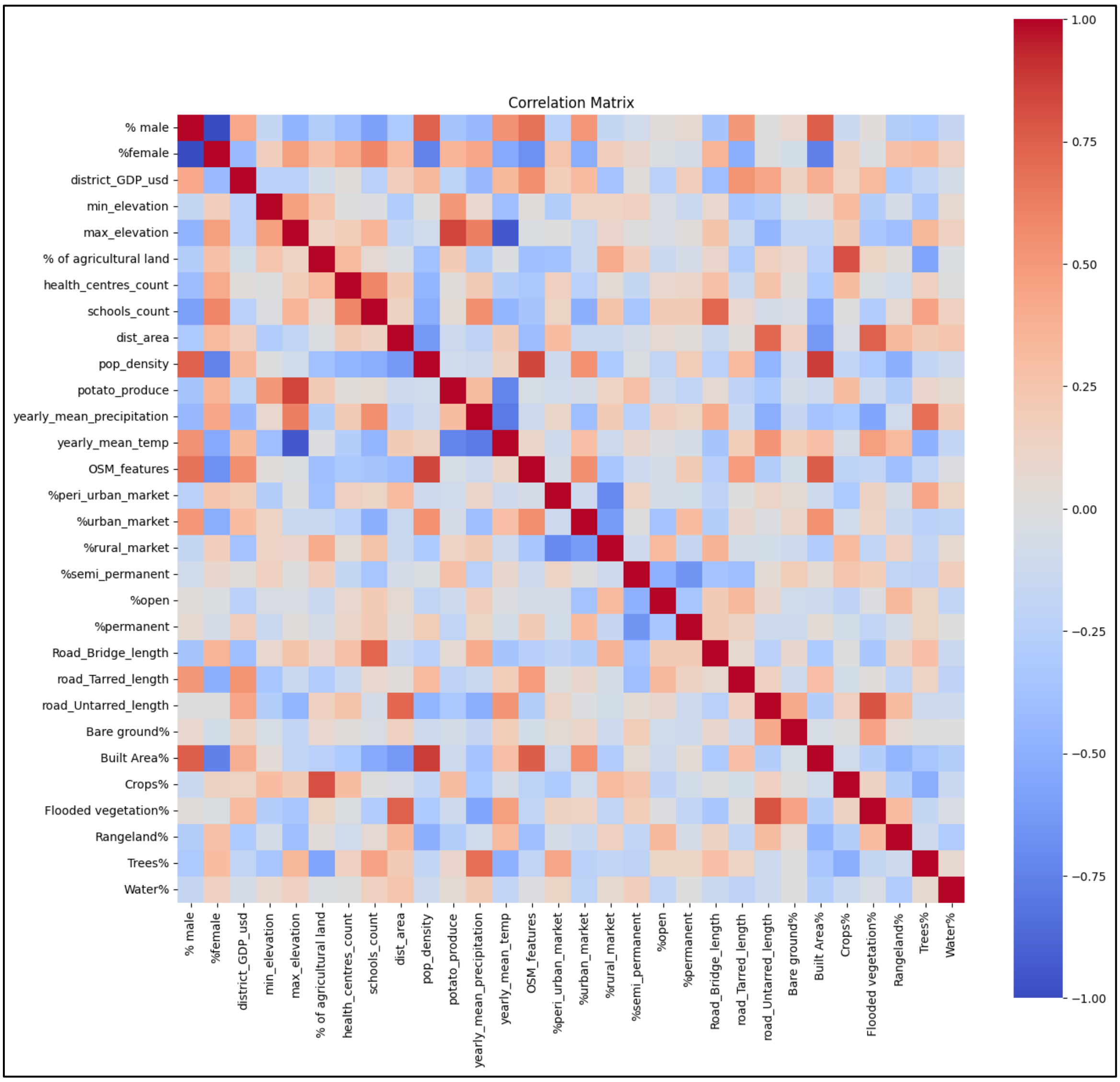


Figure 5: Features correlation matrix [(click here for an extended view)]

To understand the pattern at district level, histograms were also utilized to compare different features. Significant differences in GDP across districts were observed with Gasabo and Kicukiro exhibiting higher economic activities. Agricultural land percentage also varies notably with Gakenke and Nyagatare districts heavily relying on agriculture. While Kicukiro is more urbanized, Gakenke and Muhanga have more infrastructure like heath centres and schools. Districts like Musanze and Nyabihu have higher elevations, Nyamagabe and Nyamasheke receives more precipitation. Landuse patterns show a balance between urban development and agricultural focus , with some districts having more built-up areas and others more cropland. Figure 6 below illustrates different features comparison across 30 districts of Rwanda.

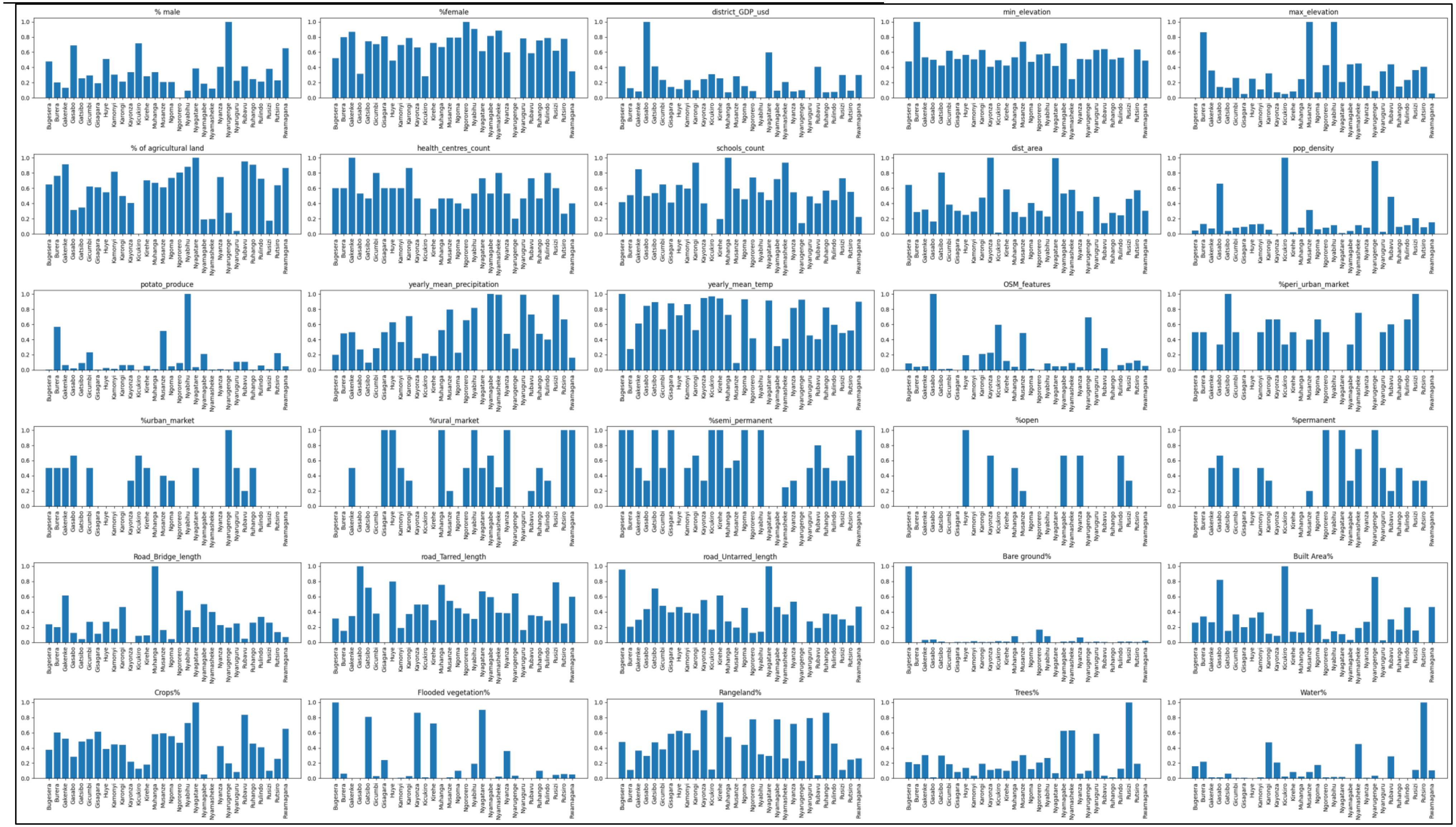


Figure 6: Feature histograms across districts [(Click here for an extended view)]

2.3. Feature Database compilation

To predict district level food flows, compiling a detailed feature database is essential. Three important datasets were of great use to achieve this goal. These datasets are existing district level potato flows extracted from the vendor survey, district centroid distances and the previously mentioned district feature properties. To create potato flow origin and destination pairs, the gravity model framework was utilized. The gravity model is commonly used to assess and predict economic factors, particularly in bilateral trade flows. The model operates on the principal that trade flow magnitude is influenced by supply factors at source and demand factors at destination, as well as dynamics propelling these trade flows (Kabir et al., 2017).

Let us define the origin and destination district of food flows as D_i and D_j, respectively, with $i, j \in [1, N]$, having N district present in the considered study area. To estimate the flow of potatoes moving from D_i to D_j, we extract a set of features which aim to model origin and destination district properties. Let x_i and x_j be the set of compiled features associated with district D_i and D_j, respectively. To estimate the flow from D_i and D_j , we exploit both origin and destination feature vectors. In addition, we include a feature typically used in gravity models, the geographic distance, $d_{i,j}$ between two districts D_i and D_j, computed as distance between district centroids. Accordingly, the feature vector used to estimate the flow between D_i and D_j is $[x_i, x_{j,} d_{i,j}]$. For each origin district D_i we computed a feature vector$[x_i, x_{j,} d_{i,j}]$ for each district D_j that could be a potential destination. i.e. $j \epsilon [1, M]$ having $M \leq N$.

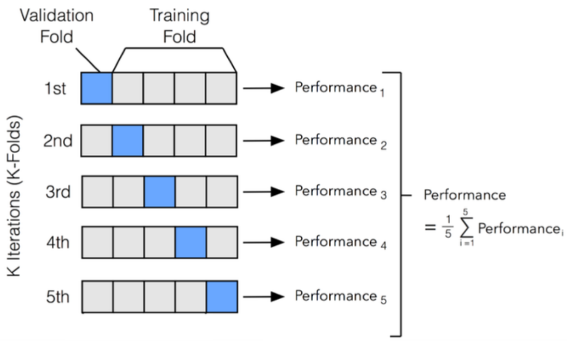
2.4. Machine Learning Models

To provide a comprehensive analysis of district-level food flows in Rwanda, this study is conducted using two machine learning models to predict district level food flows i.e., RF and SVM. This is due to the complementary strengths in handling diverse data characteristics and their proven efficacy in classification tasks (Saini & Ghosh, 2018). By integrating both models, this study aim to leverage their unique advantage to enhance the accuracy and reliability of the flow predictions. This also allows to address potential bias and limitations inherent in any single model, providing a more robust framework for the study.

I. Random Forest (RF) classification

In this study , a RF machine learning classification model is used to predict district level potato flows, outputting binary results (0 and 1) where 1 indicate presence of a flow and 0 its absence. RF is a powerful ensemble learning method that operates by constructing multiple decision trees during the training step and outputting the mode of classes or the mean prediction of individual trees (Kulkarni & Sinha, 2014).

To train the RF model, 870 out of 900 sample points were used. These points were selected after removing pairs where the distance is zero as those represented internal district flows. About 7% of 870 samples points comprised potato flows meaning 93% of all samples represented pairs where there is no potato flow. As this dataset was imbalanced, strategically splitting training samples to avoid feeding bias to our training model was essential. Stratified K-Folds was used to split the data samples as this method deemed best to handle imbalanced classes (López et al., 2014). Stratified K-folds data splitting technique combines two methods. Dividing a dataset into K equally sized subsets (folds) and ensuring that the distribution of classes is preserved across all training and validation sets (Anguita et al., 2012). Figure 8 depicts the operation of k-folds as well as the random forest model. The model is trained and validated 5K times each time using a different fold as the validation set and the remaining k-1 fold as the training set.



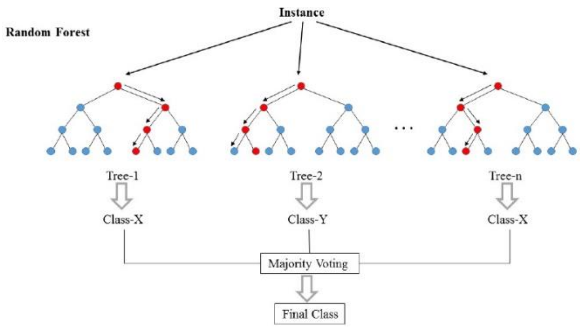


Figure 7 : K-fold splitting illustration

II. Support Vector Machine (SVM) classification

SVM is a powerful supervised machine learning model frequently used for classification tasks. The algorithm works by finding the optimal hyperplane that best separates data points of difference classes. The hyperplane is selected to maximize the margin which is the distance between the hyperplane and the closest data point from each class, also known as support vectors (Mohan et al., 2020). The margin maximization helps to improve the model's generalization capabilities on test data. In this study, SVM was used to predict potato flows due to its effectiveness in handling complex data where the relationship between features is not straight forward, which is the case for our dataset. To train the SVM model, 870 sample points were used. About 7% of 870 samples represents sample points that have flows. As this is a small portion considering the whole dataset, Stratified K-Folds was again used as a data splitting technique to handle the class imbalance in our dataset. In this study, a grid search systematic way was used to explore a range of hyperparameter values to find the best combination that can be used for the SVM Model.
Figure 9 below is a visual representation of how the SVM model works, by representing the simple case of linear kernel function. One class is represented by the blue square dots, another class by red circular dots. The back line is the decision boundary learnt by the SVM which is placed in a way that maximizes the margin between two classes.

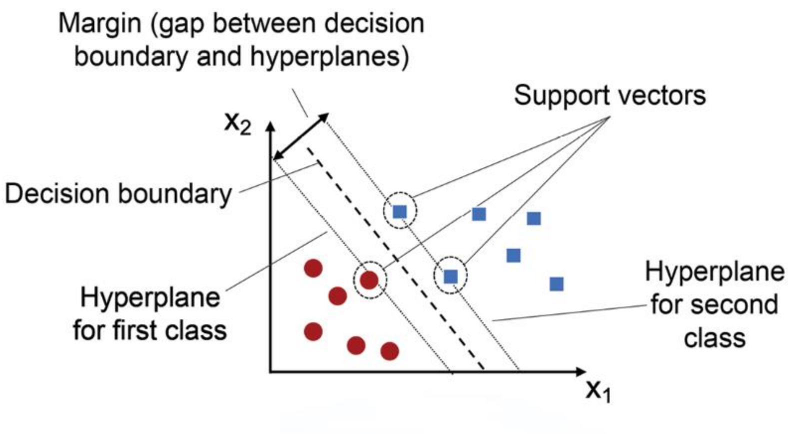


Figure 8: Support Vector Machine illustration

2.5. Explainable Artificial Intelligence (xAI) Techniques

Machine learning models are often criticized for being "black box" models where their decision process is not transparent or easily understood (Räz, 2022). This lack of clarity raises concerns especially where understanding the reasoning behind the model's prediction is crucial. xAI aims to increase machine learning model's transparency and interpretability, providing insights into the model's specific instance decisions, thus enhancing their reliability and acceptance of results (Islam et al., 2022).

xAI techniques comprise of two types, Model-agnostic and Model-specific. Model-agnostic may be utilized on any machine learning model, irrespective of its internal structure while Model specific techniques on the other hand are tailored to particular types of model. For instance, feature importance for decision tree models (Greenwell & Boehmke, 2020). These mentioned xAI types can provide global and local explanations. Global explanations provide insights into the overall behaviour of the model where they help in understanding how the model makes predictions across the entire dataset. Local explanations, on the other hand, focus on individual predictions. These provide insights into why a model made a specific prediction for a particular instance. This is particularly useful for understanding and trusting specific decisions made by the model.

One of the most commonly used model-agnostic techniques is Local Interpretable Model-agnostic Explanations (LIME). LIME technique which is primarily designed to provide local explanations, this can be used to explain individual instance predictions by locally approximating a Machine learning model, making it versatile and widely applicable (Dieber & Kirrane, 2020). In this study understanding and identifying important features that significantly influence food flows is needed hence the usage of LIME. LIME can illustrate important features that contributed significantly to the model's flow prediction, providing a clear understanding of which variables are driving specific predictions. This not only aids in verifying the model's decisions but also enhances trust in the model's outputs by highlighting the rationale behind each prediction.

In this research we structured the illustrated LIME instances by first, providing the global LIME explanations on both RF and SVM models, where we summarized the global importance in a table. Second, instances where the model predicted a flow or no flow were demonstrated and lastly instances where a flow was intended to happen but did not were also illustrated. This was done to identify and capture the different influencing features in all of these instances and also verify the models accuracies.

3. RESULTS

This Section illustrates the obtained results; first an overview of the flow predictions is given through a flow map, then, the RF and SVM flow classification results are presented. To illustrate if and how the models correctly identified the expected potato flows. To provide local explanations, on the models predictions, and to identify important features, LIME results are presented on particular instances, first where the models correctly identified the flows, then instances where the were not correctly identified. Lastly, digging deeper, we illustrate where the model was supposed to identify a flow but it failed to, we also highlight contributing features this this false predictions.

3.1. Flow description

To describe district level food flow patterns, district pairs were made and a flow map was constructed. 900 district pairs were made since our study area Rwanda, comprises 30 districts. Irish Potato was the chosen as the focus crop due to its significant relevance to Rwanda and because it was in season during the data collection period.

Figure 7 illustrates the distribution of Irish potato production across various districts of Rwanda, highlighting the movement from origin (red dots) to destination (green dots) districts. The map also shows that the flow is concentrated in certain regions with a significant movement towards central and northern districts. Musanze and Nyabihu serve as major origin points. Internal flows also suggest local distribution patterns where districts manage their own internal supply and demand. It is important to note that the production quantity is not the flow quantity coming from an origin district to a destination district. The production yield quantity represents the overall potato district production. Table 2 below illustrates the first four origin and destination districts as well as their relevant characteristics that helps in understanding the observed food flow map figure 7.

Table 2: Four Origin and destination districts and their characteristics summary

	district name	Characteristics							
		district GDP USD	min elevation	max elevation	district area(km2)	pop density	potato production quantity (MT)	yearly mean precipitation	yearly mean temperature
Flow Origin Districts	Burera	403238160	1753	4115	645.379	682	117256	102.2	16.8
	Musanze	495582880	1535	4501	527.714	1157	106130	118.8	15.7
	Nyabihu	331808880	1405	4492	537.791	642	205738	119.8	15.1
	Rubavu	568550320	1453	2913	384.136	1614	22311	115.3	17.6
Flow Destination Districts	Gasabo	914685200	1336	2075	429.260	2056	4208	91.2	20.3
	Nyarugenge	389291760	1339	1924	132.462	2830	833	91.6	20.8
	Kicukiro	511400240	1331	1816	167.020	2944	40	88.3	21.0
	Rwamagana	504351120	1327	1826	680.741	740	9511	85.4	20.6

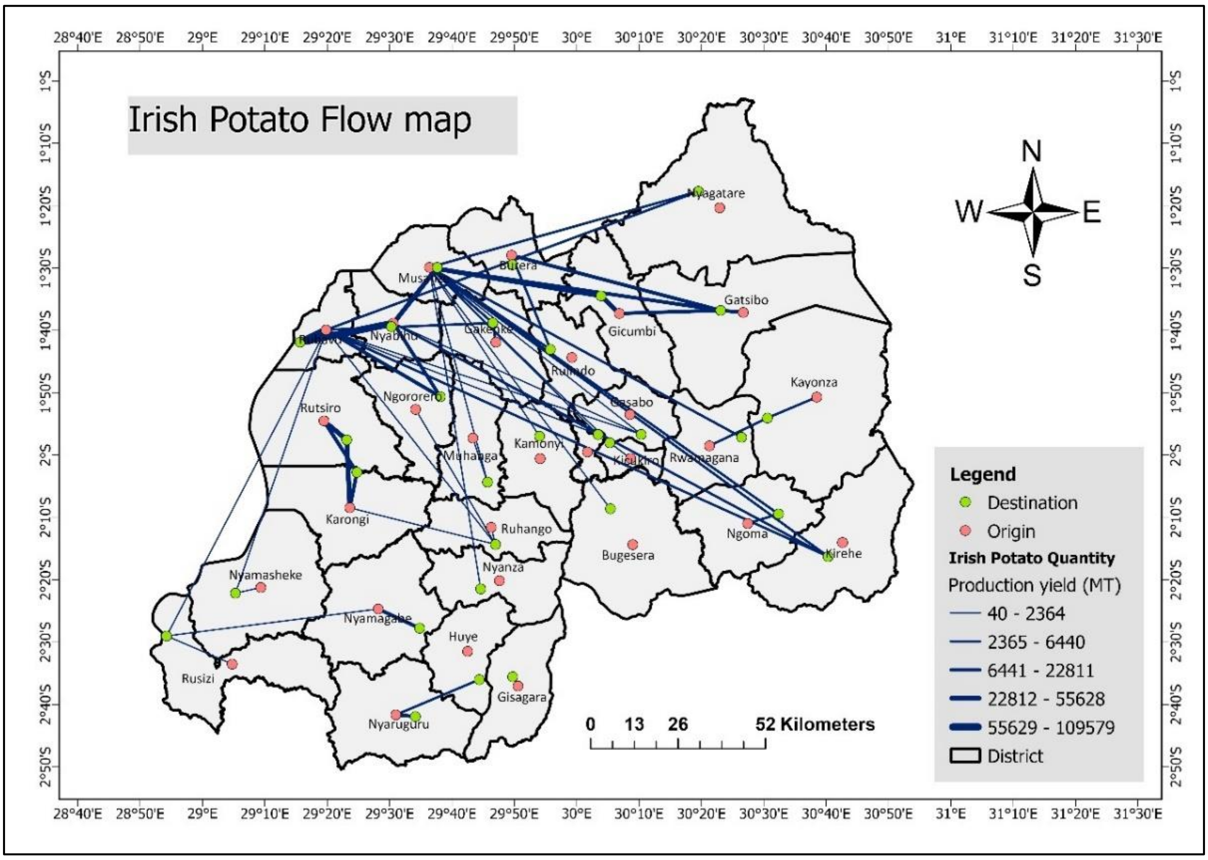


Figure 9: Irish Potato Flow Map illustrating the flow between districts in Rwanda.

Figure 10, also summarizes the district pair distances, this helps understand the spatial connectivity and logistical aspects that may happen during potato flows between districts. The longest distance between district pairs is 197.71km and is between Nyagatare and Rusizi districts. The smallest distance at same district pairs.

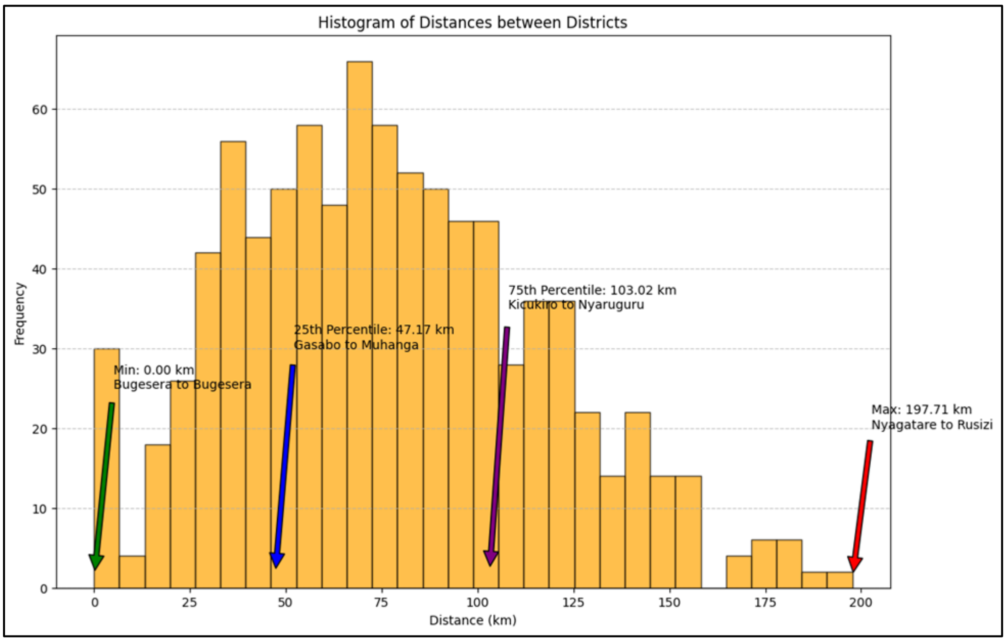
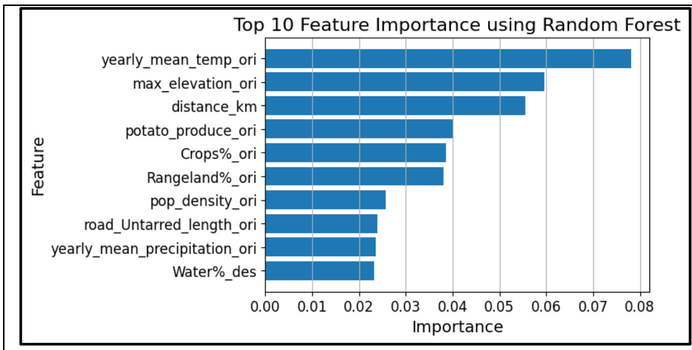


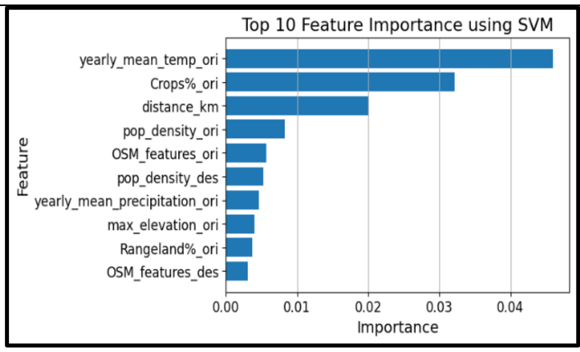
Figure 10: Distance pair quantiles

3.2 Feature importance

Different features contributed globally to the prediction of class 0 and 1 in RF and SVM models. However, some were more important than others, Table 3, illustrate RF and SVM global feature importance of the first 10 features. It is observed that the yearly mean temperature and precipitation, the distance between districts, the population density, and landcover classes area i.e: Crops and Rangeland were more important to the model's predictions. Note that full charts are displayed in the appendix.

Table 3: Feature importance of RF and SVM, top 10 first features





3.2. Random Forest Classification

Results from the RF model provide a comprehensive overview of its classification performance on predicting district level food flows in Rwanda. The training samples were initially split into five stratified folds to avoid class imbalance bias as illustrated in the methodology. To properly train the model, a grid search systematic way was used to explore a range of hyperparameter values to find the best combination. The chosen parameters were 100 trees, the maximum trees depth was set to none, which lets trees expand until all leaves are pure or contain less than min samples split, the samples were bootstrapped as well.

Table 4 Summarizes the classification output of each fold. Each fold's accuracy and the overall RF model accuracy which is the average accuracy across all five folds is presented. F1score of class, 0 (representing the absent of flow) and class 1 (representing the presence of flow) is also produced. The overall accuracy achieved by the model is 94% where F1 score of class 0 (absent of flow) is 0.96 and that of class 1 is 0.61. F1 score is the harmonic mean of the model's precision and recall metrics, it specifically measures the model's accuracy in predicting positive class instances. A higher F1 score indicates better performance in predicting the specified class.

Table 4: *: Random forest classification summary*

Fold	**Accuracy**	**F1 Score Class 0**	**F1 Score Class 1**
Fold 1	0.94	0.97	0.64
Fold 2	0.93	0.96	0.45
Fold 3	0.93	0.96	0.53
Fold 4	0.95	0.97	0.73
Fold 5	0.94	0.97	0.68
Overall	**0.94**	**0.96**	**0.61**

Confusion matrices figures are shown below, Figure 11 , all 5 folds (in blue colour) illustrate class 1(presence of flow) exhibiting less number of true predicted samples compared to class 0 (absence of flow). Overall confusion matrix (in red colour) is also show.

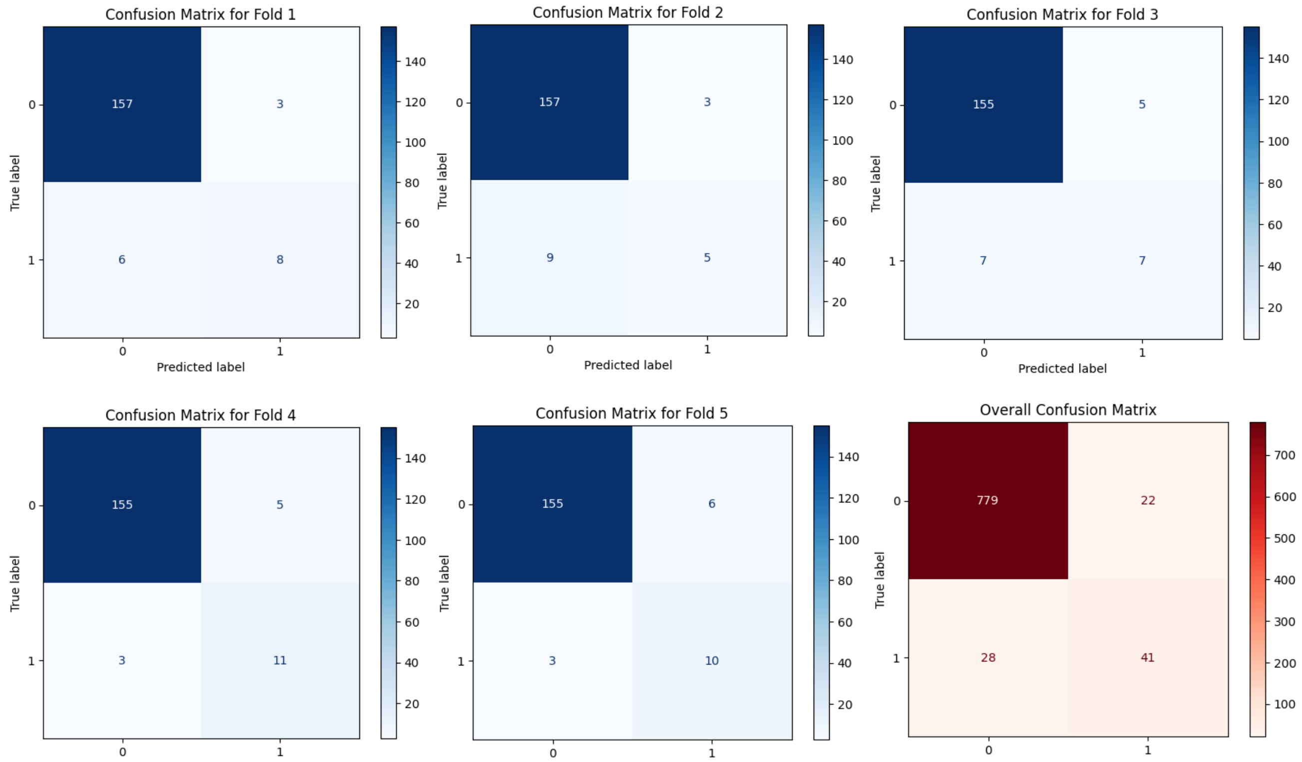


Figure 11 : RF prediction Confusion Matrices *(click here for an extended view)*

3.3. Support vector Machine(SVM) classification

Results from the Support vector Machine (SVM) model also provide a comprehensive overview of its classification performance on predicting district level food flows in Rwanda. The training samples were also split into five stratified folds to avoid class imbalance bias. The sample dataset comprised of 870 district pairs. Only 7% of these samples exhibited potato flows.
To properly train the model, a grid search was used to explore a range of hyperparameter values to find the best combination. The optimal kernel parameters (i.e., the regularization parameter C and the spread of the kernel were selected by a 5-fold cross-validation..

Table 4 Summarizes the classification output of each fold. Each fold's accuracy and the overall SVM model accuracy which is the average accuracy across all five folds is presented. F1score of class, 0 (representing the absent of flow) and class 1 (representing the presence of flow) is also produced. The overall accuracy achieved by the SVM model is 93% which is 1% less that of RF model. F1 score of class 0 (absent of flow) is 0.96 and that of class 1 is 0.59. F1 score is the harmonic mean of the model's precision (the accuracy of positive predictions) and recall (the model's ability to identify all actual positive instances) metrics, it specifically measures the model's accuracy in predicting positive class instances. A higher F1 score indicates better performance in predicting the specified class.

Results from the SVM model also provide a comprehensive overview of its classification performance on predicting district level food flows in Rwanda. The training samples were also split into five stratified folds to avoid class imbalance bias. The sample dataset comprised of 870 district pairs. Only 7% of these samples exhibited potato flows. Table 2 summarizes the classification output of each fold. Each fold's accuracy and the overall SVM model accuracy which is the average accuracy across all five folds is presented. F1score of class, 0 (representing the absent of flow) and class 1 (representing the presence of flow) is also produced. The overall accuracy achieved by the SVM model is 93% which is 1% less that of RF model. F1 score of class 0 (absent of flow) is 0.96 and that of class 1 is 0.59. F1 score is the harmonic mean of the model's precision (the accuracy of positive predictions) and recall (the model's ability to identify all actual positive instances) metrics, it specifically measures the model's accuracy in predicting positive class instances. A higher F1 score indicates better performance in predicting the specified class.

Table 5: SVM Classification Summary

Fold	Accuracy	F1 Score Class 0	F1 Score Class 1
Fold 1	0.94	0.96	0.61
Fold 2	0.92	0.95	0.48
Fold 3	0.93	0.96	0.56
Fold 4	0.94	0.96	0.66
Fold 5	0.93	0.96	0.66
Overall	**0.93**	**0.96**	**0.59**

Folds Confusion matrices figures are shown below, Figure 13 , all 5 folds (in blue colour) illustrate class 1(presence of flow) exhibiting less number of true predicted samples compared to class 0 (absence of flow). Overall confusion matrix (in red colour) is also show below.

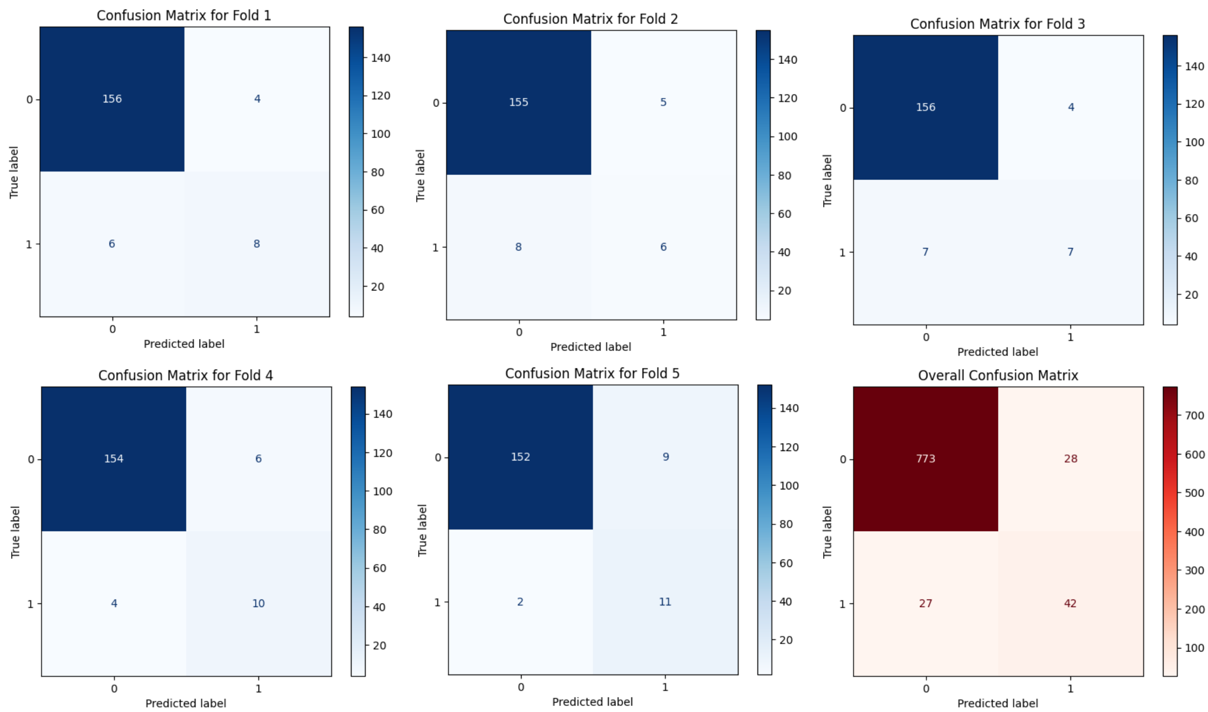


Figure 12: SVM prediction Confusion Matrices [(click here for an extend view)]

3.4. Local interpretable Model-agnostic Explanations (LIME)

LIME as an xAI technique was used to explain important features locally at specific instances of the model's predictions in this study, LIME was used to provide both global and local interpretation of instances predictions for both RF and the SVM models. For the global perspective, Table 3 below summarizes the common most important features identified in both models, RF and SVM using the LIME explanations. Yearly mean temperature, Yearly mean precipitation , Crop LULC and the population density are among the most influencing features. This showed that the environmental characteristics of the different districts played an important role in determining the food flows. Also, the population density was a drawing force of the predictions. Market characteristics i.e.: market categories (rural, urban, periurban) were found to be the least important features as they did not significantly influence flow predictions.

Table 6: Summary of common most important features on selected instances

Models	Influencing Environmental features	Influencing Socio-economic features	Least Influencing features
RF	• Yearly mean temperature • Yearly mean precipitation • Crop LULC • Rangeland LULC • Maximum elevation • Potato production quantity	• Population density	• Market characteristics (Market type & infrastructure)
SVM	• Yearly mean precipitation • Crop LULC • Yearly mean temperature • Distance • OSM features	• Population density	• Market characteristics (Market type & infrastructure)

However, to understand the reasoning behind specific predictions, the local interpretation of specific instances was also carried out. This allows the illustration of certain feature influence on local predictions, helping to pinpoint and label exactly where the flow was occurring or not.

- **LIME on RF and SVM Model: potato flow predictions**

To illustrate further, first, an instance where the model correctly predicted a flow (class1) was selected randomly for a closer observation of the prediction's influencing features. Figure 14 illustrates a LIME interpretation for an instance having potato flow coming from Burera to Bugesera and correctly classified using the RF classifier. The model outputs a prediction probability of 0.21 for Class 0, representing absence of flow and a probability of 0.79 for class 1 representing the presence of flow. The figure also illustrate 10 first prediction contributing features where the first illustrated ones (represented by the orange colour) contributed positively. Yearly mean temperature, the crop landuse landcover district percentage as well as the minimum elevation at the origin district were among the first influencing features to this prediction.
The magnitude of influencing features is also shown, highlighting the weight of each feature to the prediction.

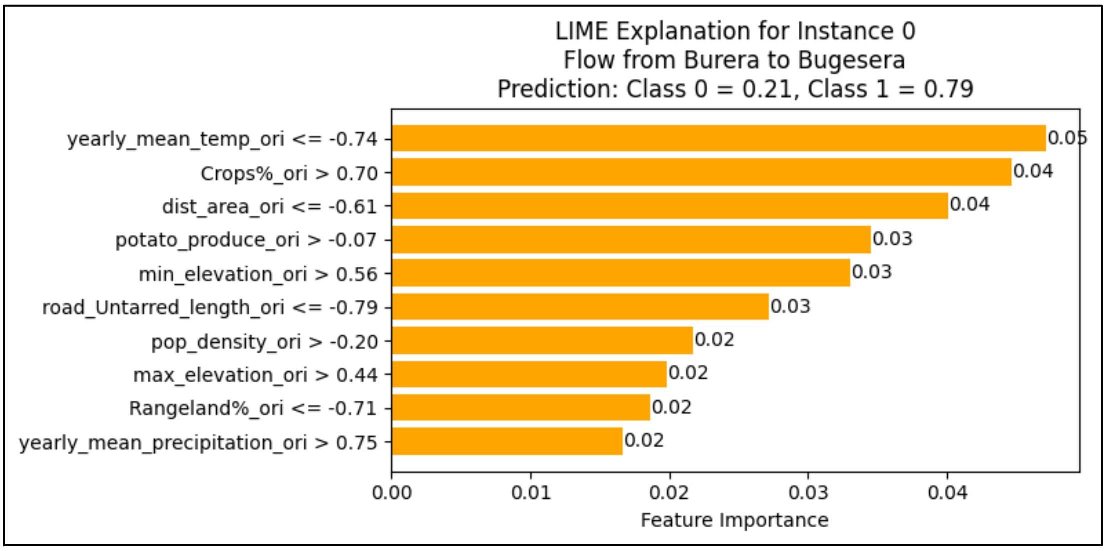


Figure 13: LIME explanation for the RF model for an instance having potato food flows and correctly predicted. The first 10 important features are also reported.

For the same instance, the SVM predicted the presence of a flow with a probability of 0.71. Figure15, illustrate 10 first contributing features where these contributed positively (represented by the orange colour). Yearly mean precipitation, the population density, the crop landuse landcover district percentage as well as the yearly mean temperature at the origin district were among the first influencing features to this SVM prediction. Comparing with RF, two features are common: the yearly mean temperature and the cropland landuse landcover percentage. This illustrated that temperature and Landuse are an important element in agricultural productivity which leads to more food flowing to different districts.

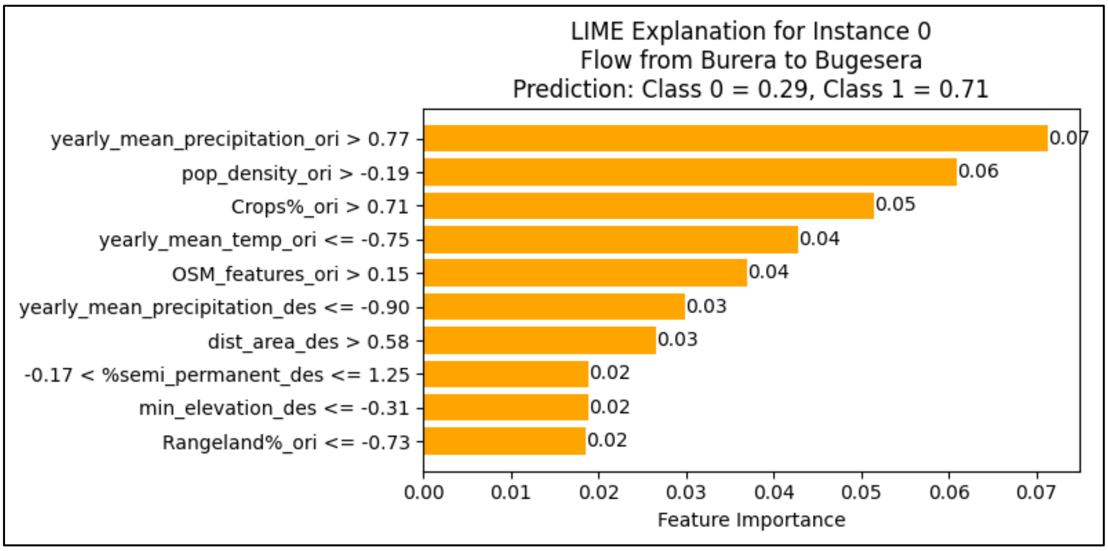


Figure 14: LIME explanation for the SVM model for an instance having potato food flows and correctly predicted. The first 10 important features are also reported.

❖ LIME on RF and SVM Model: absence of potato flow prediction

Figure 16 illustrates a LIME interpretation for a randomly chosen instance which present the absence of potato flow. The flow origin and destination districts at this instance are Rusizi and Gakenke. Here the RF model outputs a prediction probability of 1.00 and 0.00 for class 0 and 1 respectively. This indicates that the model was successfully able to predict class 0 with a maximum certainty. The figure also shows positive (orange colour) and negative (blue colour) contributing features to these predictions. Population density at origin district, yearly mean precipitation and rangeland landuse landcover percentage in a district contributed positively. Features like centroid distance, crops landcover, min elevation etc contributed negatively. Their influencing magnitude is also shown, highlighting the weight of each feature to the prediction.

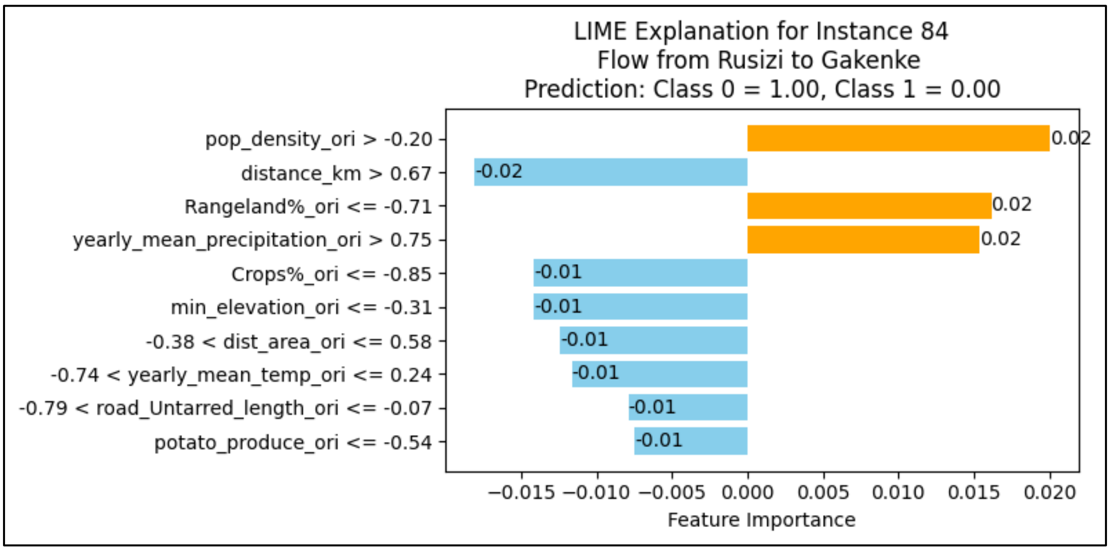


Figure 15: LIME explanation for the RF model for instance which present the absence of potato flow. The first 10 important features are also reported.

Figure 17 illustrates a LIME interpretation for the SVM model considering the same instance. Here the model outputs a prediction probability of 0.92 and 0.08 for class 0 and 1 respectively. This indicates that the model was successfully able to predict class0. The figure also shows positive(orange colour) and negative(blue colour) contributing features to these predictions. Population density at origin district, the distance between districts and OSM features contributed positively while features like yearly mean precipitation, crops landcover, yearly mean temperature and so forth contributed negatively. Their influencing magnitude is also shown, highlighting the weight of each feature to the prediction. Comparing features that contributed positively for both RF and SVM; the population density, Rangeland landuse landcover area were more important to the predictions. On the other hand, the district area, the yearly mean temperature and the crops landuse landcover had a negative contribution the overall flow classes predictions.

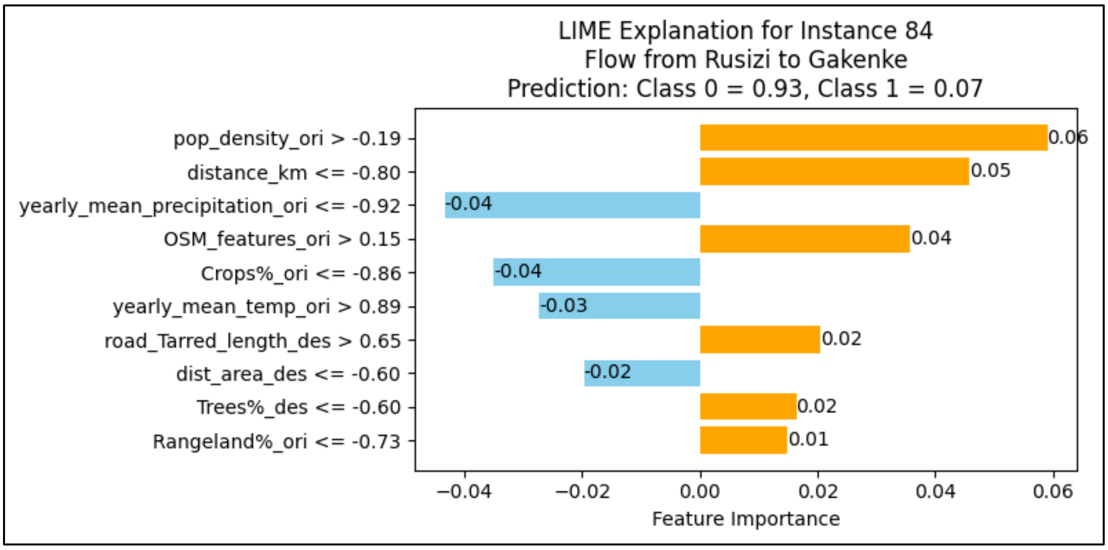


Figure 16: LIME explanation for the SVM model for instance which present the absence of potato flow. The first 10 important features are also reported.

I. Instances where a flow was expected but the model didn't spot it

Results also showed that there were cases where a flow was expected at some instances but the models could not spot it. Figure 18 shows an instance where a flow was expected and RF correctly spotted the flow with 0.12 and 0.88 for class 0 and class1 respectively. However, SVM failed to identify this flow (see Figure 19). Among the positively contributing features for the RF model were yearly mean temperature, the crop landuse landcover, the minimum elevation. For the SVM model features like the yearly mean precipitation, the district area, the minimum elevation contributed positively while features like the population density contributed negatively to this flow.

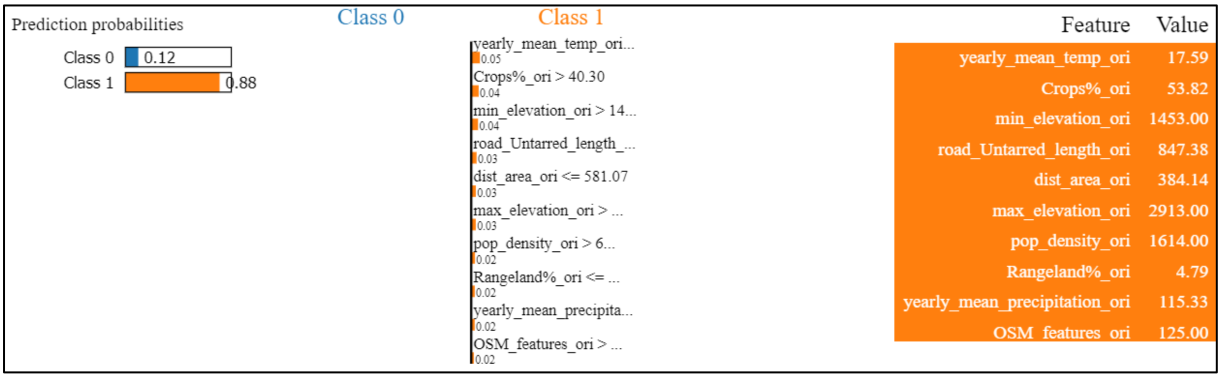


Figure 17: LIME explanation for the RF model, which successfully predicted the food flow for the 110^th^ instance.

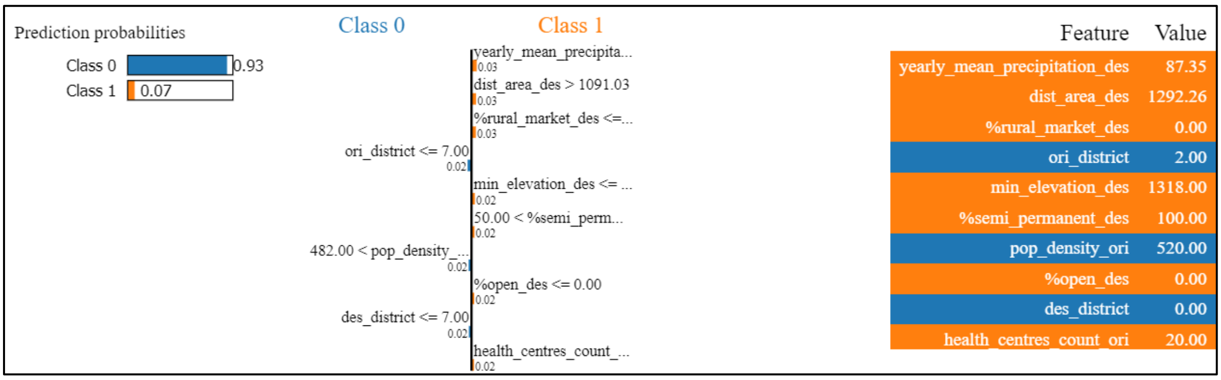


Figure 18: LIME explanation for the SVM model, which was not able to predicted the food flow for the 110th instance.

Another example (Figure 20) of an expected flow is with 120th instance. Here also a flow was expected, class 1 successful prediction. RF successfully identified the expected flow with 0.01 and 0.39 probabilities respectively for class1 and class 0. However, SVM failed to identify the expected flow (Figure 21) where it outputted a probability of 0.02 for class1 and 0.98 for class 0. Features that contributed positively to the predictions in both cases were the potato produce at origin district, the distance between district as well as yearly mean precipitation.

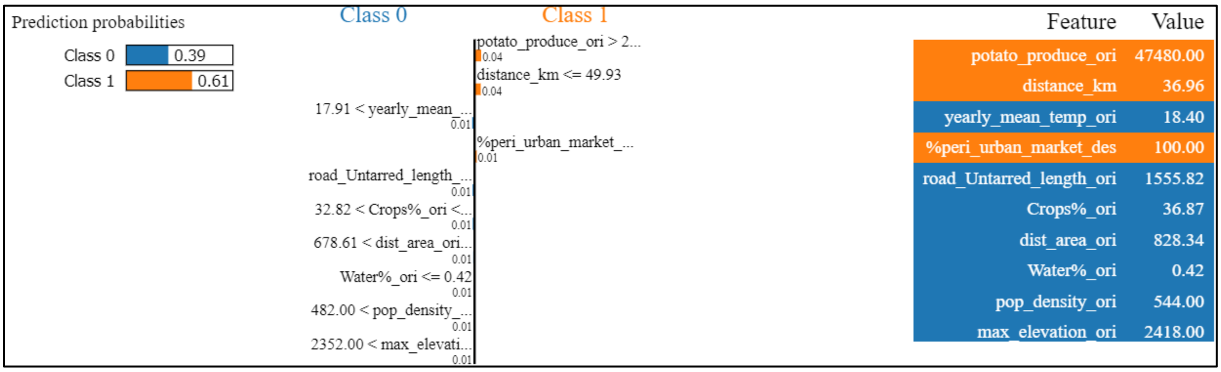


Figure 19: LIME explanation for the RF model, which successfully predicted the food flow for the 120th instance

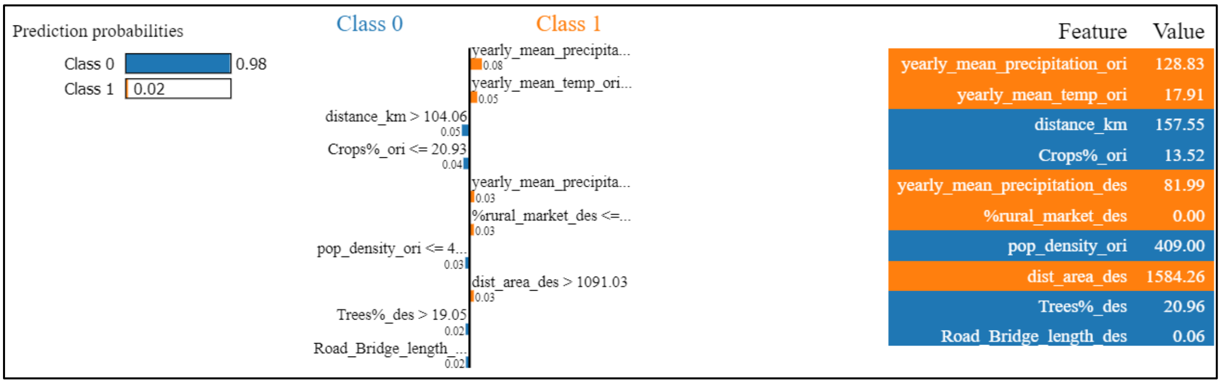


Figure 20: LIME explanation for the SVM model, which successfully predicted the food flow for the 120th instance

4. DISCUSSION

4.1. Summary:

Understanding food flows is crucial to ensuring food security, especially in rapidly urbanizing areas where demand for efficient and resilient food distribution systems are necessary (Ofori et al., 2022). Food flows are essential as they help balance supply and demand of food, ensuring surplus of production in some areas can meet needs in other areas.

This study specifically focused on predicting food flows using Random Forst and Support vector Machine as ML models. Irish potato was chosen as the focus crop and it's flow distribution among 30 districts of Rwanda. The research combined concepts of classic gravity model with machine learning techniques, relying on Explainable Artificial Intelligence techniques (xAI) to enhance the transparency of the predictive models at particular instances of the flow models. Objectives included compiling a comprehensive database of socio-economic and environmental factors along with district pair food flows and leveraging Machine learning methods to predict whether a particular district pair presents Irish potato food flows or not. It is observed that used data samples contained imbalanced flow classes where class 1 (representing the presence of flow) was underrepresented in the used dataset. Only 7% of the used data samples had a presence of flow. This affected the flow predictions on both Random forest and Support Vector machine models. Environmental features were the most flow influencers compared to socio-economic features. This was more demonstrated at the flow origin districts compared to flow destination districts mostly due to the physical characteristics of origin districts.

This research contributes significantly to the food flow studies literature, by addressing a notable gap identified in prior research. Previous studies largely concentrated on urban food flows within cities or between cities and their surrounding rural areas, often at National or international level by using broader datasets that may overlook finer details of local food flows. For instance (Moschitz & Frick, 2020) emphasized on broader city food flows analysis providing detailed information on current situation of urban food provisioning. (Mkondiwa et al., 2022) develops a spatially detailed model for district food flows in Malawi, highlighting the lack of district-level data. Similarly, (Y. Zhou et al., 2021) created an ML model to predict food insecurity in sub-Saharan Africa, while highlighting the importance of model transparency and adaptability to policy makers. Additionally (Lin et al., 2019), also developed a model that estimates food flows between USA counties, enabling detailed supply chain analysis. To the best of our knowledge, this study is one of the first to predict district level food flows in sub Saharan region at the district level using machine learning models coupled with xAI techniques, contributing to a more nuanced understanding of local food distribution. This approach also provides a replicable model for other regions facing similar challenges. By advancing the application of machine learning in food flow studies, this research supports the achievement of sustainable development Goal 2 which aims at ending hunger, improve nutrition and promote sustainable agriculture.

Key findings overview

The Random forest (RF) and Support Vector Mahine models demonstrated high accuracy in predicting district level Irish potato flow. The RF achieved an overall accuracy of 94% while the SVM model achieved 93%. Both models were trained on a comprehensive dataset of district level Irish potato flows that included environmental and socio economic features. However, it is important to note that the dataset presented imbalanced classes where district pairs that contained Irish potato flows were about 7% of the total data samples. The remaining dataset comprised the absence of Irish potato flow. F1 score, which is the harmonic mean of precision and recall , was used to evaluate the class prediction accuracy of the model. On the Random Forest (RF) model F1 score of class 0 (absence of flow) was 0.96 whereas on class 1 (Presence of flow) was 0.61. Similarly, the SVM model achieve an F1 score of 0.96 on class 0 and 0.59 on class 1. These f1 scores shows that both models were accurate at predicting the absence of Irish potato flows (class 0) than the presence of flow (class1), reflecting an imbalance in the dataset where instances of Irish potato flows were less frequent. LIME, an xAI technique was used to further investigate particular district pair instance prediction of potato flows and its most influencing features. It was observed that Environmental factors notably at the origin district were the most flow influencers compared to socio-economic features.

4.2. Interpretation of Significant factors

The analysis revealed several key factors influencing Irish potato flows in Rwanda are discussed below:

I. Yearly mean temperature & precipitation

This environmental feature emerged as important in the Irish potato flow predictions on both RF and SVM, yearly mean temperature and precipitation were the most significant flow predictors .
Irish potatoes (Solanum tuberosum) generally grows best in cool, temperate climates. The ideal temperature range for growing Irish potatoes during day time is between 15°C and 20°C. Nighttime temperatures are 10°C to 15°C. Temperatures above 30°C can inhibit tuber development , while temperatures below 10°C can slow down Irish potatoes growth significantly. Potatoes also require a frost-free period of 90 to 120 days to mature properly (Zemba et al., 2013).
The North-western districts of Rwanda are particularly well suited for Irish potato cultivation due to their climatic and geographical factors namely, altitude, temperature, rainfall , soil and seasonal variability. These regions are characterized by high altitude ranging from 1800-2500 meters above sea level figure 1 representing the study area digital elevation model can illustrate it. These altitude provides a cooler climate that is ideal for potato growth, with temperatures falling within the optimal range for potato cultivation. These regions also receive adequate rainfall, essential for potato cultivation. The soil in these regions are also generally fertile and well drained which is beneficial to potatoes. The seasonal variability also allows farmers to plan their planting and harvesting cycles to align with the optimal growing conditions ensuring good yields.

II. Maximum elevation

This factor was also highly influential in the model's predictions of Irish potato flows. This study revealed that elevation, alongside good potato climatic conditions plays an important role in predicting Irish potato flows. High elevations areas may also face challenges as cooler temperatures and rugged terrain can affect potato yield and transportation efficiency. These factors influence the volume and reliability of potato flows from these regions.

III. District pair distance

The geographic distance between districts pairs emerged as an influencing factor as well. It was fundamental to the gravity model and also reflecting the transportation hindrance associated with moving food over long distances. It was observed that shorter distances facilitated easier and effective movement of potatoes between districts. This is further enhanced by the potato flow channel sketch illustrated in figure 3 and images in appendix. Potatoes are carried locally on bicycles from the potato field to the markets close by or in between neighbouring districts.

IV. Potato Production quantity

Additionally, produced potatoes per district which was extracted from the yearly Agri-survey datafile was influential as well. This indicated that potatoes produced in a district directly impacts the availability of surplus quantity for distribution. Higher yields indicate greater production capacity, enabling more significant flows to other districts. This emphasized that areas with higher production can serve as major suppliers.

V. Land use patterns

The proportion of landuse landcover was another influencing factor of the potato flows where Rangeland and cropland portion percentages were more impactful. This reflects the district's agricultural focus and capacity. Districts with higher percentages of agricultural land were more likely exhibiting potato flows as they are likely to have influential agricultural systems. Some landuse classes also lowered the presence chance of potato flows like lower percentage of cropland coverage and more water landcover percentage present in a district.

VI. Market characteristics

Market characteristics factors were at the least influencing features where they were shown as less influential to the potato flow between districts. These features included the type and market infrastructures i.e.: urban, rural, peri-urban markets or market infrastructure being permanent, semi-permanent or open. These features exhibited less influence which illustrated that that potato flows could still be observed regardless the presence or absence of these features.

4.3. Classification results

The RF & SVM models coupled with LIME, an xAI technique provided comprehensive insights into factors influencing food flows. The RF model's overall accuracy was 94% , The feature importance global interpretation also showed that yearly mean temperatures, maximum elevation and district pair distance were most influential features to the prediction.

The SVM model, although slightly less accurate with 93% of overall accuracy, collaborated the findings of the RF model. Both models emphasized the influence of environmental and social factors in food flows.

The successful classification of the presence or absence of potato flows in between district was due to a strategic splitting technique by paying attention to the data characteristics. (Nath & Subbiah, 2018) of highlights the need for a diverse and balanced dataset splitting for effective model's predictions. This allows the model to capture relevant key information which was the case to our model. We have to note that the dataset presented an imbalance between classes where 7% of the dataset only presented the Irish potato flows. This is reflected by the F1 score from both models (RF, SVM) where particularly for class 1 (presence of flows) was significantly low compared to class 0 (absence of flows) which had enough representing samples.

4.4. Local interpretable model-agnoostic explanations (LIME)

LIME, which was used to give insights in model's particular predictions, highlighted local importance of contributing features. LIME was utilized in trying to capture the details at particular pair instance, this technique successfully revealed influencing features at particular instance predictions. 870 model instances are present representing the total training pair samples. To explain more, examples were presented in the results section, i.e.: instance number 48, representing a flow between Gakenke and Bugesera districts, LIME illustrated that the model outputted probabilities of 0.83 for class 1 and 0.17 for class 0. This shows the internal model's decision at that particular case instance which in the end affect the overall model's prediction. Positive contributing features (highlighted in orange) were the yearly mean temperature, the LULC class of cropland, the district area and minimum elevation just to name a few. Here the district area was shown as locally influencing while on the model's global interpretation, it wasn't the most influential. This again emphasizes the need for model's clarity while contributing to the ongoing research or transparent and interpretable models. global

4.5. Limitations

Class imbalance: An important challenge during this study and model's predictions was the class imbalance. This is an unequal representation of predicted classes which introduces bias and result in poor model's overall accuracy (Chakravarthy et al., 2019). The models performed well in predicting the absence of food flows(class 0) but struggled with predicting the presence of flows (class 1). This imbalance likely led to lower f1scores for class 1. Future work could focus on techniques to address class imbalance. Mores advanced sampling methods can be used (Dubey et al., 2014).

Temporal Dynamics: this study used a static temporal scale, 2023. This was due to the surveyed primary market data that had flow data. Incorporating temporal dynamics such as Potato seasonal variability and time series data could improve the model's ability to predict food flow in a more refined manner (Davis & Pineda Munoz, 2016). Also some datasets were not yet available for the considered time stamp. i.e.: Yearly Climate data like the precipitation and temperature. The availability of the corresponding time stamp data could improve the overall accuracy.

Data Scale: In this study some features important socio-economic features were not considered to the absence of data at the desired district scale. i.e.: age, economic wealth or the education level of district population. These could have explained in details the characteristics of the districts at study How do you expect this affect the results?

Flow quantity: Food item flow quantity was missing in the acquired dataset. The district produced potato quantity was used in this study. This does not represent the moved flow quantity across different districts. Having this dataset helps in the understanding of local food flows and would help to characterize better the food flows at study. The availability of this data could have improved the overall accuracy and reliability of this study. Further studies can focus on acquiring precise flow quantities.

4.6. Implications

Insights from this study reveal several implications for policy makers and practitioners aiming to enhance food flows and food security in Rwanda:

I. **Infrastructure Development :** understanding food flows on the district level can help the government plan for infrastructure development. Most potato producing regions are found in highly elevated areas as shown with the most important features, improving transportation infrastructure which could help in food movement , particularly in these areas can facilitate more efficient flows. Investment in roads or potato storage facilities can significantly reduce transportation costs which can be reflected on the final consumer price.

II. **Support for High-Yield Districts:** policies that support and enhance productivity in observed high-yield districts can help balance potato availability across the country. This may include investing in Irish potato seed multiplication. This was also noted as an input from farmers during the field surveys. This can help relieve constraints imposed by the local potato seed costs and availability while focusing on improving productivity.

III. **Market Development:** Well-functioning markets are critical for effective food distribution and can help in stabilizing prices and improving access to food (Gouel, 2013). Understanding food flows can help strengthen market infrastructure as it would give insights on local distribution dynamics. Additionally ensuring the availability of reliable market information can enhance the efficiency of distribution systems. This involves improved market facilities, supporting local market cooperatives and leveraging digital platforms for market information dissemination.

4.7. Future Study Directions

Based on findings from this study, the following recommendations are proposed further to enhance the understand the implication of this research and impact.

I. **Integration of market level flows:** While this study provides valuable insights, it has certain limitations. The reliance on specific datasets, such as market district level data flows. While the surveyed data was at market level , the dataset only comprised of reported district level food flows. Recording the finest, small scale data i.e. market level data flows can increase the resolution of details and reveal more factors that influence local food flows.

II. **Scope & temporal expansion:** while this study only focused on Irish potato flows**,** future research could expand the scope by including a wider variety of crops which are important to the local context. Further exploration of other machine learning models and xAI techniques could also enhance the robustness and interpretability of the predictions. Additionally, the temporal scale can also be increased comprising a longer year periods.

5. ETHICAL CONSIDERATIONS

This research strictly followed ethical standards, in accordance with the university of Twente's Research Ethics Policy. This commitment ensures the protection of participants' privacy, the welfare of their communities, and the preservation of the environment. All procedures involving human participants adhered to these guidelines, ensuring informed consent, confidentiality, and the right to withdraw without any consequences.

Data collection and handling was conducted with the utmost respect for privacy and mindful of the local customs and traditions. Ethical approval was sought from the relevant Institutions before commencing the study. The research team initiated contact and clearly conveyed detailed study purpose to ensure flexibility. Furthermore, no potential conflicts of interest occurred between researchers and involved communities.

Throughout the research process, the engagement with local communities, seeking their input towards food flows was done. The welfare of participants and the integrity of the research data will has always been in our foremost priorities. While adhering to these ethical principles, we aimed to conduct a responsible and socially accountable study that fosters resilient Agri-food systems and food security

6. CONCLUSION

This research delved into understanding food flows at the district level to ensure food security, particularly within the context of developing countries like Rwanda. The study was driven by the need to understand food flow at a more localized level as this can lead to efficient and robust food distribution systems. Objectives included understanding and predicting food flows at district level with a specific focus on Irish potato distribution among its 30 districts of Rwanda. To achieve this, the study employed two machine learning (ML) models, Random Forest (RF) and Support Vector Machine (SVM), coupled with Explainable Artificial Intelligence (xAI) techniques. The integration aimed at enhancing the transparency and interpretability of predicted outcomes, providing a robust framework for policy-making.

The findings revealed that both the RF and SVM models demonstrated high accuracy in predicting Irish potato food flows, achieving overall accuracies of 94% and 93%, respectively. These models were trained on a comprehensive dataset encompassing environmental and socio-economic features. Despite the challenges posed by the dataset imbalance, several significant factors emerged as key influencers of food flows especially environmental features. These included yearly mean temperature, maximum elevation, district pair distance, land use patterns. The least influencing features were market characteristics. The use of LIME, an xAI technique, clarified the decision-making processes of the ML models by identifying the contributing features for specific predictions. This enhanced the transparency value of the models, making them more useful for policy-making and practical applications.

This study supports the achievement of Sustainable Development Goal 2, aiming to end hunger, improve nutrition. The integration of Machine learning models and a detailed focus on local districts flows contributes to food flow literatures, paving the way for future research and practical applications in other regions and contexts.

LIST OF REFERENCES

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., & Ridella, S. (2012). The "K" in K-fold Cross Validation. *The European Symposium on Artificial Neural Networks*.

Bala, R. (2023). Food Security for Sustainable Future: Challenges, Strategies and Solutions. *IJFMR - International Journal For Multidisciplinary Research*, *5*(6). https://doi.org/10.36948/IJFMR.2023.V05I06.8868

Benton, T. G. (2020). COVID-19 and disruptions to food systems. *Agriculture and Human Values*, *37*(3), 577–578. https://doi.org/10.1007/S10460-020-10081-1/METRICS

Bricas, N., Barles, S., Billen, G., & Routhier, J.-L. (2019). *Urbanization Issues Affecting Food System Sustainability*. 1–25. https://doi.org/10.1007/978-3-030-13958-2_1

Capone, R., El Bilali, H., Debs, P., Cardone, G., & Driouech, N. (2014). Food System Sustainability and Food Security: Connecting the Dots. *Journal of Food Security*, *2*(1), 13–22. https://doi.org/10.12691/JFS-2-1-2

Carrillo, A., Cant'u, L. F., & Noriega, A. (2021). Individual Explanations in Machine Learning Models: A Survey for Practitioners. *ArXiv.Org*.

Chakravarthy, A. D., Bonthu, S., Chen, Z., & Zhu, Q. (2019). Predictive Models with Resampling: A Comparative Study of Machine Learning Algorithms and their Performances on Handling Imbalanced Datasets. *International Conference on Machine Learning and Applications*, 1492–1495. https://doi.org/10.1109/ICMLA.2019.00245

Condratchi, L. (2014). *Supply chains in agriculture and food production*.

Davis, M., & Pineda Munoz, S. (2016). The temporal scale of diet and dietary proxies. *Ecology and Evolution*, *6*(6), 1883–1897. https://doi.org/10.1002/ECE3.2054

Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. *ArXiv.Org*.

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J. (2014). Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study. *NeuroImage*, *87*, 220–241. https://doi.org/10.1016/J.NEUROIMAGE.2013.10.005

Engemann, H., & Jafari, Y. (2022). COVID-19 and changes in global agri-food trade. *Q Open*, *2*(1), 1–26. https://doi.org/10.1093/QOPEN/QOAC013

FAO. (2023a). Food Outlook Biannual report on global food markets. *Food Outlook – Biannual Report on Global Food Markets*. https://doi.org/10.4060/CC3020EN

FAO. (2023b). *Rwanda | African Roots and Tubers | Food and Agriculture Organization of the United Nations*. https://www.fao.org/in-action/african-roots-and-tubers/countries/rwanda/en/

FSIN. (n.d.). *Food security information Network Global report on food crises*.

Giertz, Å., Gray, G., Mudahar, M. S., Rubaiza, R., Galperin, D., & Suit, K. (2015). *RWANDA Agricultural Sector Risk Assessment Agriculture globAl PrActice technicAl AssistAnce PAPer*. www.worldbank.org

Gouel, C. (2013). *Food Price Volatility and Domestic Stabilization Policies in Developing Countries*. https://papers.ssrn.com/abstract=2245443

Greenwell, B. M., & Boehmke, B. C. (2020). Variable Importance Plots - An Introduction to the vip Package. *The R Journal*, *12*(1), 343–366. https://doi.org/10.32614/RJ-2020-013

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences*, *12*(3). https://doi.org/10.3390/APP12031353

James, S. W., & Friel, S. (2015). An integrated approach to identifying and characterising resilient urban food systems to promote population health in a changing climate. *Public Health Nutrition*, *18*(13), 2498–2508. https://doi.org/10.1017/S1368980015000610

Kabir, M., Salim, R., & Al-Mawali, N. (2017). The gravity model and trade flows: Recent developments in econometric modeling and empirical evidence. *Economic Analysis and Policy*, *56*, 60–71. https://doi.org/10.1016/J.EAP.2017.08.005

Karg, H., Akoto-Danso, E. K., Amprako, L., Drechsel, P., Nyarko, G., Lompo, D. J. P., Ndzerem, S., Sidibé, S., Hoschek, M., & Buerkert, A. (2023). A spatio-temporal dataset on food flows for four West African cities. *Scientific Data 2023 10:1*, *10*(1), 1–14. https://doi.org/10.1038/s41597-023-02163-6

Kulkarni, V., & Sinha, P. K. (2014). *Effective Learning and Classification using Random Forest Algorithm*.

Lin, X., Ruess, P. J., Marston, L., & Konar, M. (2019). Food flows between counties in the United States. *Environmental Research Letters*, *14*(8), 084011. https://doi.org/10.1088/1748-9326/AB29AE

López, V., Fernández, A., & Herrera, F. (2014). On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, *257*, 1–13. https://doi.org/10.1016/J.INS.2013.09.038

Mishra, A., Bruno, E., & Zilberman, D. (2021). Compound natural and human disasters: Managing drought and COVID-19 to sustain global agriculture and food sectors. *Science of The Total Environment*, *754*, 142210. https://doi.org/10.1016/J.SCITOTENV.2020.142210

Mkondiwa, M., & Apland, J. (2022). Inter-district food flows in Malawi. *Food Security*, *14*(6), 1553–1568. https://doi.org/10.1007/s12571-022-01302-y

Mkondiwa, M., Apland, J., Mkondiwa, M., & Apland, J. (2022). Inter-district food flows in Malawi. *Food Security: The Science, Sociology and Economics of Food Production and Access to Food*, *14*(6), 1553–1568. https://doi.org/10.1007/S12571-022-01302-Y

Mohan, L., Pant, J., Suyal, P., & Kumar, A. (2020). Support Vector Machine Accuracy Improvement with Classification. *International Conference on Computational Intelligence and Communication Networks*, 477–481. https://doi.org/10.1109/CICN49253.2020.9242572

Mor, K., & Sethia, S. (2014). Factors that influence household and individual food consumption : a review of research and related literature. *Management Insight*.

Moschitz, H., & Frick, R. (2020). City food flow analysis. A new method to study local consumption. *Renewable Agriculture and Food Systems*, *36*(2), 150–162. https://doi.org/10.1017/S1742170520000150

Moschitz, H., & Frick, R. (2021). City food flow analysis. A new method to study local consumption. *Renewable Agriculture and Food Systems*, *36*(2), 150–162. https://doi.org/10.1017/S1742170520000150

Nath, A., & Subbiah, K. (2018). The role of pertinently diversified and balanced training as well as testing data sets in achieving the true performance of classifiers in predicting the antifreeze proteins. *Neurocomputing*, *272*, 294–305. https://doi.org/10.1016/J.NEUCOM.2017.07.004

National Institute of Statistics of Rwanda (NISR). (2023, February 1). *The Fifth Rwanda Population and Housing Census, Main Indicators Report*. https://www.statistics.gov.rw/publication/main_indicators_2022

Ofori, B. D., Pabi, O., Nukpezah, D., Annan, J., & Wang, H.-C. (2022). Food system flows and distribution for the Accra metropolis: Unfolding the policy dimensions. *Legon Journal of the Humanities*, *32*(2), 145–169. https://doi.org/10.4314/LJH.V32I2.7

Puma, M. J., Bose, S., Chon, S. Y., & Cook, B. I. (2015). Assessing the evolving fragility of the global food system. *Environmental Research Letters*, *10*(2), 024007. https://doi.org/10.1088/1748-9326/10/2/024007

Räz, T. (2022). ML Interpretability: Simple Isn't Easy. *Studies in History and Philosophy of Science*, *103*, 159–167. https://doi.org/10.1016/J.SHPSA.2023.12.007

Republic of Rwanda. (2018). *Third National Communication: Report to the United Nations Framework Convention on Climate Change*. www.rema.gov.rw

Sage, C. (2019). Food Security. *International Encyclopedia of Geography*, 1–7. https://doi.org/10.1002/9781118786352.WBIEG0791.PUB2

Saini, R., & Ghosh, S. K. (2018). Crop classification on single date sentinel-2 imagery using random forest and support vector machine. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII–5*, 683–688. https://doi.org/10.5194/ISPRS-ARCHIVES-XLII-5-683-2018

Swaminathan, M. S., & Bhavani, R. V. (2013). Food production & availability - Essential prerequisites for sustainable food security. *The Indian Journal of Medical Research*, *138*(SEP), 383–391. https://doi.org/10.1142/9789813200074_0025

The World Bank Group. (2021). *Climate Risk Profile: Rwanda*. www.worldbank.org

Yang, J., Wang, J., Xu, C., Liu, Y., Yin, Q., Wang, X., Wang, L., Wu, Y., & Xiao, G. (2021). Rice supply flows and their determinants in China. *Resources, Conservation and Recycling*, *174*. https://doi.org/10.1016/j.resconrec.2021.105812

Zemba, A. A., Wuyep, S., Adebayo, A., Jahknwa, C. J., & Ambrose. (2013). Growth and Yield Response of Irish Potato (Solanum Tuberosum) to Climate in Jos-South, Plateau State, Nigeria. *Global Journal of Human-Social Science Research*.

Zhou, D., Matsuda, H., Hara, Y., & Takeuchi, K. (2012). Potential and observed food flows in a Chinese city: A case study of Tianjin. *Agriculture and Human Values*, *29*(4), 481–492. https://doi.org/10.1007/S10460-012-9374-X/TABLES/3

Zhou, Y., Lentz, E., Michelson, H., Kim, C., & Baylis, K. (2021). Machine learning for food security: Principles for transparency and usability. *Applied Economic Perspectives and Policy*, *44*(2), 893–910. https://doi.org/10.1002/AEPP.13214

Zou, T., Dawodu, A., Mangi, E., & Cheshmehzangi, A. (2023). Exploring Current Trends, Gaps & Challenges in Sustainable Food Systems Studies: The Need of Developing Urban Food Systems Frameworks for Sustainable Cities. *Sustainability (Switzerland)*, *15*(13). https://doi.org/10.3390/SU151310248

Potato Flow Channel explanations

Potato field harvesting

Bike potato transport to the collection center

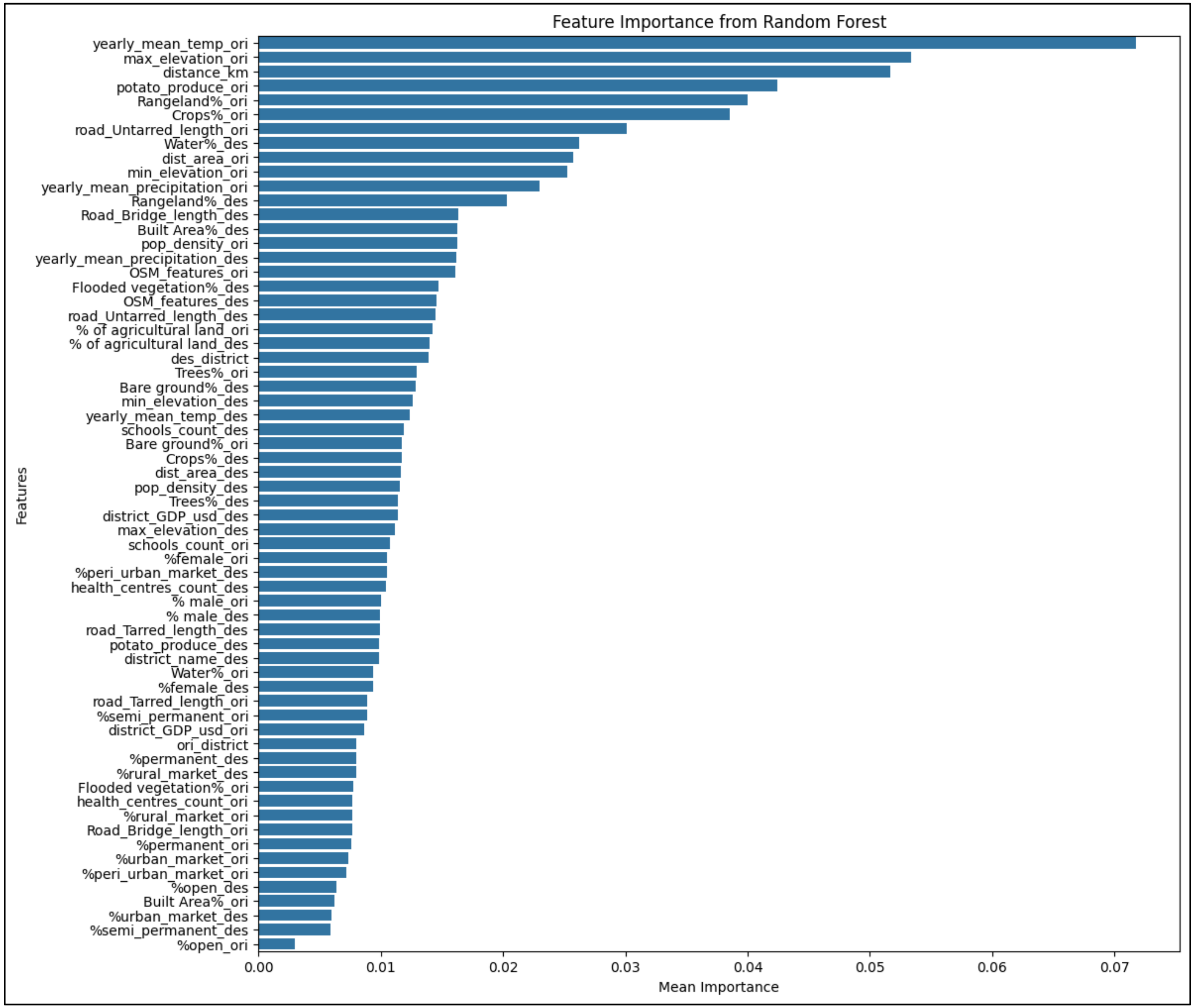
A potato collection center where wholesalers collect potatoes

Inside a potato collection center house

Truck loading for export outside the district

Visiting a potato collection center

Feature importance of RF



Feature importance of SVM