Analyzing and Predicting Urbanization Patterns in Nairobi Using Geographic Information Systems and Data Mining Techniques

By

Yvonne Makena Baariu

169522

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

Institute of Mathematical Sciences and iLab Africa

Strathmore University

Nairobi, Kenya

June, 2025

Declaration and Approval

Declaration

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

Student's Name: **Yvonne Makena Baariu**

Sign: _____ Date: 03/28/2025

Approval

The dissertation of **Yvonne Makena Baariu** was reviewed and approved by the following:

Dr. Kennedy Senagi,

28th March 2025

Lecturer, Insititute of Mathematical Sciences,

Strathmore Univeristy.

Abstract

Urbanization, characterized by the expansion and development of towns due to population growth from both natural increase and migration, has profoundly reshaped cities worldwide. Over the past four decades, the proportion of people residing in urban areas more than doubled. According to the United Nations, projections indicated that by 2050, approximately 68% of the global population would be living in cities. This growth implied that nearly 2.5 billion additional people would reside in urban areas by 2050, with the majority of this expansion occurring in Asia and Africa.

In Kenya, rapid urbanization resulted in various challenges, including the proliferation of informal settlements and inadequate infrastructure. Nairobi, as a major metropolis, experienced substantial urban growth, exacerbating these issues. This study sought to model urbanization patterns in Nairobi through the use of Geographic Information System (Geographic Information Systems (GIS)) and Remote Sensing, alongside the application of Data Mining techniques to forecast future urban expansion. The methodology involved analyzing Land Satellite (LANDSAT) 8 satellite imagery from 2014 and 2024, computing the Normalized Difference Built-up Index (Normalized Difference Built-up Index (NDBI)), and utilizing change detection techniques to assess urban development. The findings can provide crucial insights for policymakers and urban planners, facilitating sustainable urban development and addressing challenges associated with rapid urbanization. This research contributed to a broader understanding of urban dynamics in Kenya and presented a predictive framework for future urban planning efforts.

Keywords: Urbanization, GIS, Predictive Modeling, Machine Learning

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

ART-MMAP Adaptive Resonance Theory with Mixture Model Adaptive Process

API Application Programming Interface

ArcGIS Aeronautical Reconnaissance Coverage Geographic Information System

CA Cellular Automata

CSV Comma Separated Values

CRISP-DM Cross-Industry Standard Process for Data Mining

DEM Digital Elevation Model

ERD Entity Relationship Diagram

GEE Google Earth Engine

Geo-Tiff Geo-referenced Tagged Image File Format

GPS Global Positioning System

GIS Geographic Information Systems

JSON JavaScript Object Notation

LANDSAT Land Satellite

LRM Logistic Regression Model

LSI Landscape Shape Index

LULC Land Use and Land Cover

ML Machine Learning

NDBI Normalized Difference Built-up Index

NDVI Normalized Difference Vegetation Index

NDWI Normalized Difference Water Index

NIR Near Infrared

NLSI Normalized Landscape Shape Index

NDVI Normalized Difference Vegetation Index

QGIS Quantum Geographic Information System

RF Random Forest

ROC-AUC Receiver Operating Characteristic - Area Under the Curve

RS Remote Sensing

SAVI Soil Adjusted Vegetation Index

SDG Sustainable Development Goals

SMOTE Synthetic Minority Oversampling Technique

SMOTETomek Synthetic Minority Over-sampling Technique combined with Tomek Links

SVM Support Vector Machine

SWIR Short Wave Infrared

SVM Support Vector Machine

UI User Interface

UML Unified Modelling Language

XGBoost Extreme Gradient Boosting

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Kennedy Senagi, for his invaluable guidance, support, and encouragement throughout this research. His expertise and unwavering dedication have been instrumental in shaping the direction and quality of this work.

I am also grateful to Strathmore University for providing the academic environment and resources that have supported my research. Additionally, I extend my appreciation to Google Earth Engine for granting access to crucial geospatial data, which has been fundamental to the analysis and findings of this study.

I am deeply thankful to my classmates for their collaboration, insightful discussions, and continuous support throughout this journey. Their encouragement and shared knowledge have greatly contributed to my learning and the successful completion of this study.

A heartfelt thank you to my parents for their unconditional love, sacrifices, and constant encouragement. Their unwavering belief in me has been a source of strength and motivation throughout this process.

Above all, I give thanks to God for granting me the resilience, and strength to complete this research. Without His guidance and grace, none of this would have been possible.

To everyone who has contributed to this journey in one way or another, I am deeply grateful. Your support has been invaluable.

Chapter 1: Introduction

1.1 Background

Some scholars claim that urbanization is one of the most tangible indicators of shifting global human settlements in the 21st century. Prior to 1950, urbanization mostly occurred in the developed countries and it was driven mainly by industrialization in Europe and North America in the 19th and 20th centuries (Zhang, 2016). Today, more than 56% of the population lives in Urban areas, with projections estimating that by 2050, 68% of the global population will reside in cities(Ritchie and Roser, 2018).

Between the 19th and early 20th century, the U.S.A. experienced significant urbanization, a trend mirrored by most European countries during the same period(Zhang, 2016). However, since 1950, the rate of urbanization has slowed in many developed nations. In some cases, large cities have seen population declines as people migrated from urban areas to the countryside. As a result, the urbanization process in much of the developed world has nearly reached its conclusion.(Cohen, 2004). Today, the challenges of rapid urbanization mainly lie in the developing countries(Zhang, 2016).A review of global trends highlights that rapid urban expansion in cities across Asia, Latin America, and Africa often leads to urban sprawl, unplanned settlements, and environmental pressures(Cohen, 2004). These regions face the dual challenge of accommodating new urban populations while addressing infrastructural deficiencies, housing shortages, and environmental sustainability concerns.

Africa has seen dramatic rates of urbanization, often driven by rural-urban migration. In Nigeria for instance, urban expansion has exacerbated environmental problems such a deforestation and pollution and led to challenges such as overcrowding, slums, and environmental degradation (Avis, 2019). Libya's Tripoli metropolitan area has faced urban sprawl that threatens its ecological balance. Researchers have found that rapid expansion has drastically altered land use patterns(Alsharif and Pradhan, 2014).In Ethiopia, similar challenges have dramatically changed land use, contributing to the loss of agricultural lands and increased pressure on infrastructure (Dadi et al., 2016)

In Kenya, urbanization has followed this broader African trend, with Nairobi, the capital, acting as a primary hub for internal migration. The city's expanding population has lead to housing shortage, increased poverty, poor sanitation, overcrowding and en-

vironmental strain(Odhiambo, 2022).Kenya's urban landscape, as highlighted by (Alkire et al., 2011), faces an increasing challenge of urban poverty, particularly in informal settlements because the rate at which urban centers are growing often exceeds the capacity of local governments to provide adequate services and infrastructure. Despite these issues, Nairobi's growth also presents opportunities for economic expansion and improved livelihoods, but only if the city's development can be properly managed.

1.2 Problem Statement

Nairobi, like many cities in the Global South, has experienced uncontrolled urban expansion (Farhan et al., 2023). As a result, the city faces challenges such as traffic congestion, inadequate housing, and environmental degradation (Mundia and Aniya, 2007). Current urban planning methods are often reactive rather than proactive, leading to uncoordinated development Odhiambo (2022).

The loss of agricultural land threatens food security, while slum proliferation exacerbates poverty and inequality. Environmental degradation further diminish the quality of life, hindering Nairobi's potential as a sustainable and inclusive urban center. As Nairobi inevitably continues to urbanize, city planners and policy makers will need more effective methods and tools such as urban expansion forecasting models. These will enable them to implement informed, data driven and sustainable strategies (Jaad and Abdelghany, 2021). It is therefore important to develop a predictive model that uses historical and real-time data to forecast urbanization trends in Nairobi.

This research addressed this gap by combining GIS, remote sensing, and data mining techniques to create a model that can predict urban growth patterns in Nairobi.

1.3 Research Aim

To conduct a comprehensive review of the existing literature related to urbanization , with a focus on Geographic Information Systems (GIS) and data mining techniques as applied to the study of Nairobi.

1.4 Research Objectives

This research aimed to address the following objectives:

(a) To conduct a comprehensive review of the existing literature related to urbanization, with a focus on Geographic Information Systems (GIS) and data mining techniques as applied to the study of Nairobi

(b) To map historical and current urban growth using satellite imagery and GIS tools.

(c) To use remote sensing data to detect and measure land use/land cover changes over time.

(d) To develop a predictive model based on historical urbanization patterns using data mining techniques.

1.5 Research Questions

The research questions addressed in this study were as follows:

(a) What are the key spatial and temporal factors that have driven urbanization in Nairobi between 2014 and 2024, and how are these factors related to land-use, infrastructure development, and population growth?

(b) How can data mining techniques, such as regression, be effectively employed to analyze historical urbanization trends in Nairobi using GIS and remote sensing?

(c) Which machine learning models (e.g., logistic regression, neural networks) provide the highest accuracy in predicting future urban growth in Nairobi, and how does model performance vary based on the inclusion of key variables like distance to infrastructure or elevation?

(d) What is the role of remote sensing indices, such as the Normalized Difference Built-up Index (NDBI), in tracking urban expansion in Nairobi, and how accurately can these indices predict future urban growth?

1.6 Scope and Limitations of the Study

1.6.1 Scope

The scope of this research was limited to Nairobi County. The research focused on identifying urbanization patterns within Nairobi county only.

1.6.2 Limitations

Here are some of the limitations of this research work:

(a) Researching urbanization in developing nations is fraught with difficulties.

(b) The primary issue is the lack of an international classification scheme for urban environments. Most nations distinguish between populations living in urban and rural areas, but the definition of an urban area varies from nation to nation and, in some cases, even over time within a single nation.

(c) Inconsistent spatial data resolution: High-resolution satellite imagery may not be available for all areas of interest, or the images might lack clarity, making it difficult to detect subtle changes in land use.

(d) Budget limitations: Gathering high-quality spatial data, conducting fieldwork, and processing large datasets could require financial resources that might not be readily available, limiting the scale or depth of the research.

1.7 Research Justification

Urbanization in Nairobi presents both opportunities and challenges. Proper planning can foster economic growth, improve living standards, and reduce poverty. However, without appropriate tools to forecast urban expansion, Nairobi risks facing worsening environmental degradation, housing shortages, and infrastructure collapse (Arego, 2020). This study contributed to the growing body of knowledge on urbanization in African cities, and aligns with certain Sustainable Development Goals (SDG) goals for example: SDG 11: Sustainable Cities and Communities which aims to make cities inclusive, safe, resilient, and sustainable.

A working model can be used by planners and other relevant stakeholders to plan for the natural expansion of cities and also future cities, without the unwanted consequences discussed (Friesen et al., 2018)

Chapter 2: Literature Review

Urbanization is a defining global trend of the 21st century, with cities expanding at an unprecedented rate to accommodate increasing populations (Henderson, 2005). Nairobi, the capital and largest city of Kenya, has undergone significant urban growth in recent decades. Understanding the dynamics and underlying factors of this urban expansion is essential for informed urban planning, policy formulation, and sustainable development.

The analysis of urbanization patterns, especially in rapidly growing cities like Nairobi, has greatly benefited from advancements in Geographic Information Systems (GIS) and data mining techniques. Traditional urban growth planning methods predominantly relied on statistical models and land use maps, which proved to be inadequate for analyzing spatial and temporal urban expansion (Obura, 2009).

Geographic Information Systems (GIS) play a crucial role in spatial analysis by leveraging data from remotely sensed images (Xiuwan, 2002), offering valuable insights into the geographic distribution of urban expansion. Meanwhile, data mining techniques enable the extraction of meaningful patterns and trends from extensive datasets. When integrated, these technologies provide a robust approach to studying and forecasting urbanization, aiding in the effective management of urban growth.

This review presents an in-depth examination of existing studies on urbanization patterns, with a particular focus on the application of GIS, remote sensing, and data mining techniques. It discusses key theoretical frameworks, empirical research, and methodologies utilized in the field. By evaluating the current body of knowledge, this review aims to identify research gaps and highlight areas for future exploration.

Through synthesizing prior research, this literature review establishes a foundation for the present study, which seeks to analyze and predict urbanization trends in Nairobi by leveraging GIS, remote sensing, and data mining techniques.

2.1 Theoretical Framework

2.1.1 Urbanization Theories

Urbanization refers to changes in city size, density, and diversity. According to (Vlahov and Galea, 2002), it encompasses transformations in population distribution and settle-

ment patterns. In developing countries, urbanization is often characterized by a shift from rural to urban living, typically occurring as nations transition from agricultural economies to industrial and service-based economies (Turok and McGranahan, 2013). Developing regions, in this context, include Africa, East Asia, South Asia, Western Asia, Latin America, and the Caribbean (Bodo, 2019).

The scale and pace of urbanization in these regions are reflected in projections indicating an additional 2.5 billion people in urban areas by 2050, with approximately 90% of this growth occurring in Asia and Africa (Benna and Benna, 2018). Moreover, cities such as Lagos, Delhi, and Dhaka experienced urban growth rates of 85, 79, and 74 people per hour, respectively, in 2015 (Benna and Benna, 2018).

Region/Area	**Average annual rate of the urban population**				
(lr)2-6	1995-2000	2000-2005	2005-2010	2010-2015	**Entire period** 1995-2015
World	2.13%	2.27%	2.20%	2.05%	2.16%
High-income countries	0.78%	1.00%	1.00%	0.76%	0.88%
Middle-income countries	2.74%	2.77%	2.61%	2.42%	2.63%
Low-income countries	3.54%	3.70%	3.70%	3.77%	3.68%
Africa	3.25%	3.42%	3.55%	3.55%	3.44%
Asia	2.79%	3.05%	2.79%	2.50%	2.78%
Latin America and the Caribbean	2.19%	1.76%	1.55%	1.45%	1.74%
Europe	0.10%	0.34%	0.34%	0.33%	0.31%
North America	1.63%	1.15%	1.15%	0.04%	1.24%
Oceania	1.43%	1.49%	1.78%	1.44%	1.53%

Table 2.1: Average annual rate of the urban population, 1995-2015

Currently, 54% of the global population resides in urban areas, and this proportion is expected to rise to 66% by 2050. More than 90% of this urban growth will be concentrated in Asia and Africa. Notably, India and China, the two most populous nations in Asia, are projected to witness their urban populations increase by 404 million and 292 million, respectively, between 2014 and 2050 (un-, 2018)

While economic expansion in urban areas is widely recognized as a key driver of rural-to-urban migration, a purely economic perspective does not fully explain the phenomenon of "urbanization without growth," which was prevalent in sub-Saharan Africa during the 1980s and 1990s (Fox, 2012). Fox further argues that technological advancements have also been instrumental in driving rural-to-urban migration.

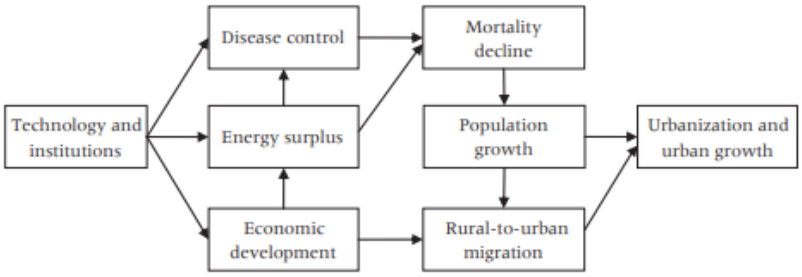


Figure 2.1: Urbanization Theory Influenced by Information Technology(Ritchie and Roser, 2018)

Figure 2.1 provides a stylized diagram of this theory of urbanization. In brief, according to (Fox, 2012), the fundamental drivers of urban transition include technological advancements and institutional developments that enhance disease control and ensure surplus energy availability. These factors contribute to a decline in mortality rates in both rural and urban areas, directly boosting urban population growth through natural increase and indirectly accelerating rural-to-urban migration (Alsharif and Pradhan, 2014).

Many classical urbanization theories originated from the industrial revolution experiences of European and American cities (Benna and Benna, 2018). Several key theories have been proposed to explain the process of urbanization:

Self-Generated Urbanization Theory: Historically, urban development in Western Europe and the United States has been closely linked to industrialization, serving as a model for developing nations. This perspective aligns with the original Lewis theory of labor transfer (Todaro, 1997), which posits that urbanization is primarily driven by rural-to-urban migration. This migration is stimulated by a wage disparity between rural and urban areas that emerges in the early stages of industrialization (Fox, 2012).

Modernization Theory: This theory argues that urbanization arises from societal innovations introduced through industrialization, technological advancements, information dissemination, and cultural diffusion (Bodo, 2019). It frames urbanization as an outcome of modernization, emphasizing the transition of societies from primitive stages (such as the Stone Age) to contemporary modes of operation (Kasarda and Crenshaw, 1991). The theory underscores the role of technology in shaping social organization and assumes that technological advancement is more influential than societal structure in driving urbanization.

Dependency/World-System Theory: This perspective emerged in response to the limitations of modernization theory in explaining urbanization patterns in developing countries (Bodo, 2019). It contends that urbanization in these regions is driven by external forces, either through deliberate coercion or the inherent mechanisms of global capitalism. The theory asserts that social changes in developing nations result from the structures and processes of the capitalist world system (Dear and Scott, 2018). It further argues that this system perpetuates inequalities, as developed world cities exploit primate capitals in developing regions as centers for wealth accumulation and transmission (Bradshaw, 1987).

Theory of Urban Bias: This theory suggests that rural populations are economically disadvantaged due to urban favoritism. Urban dwellers benefit from inexpensive goods produced in rural areas and enjoy urban infrastructure funded by tax revenues collected from rural populations (Dixon and McMichael, 2017). This urban bias creates significant disparities between rural and urban areas in terms of consumption, wages, and productivity levels. Consequently, many rural inhabitants migrate to cities in pursuit of better economic opportunities and improved living standards (Bradshaw, 1987).

2.1.2 Remote Sensing in Urbanization Studies

Remote sensing serves as a crucial tool for analyzing urbanization, allowing researchers to utilize large-scale, multi-temporal satellite imagery to track urban expansion patterns. However, Land Use and Land Cover (LULC) classification in Africa remains challenging due to spectral confusion arising from the complexity and heterogeneity of urban landscapes (Hadeel et al., 2011). In sub-Saharan cities, the mixed-pixel problem—where a single pixel represents multiple land cover types—is particularly prevalent due to the intricate spatial structures and the diverse materials used in construction (Simwanda and Murayama, 2017; Ngolo and Watanabe, 2023).

Although several pixel-based algorithms have been designed to mitigate this issue in medium-resolution imagery, their classification accuracy often remains suboptimal. Supervised classification methods, while effective, require significant expertise and time, with their performance largely dependent on the quality of training data (Yang and Huang, 2021). An alternative approach involves leveraging band combinations and spec-

tral indices, such as the Normalized Difference Built-up Index (NDBI), to improve the identification of built-up areas (Hurd, 2015). NDBI, a spectral index commonly used in spatial data analysis, enhances the distinction between urban environments—such as roads and buildings—and other land cover types like vegetation, water bodies, and bare soil (Ngolo and Watanabe, 2023).

The NDBI equation is: (1):

$$NDBI = (SWIR–NIR)/(SWIR + NIR) \tag{1}$$

The NDBI values lie between -1 to +1. Negative value of NDBI represent water bodies but higher value represent build-up areas(Kshetri, 2018).

Studies like (Mwakapuja et al., 2013), (Odhiambo, 2022), (Ngolo and Watanabe, 2023) have used spectral indices like NDBI, Normalized Difference Vegetation Index (NDVI), and Normalized Difference Water Index (NDWI) on Landsat images to extract land-cover information.

2.1.3 Geographic Information Systems (GIS) in Urban Studies

Geographic information is fundamental to urban planning. While GIS was introduced in the late 1960s, its early adoption was hindered by high hardware costs and software constraints (Masser and Ottens, 2019; Yeh, 1999). Today, GIS is widely employed for database management, visualization, spatial analysis, and modeling in urban development (Yeh, 1999). This study leverages GIS and remote sensing to examine urban expansion by tracking LULC changes, using satellite imagery to measure the extent, nature, and spatial distribution of these transformations (Fischer and Nijkamp, 1992). GIS provides a versatile platform for storing, visualizing, and analyzing digital data essential for detecting changes over time (Wu et al., 2006).

Analyzing LULC changes requires time-series data, with models such as Agent-Based Models (ABM) playing a crucial role in predicting urban expansion (Rajan and Shibasaki, 2000). The integration of GIS is vital for managing spatial datasets and recognizing patterns of spatial heterogeneity (Pijanowski et al., 2002). Many GIS-based models utilize

raster data structures, which represent space through a grid of uniformly sized cells, simplifying spatial analysis (Pijanowski et al., 2002).

2.1.4 Data Mining Techniques in Urban Analysis

Spatial Data Mining involves uncovering knowledge, spatial relationships, or patterns that are not explicitly stored in spatial databases (Rajesh, 2011).

Various methods have been developed to model urban expansion, including cellular automata (Cellular Automata (CA))–based simulations (Clarke and Gaydos, 1998) and multivariate statistical models (Wu and Yeh, 1997). CA models indicate that the relationship between urban growth and influencing factors is often non-stationary, meaning it differs across regions (Luo et al., 2008). This variability implies that identical factors may lead to distinct outcomes in different locations, highlighting the importance of modeling spatial dynamics for deeper insights (Luo et al., 2008).

Spatial data mining techniques for land-use and land-cover (LULC) changes can be broadly categorized into regression-based and transition-based models (Liu and Seto, 2008). Regression-based models establish functional relationships between land-use change and predictor variables to determine where future changes may occur (Abebe et al., 2023). A key advantage of these models is their ability to quantify the contribution of different variables to land-use change predictions (Pijanowski et al., 2002).

Transition-based models, on the other hand, focus on analyzing and forecasting spatial changes over time, particularly shifts in the state of spatial units. A notable example is CA, which is widely applied in simulating urban growth (Pijanowski et al., 2002) and can be used to model urban sprawl in Nairobi based on the influence of neighboring urban cells.

There is no single method for categorizing spatial data mining techniques. However, based on general data mining tasks, they can be grouped into:

Clustering And Outlier Detection: Spatial clustering involves grouping spatial objects into distinct clusters (Sumathi et al., 2008)

Classification: This technique identifies rules that divide a dataset into predefined classes. It is a predictive spatial data mining approach that first develops a model to analyze the

dataset (Alshalabi et al., 2013).

Trend Detection: This method identifies patterns in attribute changes relative to the surrounding spatial objects. Examples include logistic regression and CA (Sumathi et al., 2008).

2.1.5 Integrating Remote Sensing, GIS and Data Mining for Urbanization

The integration of GIS and remote sensing has greatly benefited from advancements in computing, global positioning systems (Global Positioning System (GPS)), and artificial intelligence (Artificial Intelligence (AI)) techniques (Ngolo and Watanabe, 2023). Notably, machine learning (Machine Learning (ML)) has enabled the rapid analysis of vast amounts of imagery, surpassing the efficiency of traditional methods.

Remote sensing datasets, including Landsat, Sentinel, and Spot imagery, play a crucial role in visualization, classification, and spatial analysis of various regions (Kshetri, 2018).

Integrating GIS with data mining techniques enables spatio-temporal analysis, allowing for the examination of land use changes and urban expansion over both space and time (Pampoore-Thampi et al., 2021). Applying this approach to Nairobi provides insights into the city's historical development while facilitating predictions of future urban growth, aligning with the core objectives of this study.

Predictive models such as the Adaptive Resonance Theory with Mixture Model Adaptive Process (ART-MMAP) neural network-based model (Liu and Seto, 2008) streamline the process by eliminating the need for multiple iterations to compute urban transition probabilities for individual cells in Nairobi. Consequently, this model effectively forecasts both the location and extent of future urban expansion (Liu et al., 2007).

2.2 Empirical Review

2.2.1 Studies on Urbanization in Nairobi and Similar Cities

Urbanization patterns vary significantly both across and within countries, highlighting the absence of a universal "urban transition" (Elmqvist et al., 2013). The global landscape of urbanization has also shifted, with the 20 fastest-growing urban regions now located in Asia and Africa rather than Europe or North America (Elmqvist et al., 2013).

In Africa, the urban population is projected to more than triple over four decades, rising from 395 million in 2010 to 1.339 billion by 2050, accounting for 21% of the world's estimated urban population (Güneralp et al., 2017). While Africa's rapid urbanization mirrors trends in other fast-urbanizing regions, the processes driving urban growth on the continent are distinct from those observed elsewhere (Elmqvist et al., 2013).

Despite the high rate of urban population growth, many African nations continue to exhibit strong urban primacy, where a single city—typically the capital—dominates in terms of population, economic activity, and political influence, significantly surpassing the next largest city (Güneralp et al., 2017).

In Kenya, the rapid growth of the urban population is driven by both natural population increase among urban residents (52%) and rural-to-urban migration (48%). As a result, Nairobi's urban primacy index has steadily risen between 1979 and 2010, indicating that a significant proportion of Kenya's urban population resides in the capital (Mundia and Aniya, 2007).

According to (Juma et al., 2019), this urban expansion has been described as "rapid, unplanned, and unmanaged," with 62% of Africa's urban population living in informal settlements. These areas are often marked by poor living conditions, overcrowding, and makeshift housing constructed from substandard materials such as mud, scrap wood, or corrugated metal (Juma et al., 2019).

A study conducted in Nigeria found that more than seven cities have populations exceeding one million, with Lagos and Kano each surpassing nine million inhabitants. Additionally, the proportion of Nigeria's population living in urban centers increased from 15% in 1960 to 50% in 2013, with projections indicating a further rise to 60% by 2015 (Adekola, 2016).

This rapid urbanization has resulted in challenges similar to those observed in other rapidly growing urban areas worldwide. Contemporary Nigerian cities are characterized by inadequate and substandard housing, the proliferation of slums, insufficient infrastructure, transportation difficulties, low productivity, poverty, crime, and juvenile delinquency (Nnaemeka-Okeke, 2016; Bodo, 2019).

2.2.2 GIS and Remote sensing in mapping Urbanization patterns

Urban growth or expansion can be effectively analyzed using GIS. It is generally understood as a land-use transition from non-urban to urban categories. To assess urban growth patterns, GIS is essential for detecting changes between images and monitoring land-use transformations (Aburas et al., 2017).

Various indices, including mathematical and statistical measures, have been employed to quantify urban growth. These indices are integrated within a GIS framework alongside remote sensing data and techniques to analyze land-use changes both spatially and temporally (Aburas et al., 2017).

For instance, (Ramachandra et al., 2013) utilized the Landscape Shape Index (Landscape Shape Index (LSI)) and the Normalized Landscape Shape Index (Normalized Landscape Shape Index (NLSI)) to assess, measure, and compute urban growth patterns based on different analytical matrices.

Other studies have explored alternative methods, such as the one proposed by (Zha et al., 2003), who introduced the Normalized Difference Built-up Index (NDBI) to automate the mapping of built-up areas. He argued that, compared to the maximum likelihood classification method commonly used in LULC analysis, NDBI offers a faster and more objective approach for identifying built-up regions (Zha et al., 2003).

(Xu, 2007) enhanced the NDBI method by integrating it with other indices, demonstrating its effectiveness in analyzing urban expansion. Xu's study supports the application of NDBI both as a standalone method and in combination with other indices to monitor urban growth over time. The improved NDBI approach achieved a mapping accuracy of 92.6%, highlighting its reliability in fulfilling urban mapping objectives.

(Hadeel et al., 2011) analyzed the urban boundaries of Basrah Province counties between 1990 and 2003 using NDBI and found that the study area experienced rapid urban expansion over the 13-year period.

Similarly, (Elnaggar et al., 2020) conducted a classification using three indices—NDVI, NDBI, and Soil Adjusted Vegetation Index (SAVI)—and determined that NDBI provided the highest accuracy in detecting urban areas.

NDBI enhances the mapping of urban land cover by effectively highlighting built-up areas with greater accuracy and objectivity (Xu, 2007). Its values range from -1 to 1, where negative values indicate water bodies, positive values correspond to built-up and paved surfaces, and values near zero represent vegetated areas (He et al., 2010).

However, (Xu, 2007) pointed out a limitation of NDBI: it cannot differentiate between industrial, commercial, and residential zones within urbanized or developing areas, as these surfaces share similar waterproof characteristics and reflectance properties.

In change detection research, the vast and otherwise hidden information within multi-spectral images can be extracted using various image processing techniques, including spectral mixture analysis (Wu and Murray, 2003), machine learning-based classification (Maxwell et al., 2018), and artificial neural networks (Şenkal, 2010), among others. Spectral index-based approaches are among the most widely used techniques due to their ease of implementation and ability to produce reliable results (Bijeesh and Narasimhamurthy, 2019). The application of GIS in detecting and analyzing these changes follows a systematic approach, as outlined below.

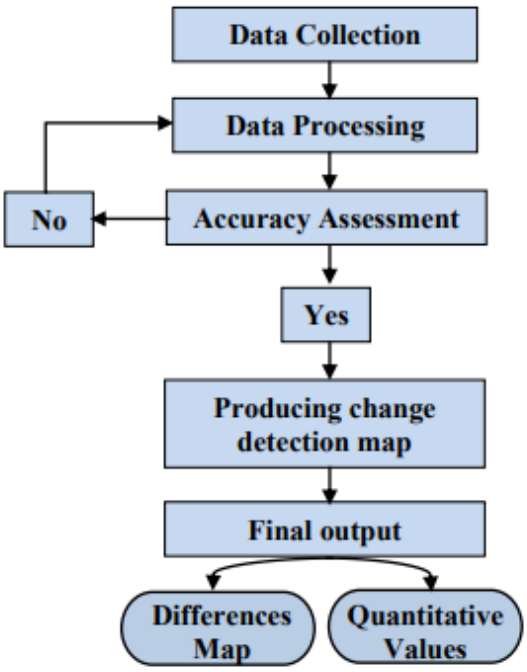


Figure 2.2: Framework of land use change detection using GIS

2.2.3 Data Mining and Predictive Modeling for Urban Growth

The final output of remote sensing products can be leveraged to address various societal challenges. For instance, predictive models based on change detection can provide insights into long-term phenomena such as urban expansion (Ngolo and Watanabe, 2023).

Selecting the right variables that explain urban expansion is crucial for model performance. Some studies highlight key factors, including: (i) Transport Network, (ii) Spatial Interactions in the form of neighborhood indices (e.g. the presence of industrial, commercial or non-urban land in surrounding areas),(iii) Policy Variables, typically the land-use zoning, are other significant factors as they determine regulations for future land use (iv) Topography(Linard et al., 2013).

Variable	Description	Resolution
ACCESS	Travel time to CBD (scaled and centered)	3 arc-seconds (100 m)
N150	Proportion of urban pixels in a 150 m buffer	3 arc-seconds (100 m)
N1000	Proportion of urban pixels in a 1 Km buffer	3 arc-seconds (100 m)
N5000	Proportion of urban pixels in a 5 Km buffer	3 arc-seconds (100 m)
DISTURB	Distance to the nearest urban pixel (meters)	3 arc-seconds (100 m)
SLOPE	Slope (%)	1 arc-second (30 m)

Table 2.2: Variables used in (Linard et al., 2013)

(Eyoh et al., 2012) classified three Landsat images of Lagos, Nigeria, from 1984, 2000, and 2005 using the K-means unsupervised algorithm. They modeled and predicted future urban expansion by employing logistic regression, considering 10 explanatory variables: Distance to Water,Distance to Residential Structures, Distance to Industrial and commercial centers, Distance to major roads,Distance to railway, Distance to Lagos Island,Distance to international airport, Distance to University of Lagos, Distance to University of Lagos State University, Distance to international seaport.

(Alsharif and Pradhan, 2014) selected Distance to Main Active Economic Centers, Distance to Central Business District, Distance to the nearest urbanized area , Distance to educational area, Distance to roads,Distance to urbanized areas , Easting Coordinates, Northing Coordinates, Slope, Restricted area, Population Density as possible drivers of urban sprawl and generated probability maps using data from 1984 to 2002 in the calibration phase and data from 2002 to 2010 in the validation phase resulting in an accuracy

rate of 0.86.

(Traore and Watanabe, 2017)utilized a Logistic Regression Model (LRM), GIS and Remote Sensing (RS) were used to analyze and quantify urban growth patterns and investigate their relationship with driving factors to simulate an urban growth probability map for Conakry. The results of the Logistic Regression Model (LRM) indicated that elevation and distance to major roads were the most significant variables, producing the best fit and highest regression coefficients.

Experiments have shown that Machine Learning (ML) algorithms outperform parametric techniques in remote sensing studies, with Random Forest (RF) and Support Vector Machine (SVM) achieving the highest average accuracy. However, implementation challenges often limit their widespread use (Ngolo and Watanabe, 2023).

(Linard et al., 2013) applied multiple urban expansion models using data from several large African cities and found that Boosted Regression Trees (BRT) outperformed distance-based urban growth models. They also observed that urban expansion in small, compact, and fast-growing cities was easier to model compared to large, sprawling cities.

2.3 Gaps and Opportunities

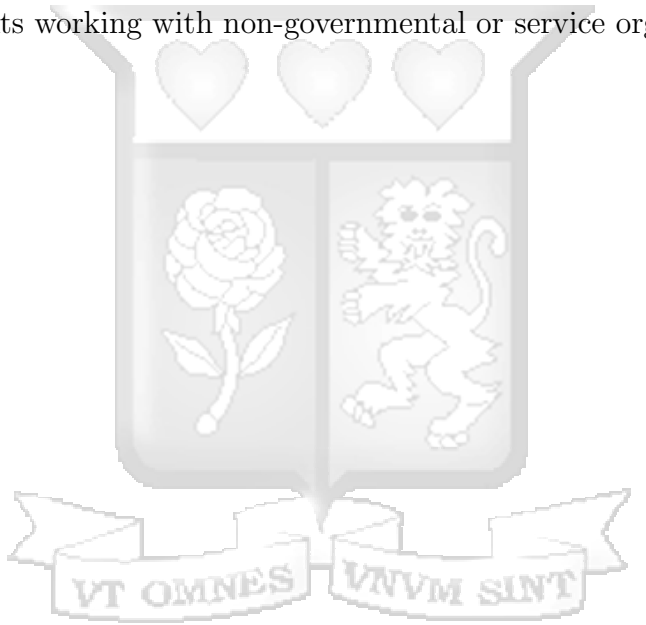
One of the main challenges in studying urbanization and city growth is defining and measuring what qualifies as "urban." The criteria used to classify an area as urban can vary across studies, leading to possible confusion and inconsistencies in urban growth analysis.

The study of land use and land cover (LULC) changes is important not only for land management but also for detecting environmental changes and developing sustainable strategies (Barnsley, 1997). However, traditional modeling approaches that rely only on demographic trends have limitations. Studies such as (Frimpong and Molkenthin, 2021) have struggled to capture recent urban growth patterns, often failing to consider how cities change over time and social preferences for different living environments (Kamusoko and Gamba, 2015).

To address these gaps, newer approaches integrate spatial data analysis, GIS, and remote sensing to provide a broader understanding of urban expansion and its influencing factors.

The limited availability of data remains a key challenge in applying GIS (Yeh, 1999). As an information system, GIS requires both graphic and textual data to operate effectively. However, obtaining reliable data on urban growth is difficult, and past projections for cities in developing regions, such as Mexico City and Lagos, have contained significant errors (Cohen, 2004). Due to these challenges, forecasting urban development involves a high level of uncertainty, often necessitating more advanced tools.

A recent review on urbanization research in Africa highlighted that many regions have lagged in both data quantity and quality since the 1970s. This data gap is linked to ongoing economic struggles and weakened institutional support for scientific research. A large portion of urban studies is conducted not through universities or government agencies but by consultants working with non-governmental or service organizations (Cohen, 2004).



Chapter 3: Methodology

This study applied Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology which is an industry-independent process model for data mining, consisting of six iterative phases from business understanding to deployment.(Schröer et al., 2021)

(a) **Business Understanding:** This phase established the project's objectives and business goals to ensure they align with the intended outcomes.

(b) **Data Understanding:** Data is collected and analyzed to gain comprehensive insights into its characteristics and relevance to the project.

(c) **Data Preparation:** The dataset is preprocessed through techniques such as feature engineering and selection to make it suitable for modeling.

(d) **Modeling:** Various machine learning models are applied to the processed data to develop predictive or descriptive models.

(e) **Evaluation:**The performance of the models is assessed to ensure they meet the project's goals and provide reliable results.

(f) **Deployment:** Once an optimal model is achieved, it is integrated into business processes to generate actionable insights and support decision-making.

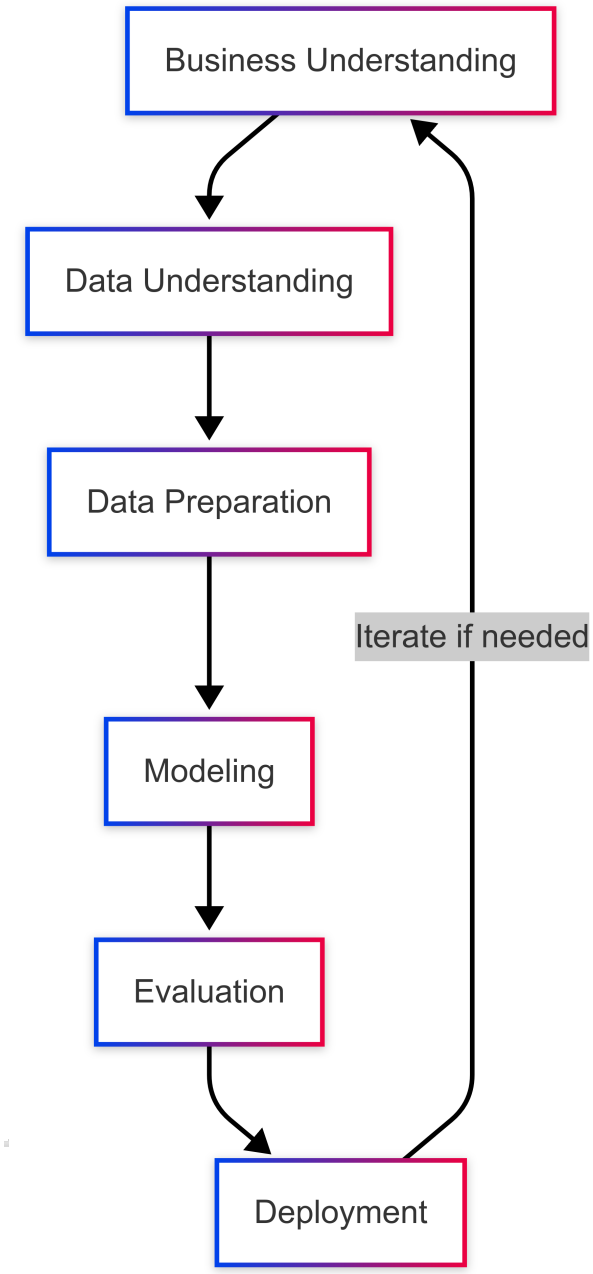


Figure 3.1: CRISP-DM framework

3.1 Business Understanding

(a) **Identifying Stakeholders:** Understanding the primary beneficiaries of this study was essential. A thorough review of the urbanization landscape in Nairobi helped identify relevant stakeholders, including urban planners, policymakers, and local government agencies. These groups stand to gain valuable insights from the predictive urbanization model, which is elaborated further in the UML diagram in Figure 4.1.

(b) **Defining Objectives:** To ensure a focused and structured approach, clear research objectives were established. These objectives served as the foundation for the study and are elaborated in Chapter 1, Section 1.3.

(c) **Determining Requirements:** Based on the research objectives, the key technical and functional requirements of the predictive urbanization model were defined. The system architecture and design were carefully structured to provide a clear framework for implementation. Preliminary visual representations, including low-fidelity wire frames and an Entity Relationship Diagram (ERD), were created to illustrate the user interface and database structure. These diagrams are provided in Chapter 4, Section4.2.2 and Figure 4.3 respectively.

3.2 Data Understanding

The project's study area, is Nairobi County and is the capital of the Republic of Kenya. It extends between 36°4' and 37°10'E and approximately between 1°9' and 1°28'S (Mundia and Aniya, 2007). With an estimated population of 5,454,000 in 2024.

Figure 3.2: Map of Nairobi County(Ngetich et al., 2016)

3.2.1 Data Collection

The primary data source for this research was satellite imagery from LANDSAT 8, obtained through Google Earth Engine. The chosen time frames, 2014 and 2024, were selected to capture a decade of urban expansion in Nairobi. LANDSAT 8 offers a resolution of 30m x 30m, with specific bands (e.g., Near-Infrared and Short-Wave Infrared) used to calculate indices such as NDBI, which is well-suited for detecting urbanization trends. In addition to LANDSAT data, secondary datasets including Nairobi's population growth and transportation infrastructure were incorporated to enhance the predictive power of the models.

3.2.2 Description of Dataset

(a) **Satellite Imagery (Primary Dataset)**

LANDSAT 8 satellite images were extracted from Google Earth Engine. The timeframe it was taken from is 2014 and 2024, chosen to analyze urban expansion over a decade. The resolution of the LANDSAT images were 30m × 30m and suitable for urbanization analysis.

The bands used are Near Infrared (NIR)and Short Wave Infrared (SWIR) for calculating the NDBI to detect built-up areas.

Steps were taken to transform images from raster (pixel-based) data to Comma Separated Values (CSV) format for machine learning.

(b) **Population Growth Data (Secondary Dataset)** Useful for understanding urbanization trends by linking population density to built-up expansion.

(c) **Distance to Airports (Secondary Dataset)** By linking distance from airport to built-up expansion, one can derive some insights

(d) **Distance to major Urban Centers (Secondary Dataset)** Helps understand urbanization trends by linking proximity to Urban centers to built-up expansion.

(e) **Distance to major roads (Secondary Dataset)** Helps understand urbanization trends by linking distance to major roads to built-up expansion.

(f) **Elevation (Secondary Dataset)** Helps understand urbanization trends by linking

distance to major roads to built-up expansion.

(g) **Slope (Secondary Dataset)** Helps understand urbanization trends by linking distance to major roads to built-up expansion.

3.3 Data Preprocessing

3.3.1 Data Preprocessing using GIS

The raw satellite images underwent atmospheric correction using QGIS to eliminate cloud cover and haze, improving the clarity of the data.

The images were then taken through final preprocessing operations such as layer stacking, image sub-setting, and image resampling before being subjected to analysis(Onyango and Opiyo, 2022).The images were then calibrated to surface reflectance based on the methods described by (Vermote et al., 2016) to ensure consistency when comparing indices over time and from different sensors(Guo et al., 2020).

(a) **Calculation of Spectral Indexes from Landsat 8 Data**

(i) Landsat-8 satellite imagery was uploaded into QGIS.

(ii) The Raster Calculator tool was used to compute NDBI for each time period.

(iii) The NDBI formula was applied in Raster Calculator as shown below:

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \quad \text{(Katyambo and Ngigi, 2017)} \qquad (2)$$

(b) **Reclassification**

(i) NDBI values range from -1 to 1.

(ii) According to (Kshetri, 2018),(Onyango and Opiyo, 2022):

Negative values indicate water bodies.

Higher values indicate built-up areas.

(iii) Reclassification of values using a threshold of 0 was done:

-1 to $0 \rightarrow$ "No Build-Up Areas"

0.1 to $1 \rightarrow$ "Build-Up Areas"

(iv) Classification was done using QGIS or ArcGIS Pro.

(c) **Change Detection**

(i) Classification of both 2014 and 2024 images was done.

(ii) The two images were overlayed and subtraction of corresponding pixel values in QGIS was done.

(iii) The change in NDBI was computed:

$$\text{NDBI Change} = \text{NDBI}_2 - \text{NDBI}_1 \tag{3}$$

(iv) The following categories were used to classify changes:

2014	2024	Change	Class
Developed	Developed	-	**-**
Developed	Undeveloped	-	**-**
Undeveloped	Undeveloped	No Change	**1**
Undeveloped	Developed	Changed	**2**

Table 3.1: Change Detection Table

(v) Pixels labeled as "1" and "2" were used for modeling.

(vi) A binary LULC map representing built-up changes was generated.

(d) **Selection of Variables** There is no standardized formula for selecting land-use drivers exists (Eyoh et al., 2012). The predictor variables are chosen based on studies in African cities. The following seven drivers were considered:

(i) Latitude and Longitude

(ii) Distance from major roads

(iii) Distance from urban centers

(iv) Distance from major airports

(v) Slope

(vi) Elevation

(e) **Slope and Elevation Calculation**

These two do not exist as shape files and have to be derived. To derive slope we used a Digital Elevation Model (DEM). Extraction of values was done directly from the DEM.

For slope, a Point Sampling Tool was used to extract slope and elevation values for each NDBI change detection point.

(f) **Final Data Preparation**

The Extracted data layers were combined into a single raster image. This was done to ensured that each raster cell contains all predictor variables.

The final image was saved in Geo-referenced Tagged Image File Format (Geo-Tiff) format. It was then converted and the data exported into a CSV format for further analysis.

3.3.2 Data Preprocessing using Python

(a) **Checking for Missing Values**.

A check was performed to identify any missing values in the dataset. The number of missing values in each column was recorded to assess data completeness

(b) **Dropping Unnecessary Columns**

"Distance_to_industrial_Zones(Metres)" was removed as it was not considered a key predictor. The *"id"* column was also dropped since it was only an identifier and did not contribute to the analysis. Additionally, *"Road_Name"* and *"Airport_Name"* were removed, as categorical location names are not useful in a numerical model.

(c) **Handling Missing Data**

The *"Slope"* column contained missing values, which were imputed using the median value to maintain consistency and avoid skewing the data.

Any remaining missing values in the *"persons per pixel 1km²"* column were removed since this was an important variable for urbanization analysis.

(d) **Reclassification of the Change Detection Variable**

The column *"change_detection_value"*, which indicated urbanization change, was originally labeled as:

(a) **1.0** → Areas that remained unchanged.

(b) **2.0** → Areas that transitioned to urban development.

This column was reclassified into a binary format for machine learning models:

1.0 was changed to **0.0** (representing "No Urbanization").

2.0 was changed to **1.0** (representing "Urbanized Area").

This transformation was necessary for models like XGBoost, which require binary classification labels.

(e) **Checking Data Shape After Cleaning**

The shape of the dataset (number of rows and columns) was checked after the cleaning process to confirm that the correct transformations had been applied.

3.3.3 Data Transformation

(a) **Feature Selection** Feature selection was performed to retain the most relevant variables for urbanization prediction. The following steps were taken:

(i) **Identifying Correlated Features**

(a) A correlation matrix was generated to examine the relationships between variables.

(b) Elevation and latitude (x-coordinate) were found to be highly correlated, with a correlation coefficient of 0.96.

(c) To avoid multicollinearity, a decision was made to retain only one of these variables.

(ii) **Measuring Feature Importance**

(a) A Random Forest model was trained to assess the importance of each feature in predicting urbanization.

(b) The top-ranked predictors were retained for modeling, while less significant features were considered for removal.

(iii) **Feature Engineering** Several ratio-based features were created to capture the relative influence of distance-based attributes:

(a) **Road-to-Airport Ratio:** Distance to roads divided by distance to airports.

(b) **Urban-to-Airport Ratio:** Distance to urban areas divided by distance to airports.

(c) **Road-to-Urban Ratio:** Distance to roads divided by distance to urban areas.

(d) A **proximity score** was computed as the sum of these ratios.

(e) An **urban access index** was generated using the ratio of population density to distance from urban areas.

(b) **Feature Scaling**

Since the dataset contained numerical variables with different ranges, Min-Max Scaling was applied to normalize the features. This ensured that all variables contributed equally to the model and prevented dominance by features with larger numerical scales.

Scaling Approach used was MinMaxScaler from `sklearn.preprocessing`. Feature scaling was applied only to the training data, while the test data was transformed using the same scaling parameters.

Class Imbalance Handling

(a) The dataset exhibited an imbalance in the urbanization class labels.

(b) Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset.

(c) Additionally, SMOTETomek was used as an alternative resampling strategy to reduce class imbalance while improving model generalization.

3.4 Model Selection and Training

There are different indicators to predict urban growth phenomena and no single model appears to perform consistently well when applied to different geographical locations(Ngolo and Watanabe, 2023). Because of this, different urban growth models were applied and their performances compared.

This section outlines the steps taken in training machine learning models for urbanization prediction in Nairobi. Three classification models were considered: Random Forest, Logistic Regression, and XGBoost. The models were trained, evaluated, and optimized through hyperparameter tuning.

3.4.1 Random Forest

The use of Random Forest classifier ensemble is based upon the basic premise that a set of classifiers do perform better classifications than an individual classifier does (Rodriguez-Galiano et al., 2012)

$$\hat{y} = \arg\max_{c} \sum_{t=1}^{T} 1(f_t(x) = c) \tag{4}$$

(Rodriguez-Galiano et al., 2012)

A Random Forest Classifier was selected as an initial baseline model.

The model was trained using a resampled dataset where Synthetic Minority Over-sampling Technique combined with Tomek Links (SMOTETomek) was applied to handle class imbalance. The model was trained on 80% of the data, and predictions were made on the 20% test set.

To optimize the model's performance, RandomizedSearchCV was used to tune hyperparameter such as Number of trees ($n_estimators$).

Predictions were evaluated using Precision, Recall, and F1-score. A Confusion matrix was used to analyze false positives and false negatives but the final model selection was based on recall to ensure minimal misclassifications of urbanized areas.

3.4.2 Logistic Regression

(Mahabir et al., 2018),(Ngolo and Watanabe, 2023) will be used to predict urbanization(Ngolo and Watanabe, 2023). According to (Ngolo and Watanabe, 2023),logistic regression might not give accurate predictions because predictor variables change over time and might create errors. As shown in Equation (5), the logistic regression model maps inputs to a probability between 0 and 1.

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{5}$$

(Kirasich et al., 2018)

A Logistic Regression model was trained to provide an interpretable comparison against the Random Forest model.The same SMOTETomek resampled dataset was used for training. The model was optimized to handle binary classification of urbanized vs. non-urbanized areas. Hyperparameter tuning was done to find the most optimal parameters for this model. The best-performing logistic regression model was selected based on Classification report metrics and F1-score for both urbanized and non-urbanized areas.

3.4.3 XGBoost

EXtreme Gradient Boosting or XGBoost is a library of gradient boosting algorithms optimized for modern data science problems and tools. Some of the major benefits of XGBoost are that it's highly scalable/parallelizable, quick to execute, and typically outperforms other algorithms and uses a more regularized model formalization, to control over-fitting, which gives it better performance(Sinha et al., 2020).

$$L_t = \sum_{i=1}^{n} [l(y_i, \hat{y}_{t-1,i} + f_t(x_i))] + \Omega(f_t) \tag{6}$$

(Ergul Aydin and Kamisli Ozturk, 2021)

The model was trained using the above mentioned resampled dataset to handle class imbalance. Initial training was performed using default XGBoost settings before hyper-parameter tuning.

RandomizedSearchCV was applied to optimize key parameters and the best XGBoost model was selected.

The model was evaluated against recall and F1-score for urbanization prediction. A classification report and confusion matrix were generated. Of all the model we have trained, it provided the highest Recall and F1 score which are the critical metrics for our task.

Key evaluation metrics used are:

a) Precision and Recall: Given the imbalanced nature of the dataset, recall was prioritized to minimize false negatives.

b) F1-score: A balance between precision and recall was used to assess overall model effectiveness.

c) Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) Score: The area under the receiver operating characteristic curve was computed to evaluate each model's ability to distinguish between urbanized and non-urbanized areas.

3.5 Model Deployment

3.5.1 Deployment Strategy

The deployment strategy was designed to facilitate real-time predictions while maintaining efficiency and scalability. The system was structured into two primary components:

(a) **Backend API**: Implemented using *FastAPI*, serving the machine learning model and handling requests from the frontend.

(b) **Frontend User Interface**: Built using *Streamlit*, providing a web-based platform for users to interact with the model.

(c) **Hosting and Deployment**: The FastAPI backend was deployed on *Render*, while the Streamlit UI was also hosted on Render for seamless access.

This approach allows for a modular design, ensuring that the backend and frontend components can be updated independently without disrupting the entire system.

The API was designed based on key principles to ensure efficiency and security:

(a) **Scalability**: The API supports multiple concurrent requests without significant latency.

(b) **Simplicity**: RESTful principles were followed to maintain an intuitive and well-structured API.

(c) **Security**: Basic authentication mechanisms were implemented to restrict unauthorized access.

(d) **Efficiency**: The model was optimized for fast inference, reducing response time.

These principles guided the development of a robust and reliable system for serving model predictions.

3.5.2 API Architecture

The API follows a RESTful architecture, allowing structured interaction between clients and the model. The key components include:

(a) **Request Handling**: The FastAPI framework processes incoming requests efficiently.

(b) **Model Prediction Pipeline**: The trained machine learning model receives input features, processes them, and returns urbanization predictions.

(c) **Response Format**: All responses are structured in JavaScript Object Notation (JSON) format to ensure compatibility with various frontend applications.

This architecture enables efficient communication between the UI and backend while ensuring scalability and maintainability.

3.5.3 Framework and Tools

Several frameworks and tools were used to develop and deploy the system:

(a) **FastAPI** – Used to build the backend API due to its speed and ease of implementation.

(b) **Streamlit** – Used for the frontend UI, providing an interactive platform for users to visualize predictions.

(c) **Render** – A cloud-based deployment platform used to host both the FastAPI backend and Streamlit UI.

(d) **Joblib** – Used to serialize and deserialize the trained XGBoost model for deployment.

These tools collectively ensure a robust, scalable, and user-friendly deployment setup.

3.5.4 Model Integration

The trained XGBoost classification model was integrated into the FastAPI backend. The integration process involved:

(a) **Loading the trained model**: The XGBoost model was serialized using Joblib and loaded into the API for inference.

(b) **Preprocessing input data**: Incoming requests were validated and transformed to match the model's expected input format.

(c) **Generating predictions**: The processed data was fed into the model, and the predicted urbanization status was returned.

(d) **Returning structured responses**: Predictions were sent back in JSON format, including confidence scores where applicable.

3.5.5 API Endpoints

The FastAPI backend exposes multiple endpoints to facilitate interaction with the model. The key endpoints include:

(a) `predict` – Accepts input features and returns urbanization predictions.

(b) `health` – A health check endpoint to monitor the status of the Application Programming Interface (API).

(c) `docs` – Auto-generated API documentation using Swagger UI, available by default in FastAPI.

These endpoints allow the front-end User Interface (UI) and external applications to query the model efficiently.

3.5.6 Deployment Environment

The system was deployed in a cloud environment to ensure availability and scalability. The deployment setup consists of:

(a) **FastAPI Backend**: Hosted on Render, allowing remote API access.

(b) **Streamlit UI**: Also deployed on Render, providing a user-friendly interface.

(c) **Model Storage**: The trained model is stored as a serialized file (`.joblib`) and loaded dynamically for predictions.

This cloud-based approach ensures that the model is accessible from anywhere, with minimal downtime and high availability.

The deployment of the urbanization prediction model was structured to provide an efficient, scalable, and user-friendly interface. The combination of FastAPI for the backend and Streamlit for the frontend, both hosted on Render, ensures a smooth experience for users interacting with the model. The architecture allows for easy updates, maintenance, and potential future expansions, such as integrating new models or additional security measures.

Chapter 4: System Design and Architecture

4.1 System Modeling

For the urbanization prediction system, a Hybrid System Model is recommended, combining Architectural Modeling (3-Tier Architecture) and Unified Modelling Language (UML) to ensure a structured, scalable, and efficient deployment.

The 3-Tier Architecture consists of a Presentation Layer (Streamlit UI) for user interaction, an Application Layer (FastAPI backend) for request handling and model inference, and a Data Layer (Machine Learning Model and Storage) for processing spatial data. This modular approach allows independent updates to different components, improving maintainability and scalability. Additionally, UML diagrams, such as Use Case and Data Flow Diagrams, visually represent system workflows, making it easier to design, debug, and enhance the system. By integrating FastAPI, Streamlit and the deployed XGBoost model, this architecture enables real-time urbanization predictions, enhances user accessibility, and ensures the system remains robust and adaptable for future improvements.

Figure 4.1 illustrates the UML diagram, showcasing the various components, relationships, and behaviors within the system.

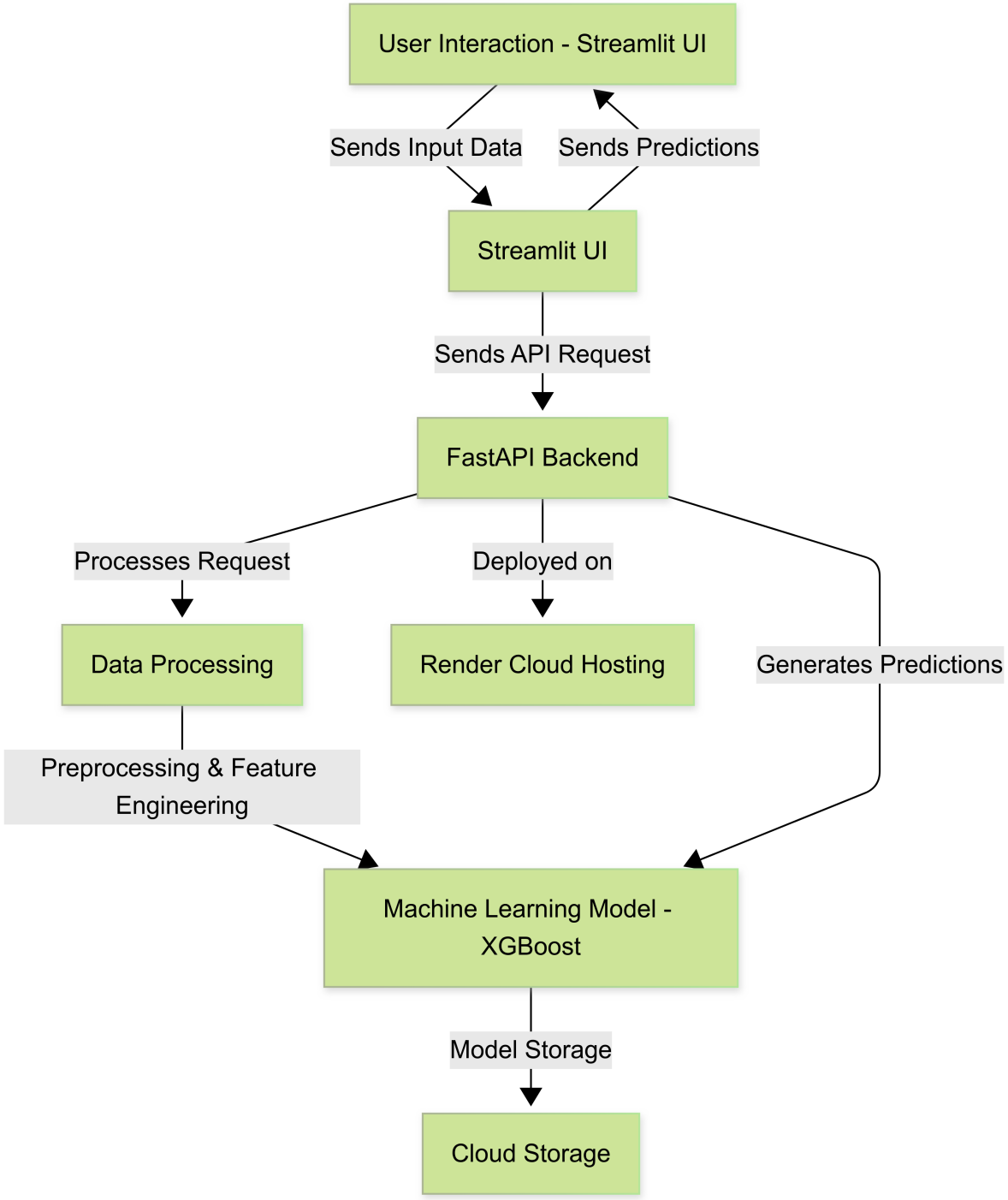


Figure 4.1: UML Diagram

4.2 System Components

During the study, the system developed comprised 5 components: User Interface (Streamlit UI), API Backend (FastAPI), Machine Learning Model (XGBoost), Data Processing & Preprocessing, Deployment & Cloud Infrastructure.

4.2.1 User Interface (Streamlit UI)

In this study, the UI was developed using Streamlit, a lightweight framework designed for building interactive web applications for data-driven tasks. The Streamlit UI facilitated a seamless user experience by allowing users to upload csv or excel datasets, request urbanization predictions, and visualize results dynamically on Aeronautical Reconnaissance Coverage Geographic Information System (ArcGIS).

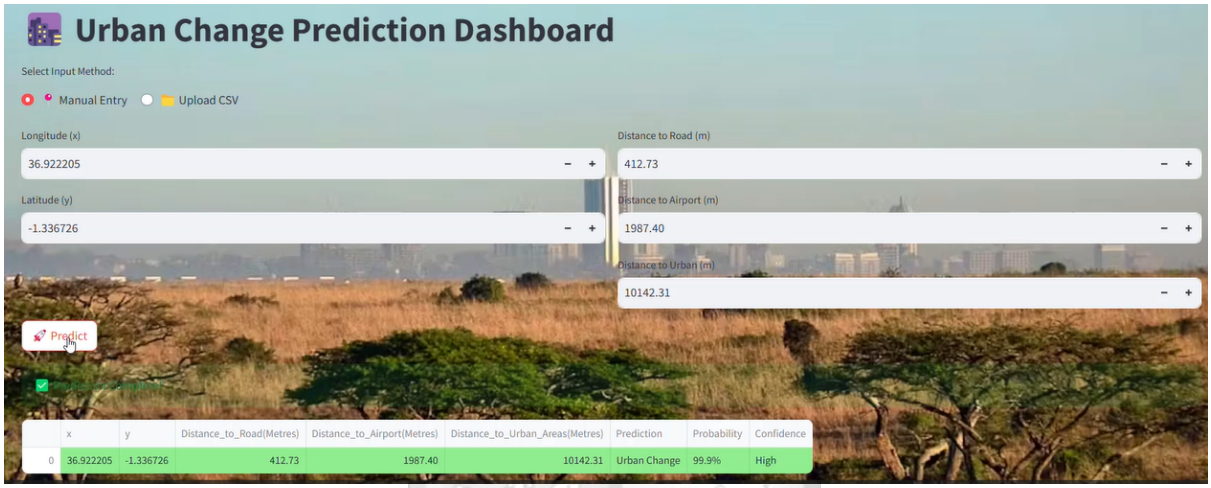


Figure 4.2: User Dashboard

4.2.2 API Backend

The API backend was developed using FastAPI, a modern, high-performance web framework for building APIs with Python. The FastAPI backend serves as the communication bridge between the user interface and the machine learning model, enabling seamless interaction between system components.

The backend is responsible for handling user requests, processing input data, and generating urbanization predictions.

An ERD, shown in Figure 4.3 visually depicted the relationships between these entities, outlining their attributes and connections within the system.

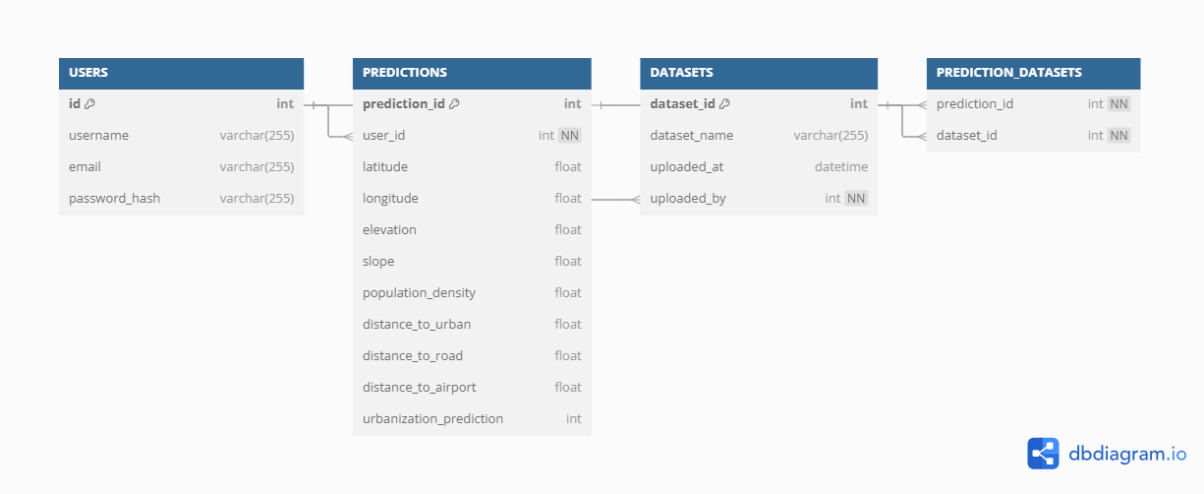


Figure 4.3: Entity relationship diagram

4.2.3 Machine Learning Model

The core prediction engine, trained on urbanization data, which receives input data and returns urbanization predictions.

These predictions were implemented using XGBoost. In deployment, the trained XGBoost model is stored as a serialized Joblib file, allowing efficient loading and inference. When a user submits geospatial data via the Streamlit UI, the data is preprocessed and sent to the FastAPI backend, where it is fed into the XGBoost model for classification. The model returns binary predictions (1 = urbanized, 0 = not urbanized) along with confidence scores, which are then formatted into a structured response and sent back to the UI for visualization.

4.2.4 Data Preprocessing

Handles preprocessing, feature engineering, and transformation of the input spatial data to align it with the model's expected structure.

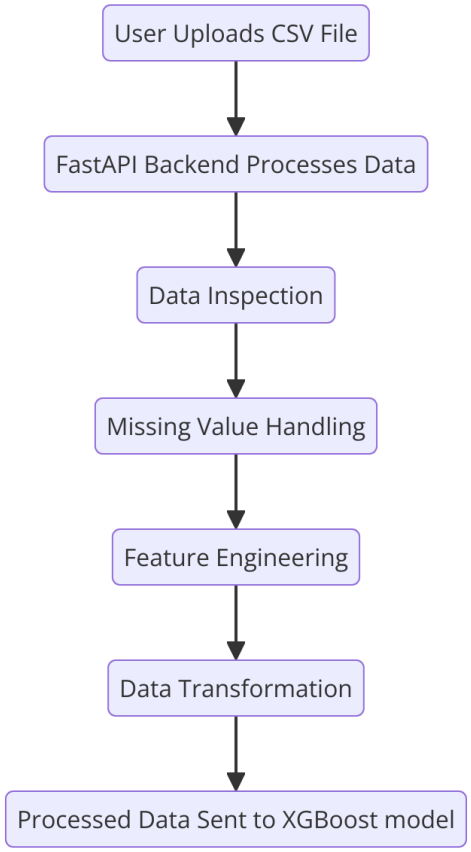


Figure 4.4: Data Preprocessing

Feature Engineering ensures the input data matches the features used in training, Scaling adjusting feature values for consistency in model input, spatial Data Processing handles geospatial attributes like coordinates and elevation to align with urbanization patterns, format Standardization ensuring the data format (CSV/Excel) is compatible with the API and ML model, quality Assurance removes duplicates, outliers, or irrelevant data to improve model performance.

This step is crucial for accurate urbanization predictions and ensures smooth integration with the FastAPI backend and XGBoost model.

4.2.5 Deployment & Hosting (Render)

In this study, the deployment and hosting architecture was designed to ensure seamless interaction between the machine learning model, API backend, and user interface while maintaining scalability and efficiency. The deployment pipeline integrates cloud hosting, containerization, and model serving techniques to enable real-time predictions and visualization of urbanization patterns.

Through this deployment strategy, the system enables seamless integration of machine learning predictions into real-world applications, allowing policymakers and urban planners to analyze and visualize urbanization trends effectively.

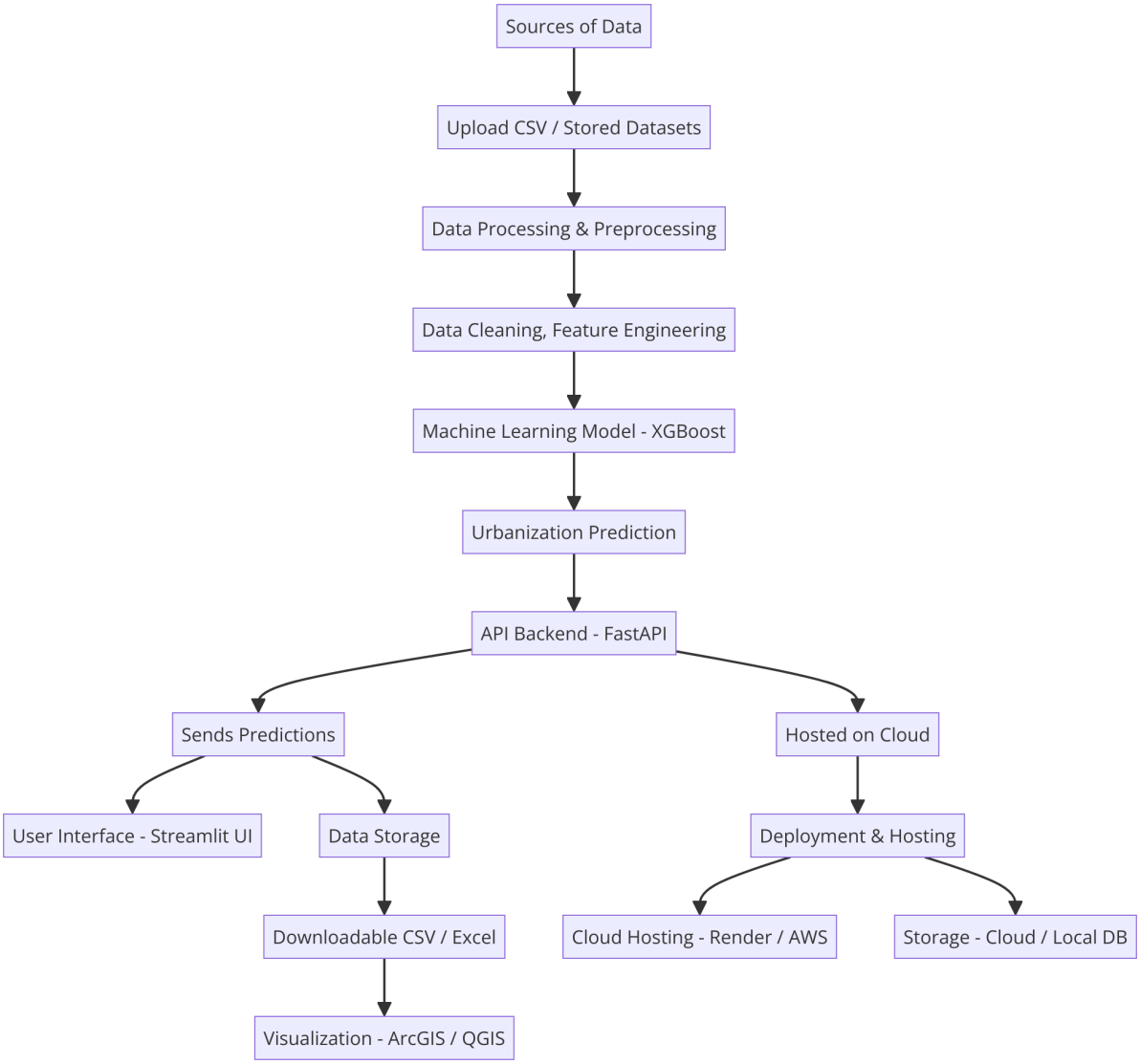


Figure 4.5: System Architecture

Chapter 5: System Implementation and Testing

This chapter provides an in-depth exploration of the development process and technical aspects of the urbanization prediction system. It delves into the system architecture, design decisions, and implementation details of the data processing pipeline, machine learning model, and web-based interface. By examining the methodologies, tools, and technologies employed, this chapter offers insights into the systematic approach taken to process spatial data, generate urbanization predictions, and visualize the results effectively.

Through an analysis of the system's components and implementation strategies, a deeper understanding is gained of how the system was conceptualized, designed, and deployed to provide accurate and actionable insights for urban planning. Additionally, this chapter will discuss testing methodologies used to evaluate the system's accuracy, performance, and usability, ensuring that it effectively addresses the challenge of predicting urbanization trends and supports decision-making in infrastructure planning.

5.1 System Implementation

5.1.1 Data Processing and Preprocessing

Effective data processing and preprocessing are crucial for ensuring high-quality inputs into the machine learning model. In this study, the dataset comprises spatial data, including latitude, longitude, elevation, slope, distance from roads, distance from urban centers, distance from airports and population density, which are essential for predicting urbanization patterns. The preprocessing workflow involves multiple stages, from data ingestion to transformation, ensuring the dataset is optimized for model training and inference.

The system accepts spatial datasets in CSV format, containing geospatial attributes and features relevant to urbanization prediction. The uploaded dataset must match the structure of the training data to maintain consistency and accuracy in predictions. The user uploads the file through the UI, after which it is processed and validated to ensure compliance with the expected schema.

Upon ingestion, the dataset undergoes several preprocessing steps:

Feature	Description
Latitude and Longitude	Spatial coordinates that provide locational context.
Elevation	The height above sea level, which influences settlement patterns.
Slope	The degree of steepness, affecting construction feasibility.
Population Density	A critical indicator of urbanization intensity.
Distance to Urban Areas	Measures proximity to existing urban centers, influencing urban expansion.
Distance to Roads	Indicates accessibility and potential infrastructure development.
Distance to Airports	Represents connectivity and the potential for economic growth.

Table 5.1: Geospatial Features Used in Urbanization Prediction

a) **Handling Missing Values**: Missing geographic attributes such as elevation or population density are imputed using spatial interpolation techniques or median imputation.

b) **Feature Scaling**: Continuous variables such as elevation and slope are normalized using Min-Max Scaling to ensure uniform value ranges across different attributes.

c) **Outlier Detection**: Anomalous values, such as extreme elevation points or unusually high population densities, are identified and treated using inter quartile range (IQR) filtering.

d) **Correlation Analysis**: Since latitude and elevation were found to be highly correlated, feature selection techniques are applied to reduce redundancy.

5.1.2 Machine Learning Model (XGBoost)

XGBoost (Extreme Gradient Boosting) was selected as the predictive model for urbanization due to the following reasons:

Efficient and Scalable as it is optimized for speed and performance.

Handling of Imbalanced Data, since the target class is highly imbalanced, where there are more non-urbanized areas than urbanized ones.

Feature Selection and Engineering.Correlation analysis was performed to ensure that highly correlated features, such as latitude and elevation, do not introduce redundancy.

Additionally, categorical encoding techniques were applied where necessary.

Model Training and Hyperparameter Optimization. The Extreme Gradient Boosting (XGBoost) model was trained using a random search and cross-validation approach to optimize its hyper parameters. The key hyperparameters tuned include:

Hyperparameter	**Description**
Learning Rate (η)	Controls the step size during optimization to balance convergence speed and model accuracy.
Maximum Depth (*max_depth*)	Defines the complexity of each decision tree in the ensemble.
Subsample Ratio (*subsample*)	Determines the fraction of data points used per boosting iteration to prevent overfitting.
Regularization Parameters (λ, α)	Applied to avoid overfitting by penalizing large coefficients.

Table 5.2: XGBoost Hyperparameters and their Descriptions

The model was evaluated using a stratified k-fold cross-validation approach, ensuring that training and validation data distributions remained balanced.

Prediction and Output Generation. Once trained, the model receives new spatial data as input and outputs a binary classification:

a.) **0 – Not Urbanized:** Areas predicted to remain non-urban.

b.) **1 – Urbanized:** Locations projected to experience urbanization.

The results are stored in a downloadable CSV file, enabling further spatial visualization in ArcGIS or Quantum Geographic Information System (QGIS). This allows stakeholders to overlay predictions on existing maps for better urban planning decisions.

Implementation Tools. The model is implemented using the following technologies:

a.) **XGBoost Library:** For efficient gradient boosting tree-based learning.

b.) **Scikit-learn:** Used for preprocessing, hyperparameter tuning, and model evaluation.

c.) **NumPy and Pandas:** For handling numerical computations and dataset manipulation.

d.) **FastAPI Backend:** Facilitates communication between the model and the user interface.

5.1.3 Deployment

The deployment of the system involved hosting both the machine learning model and the UI to ensure accessibility and usability. The deployment strategy focused on scalability, accessibility, and automation, enabling users to interact with the urbanization prediction model seamlessly. Model Loading: The trained XGBoost model was saved and loaded onto streamlit.

The overall deployment process was structured as follows:

a.) Deploy the Streamlit UI, connecting it to the backend.

b.) Users interact with the system, uploading CSVs and receiving urbanization predictions as csv file.

5.1.4 User Interface (Streamlit UI)

The user interface, built using Streamlit, was designed to allow users to interact with the model via a web-based platform. Deployment involved:

Direct API Communication: The Streamlit frontend was connected to the FastAPI backend, sending user-uploaded CSV data for processing. Real-Time Predictions: The API processed the data and returned predictions on whether specific locations were urbanized or not. Visualization Support: Users could view prediction results and download them for further analysis in GIS platforms like ArcGIS or QGIS.

5.2 Testing

Testing is a critical phase in the development of the urbanization prediction system, ensuring that the deployed application functions as expected, provides accurate predictions, and delivers a seamless user experience. Since the system relies on spatial data, machine learning models, and an interactive Streamlit interface, rigorous testing is essential to validate its functionality, usability, performance, compatibility, security, and effectiveness in predicting urban expansion. Functionality, usability, compatibility, security, and effectiveness were tested to ensure the system actually addresses the identified problem statement.

5.2.1 Functionality Testing

Functionality testing was done to ensure that the system's core features work as expected.

The Streamlit front-end was evaluated to verify that users could seamlessly upload datasets, visualize predicted urbanization areas, and compare results with actual urban expansion patterns. The model underwent testing to evaluate its prediction accuracy. It was tested using a validation dataset to check whether it correctly classified areas as urbanized or non-urbanized based on historical data. The performance was assessed using metrics such as recall, given the project's priority on minimizing missed urbanization cases for better infrastructure planning.

To validate the real-world applicability of the predictions, the generated urbanization maps were exported and imported into ArcGIS. The predicted urbanized areas were then overlaid on a 2024 satellite map to visually compare model predictions against actual urban expansion. This overlay analysis helped assess whether the model correctly identified regions that had undergone urbanization or remained unchanged.

5.2.2 Usability Testing

Usability testing was conducted to assess the user-friendliness and accessibility of the Streamlit-based urbanization prediction system. The primary goal was to ensure that users could easily navigate the interface, upload datasets, run predictions, and visualize results without technical difficulties. Test users, including individuals familiar with urban planning and GIS analysis, were asked to interact with the system and provide

feedback on its intuitiveness. The layout, button placements, file upload process, and result visualization were evaluated to ensure a seamless user experience. Additionally, the responsiveness of the UI was tested across different devices and screen sizes to confirm that users could access the system efficiently from both desktops and laptops. Based on user feedback, minor UI refinements were made to improve clarity, including adjusting labels and refining instructions to guide users through the workflow. This testing phase ensured that the system was accessible to both technical and non-technical users, making it more effective for urbanization analysis.

5.2.3 Compatibility Testing

Compatibility testing was conducted to ensure that the Streamlit-based urbanization prediction system functioned seamlessly across different platforms, web browsers, and devices. The application was tested on multiple web browsers, including Google Chrome, Mozilla Firefox, Microsoft Edge, and Safari, to verify that all interface components rendered correctly and that there were no inconsistencies in layout or functionality.

Additionally, mobile responsiveness was evaluated by accessing the system on various screen sizes, including desktops, laptops, and tablets, to check for adaptability and usability. While Streamlit is primarily optimized for desktop use, testing ensured that users could still access results on mobile devices when necessary. Furthermore, different operating systems, including Windows, macOS, and Linux, were used to verify system stability and performance consistency. Any UI misalignment or slow-loading issues were documented and optimized to improve accessibility, ensuring that the system could be effectively utilized across different environments.

5.2.4 Security Testing

Security testing focused on protecting user data, ensuring safe file uploads and downloads, and preventing unauthorized access to the Streamlit application. Since the system allows users to upload CSV files, testing was conducted to prevent malicious file injections by validating file formats and restricting unsupported file types.

Furthermore, since users can download prediction results, testing ensured that exported files were correctly formatted, free from corruption, and did not contain any unintended

security vulnerabilities. The system was tested to confirm that all downloaded CSV or raster files contained only the intended prediction data without exposing sensitive internal processes.

Chapter 6: Discussion of Results

This chapter reviews the results obtained at each stage of the CRISP-DM framework, from data understanding to model deployment. It presents key insights gained from data exploration and analysis, emphasizing the patterns and trends identified during data preprocessing and modeling. Additionally, it evaluates the performance metrics and overall effectiveness of the developed models in meeting the research objectives outlined in Chapter 1 section 1.3.

6.1 Data Understanding

6.1.1 Data Extraction

The extracted dataset comprises key spatial features relevant to urban prediction, including latitude, longitude, elevation, slope, distance to road, distance to urban area, distance to airport and population density. These features were selected based on their influence on urban growth patterns, where elevation and slope determine land suitability for development, and population density serves as a proxy for human settlement concentration. The dataset spans Nairobi County as illustrated in Figure 6.1

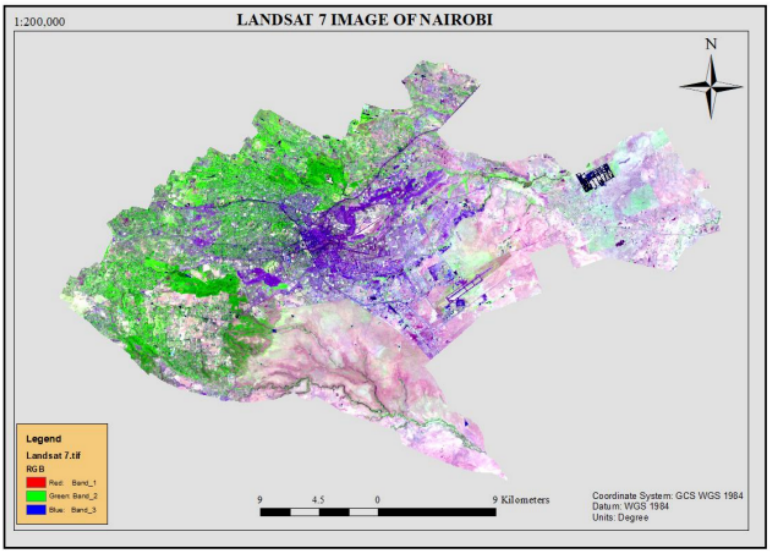


Figure 6.1: Landsat Image Nairobi 2014

The LANDSAT images extracted when analyzed showcase areas undergoing transformation. Since Google Earth Engine (GEE) provides multi-temporal satellite-derived data, the extracted features reflect land surface conditions over time, allowing for an analysis

of urban expansion dynamics. Given the reliance on satellite imagery and computational geospatial analytics, the dataset was structured to ensure compatibility with machine learning techniques.

6.1.2 Urban Expansion Trend

Urbanization in Nairobi exhibits a distinct spatial pattern, with significant growth occurring along major transportation corridors. As illustrated in Figure 6.2, the urbanized clusters (green points) are predominantly aligned with key road networks, including the A8 and A8/E highways.

This alignment suggests that infrastructure development plays a crucial role in shaping urban expansion, as proximity to well-established roads and economic hubs facilitates accessibility and attracts settlement growth.

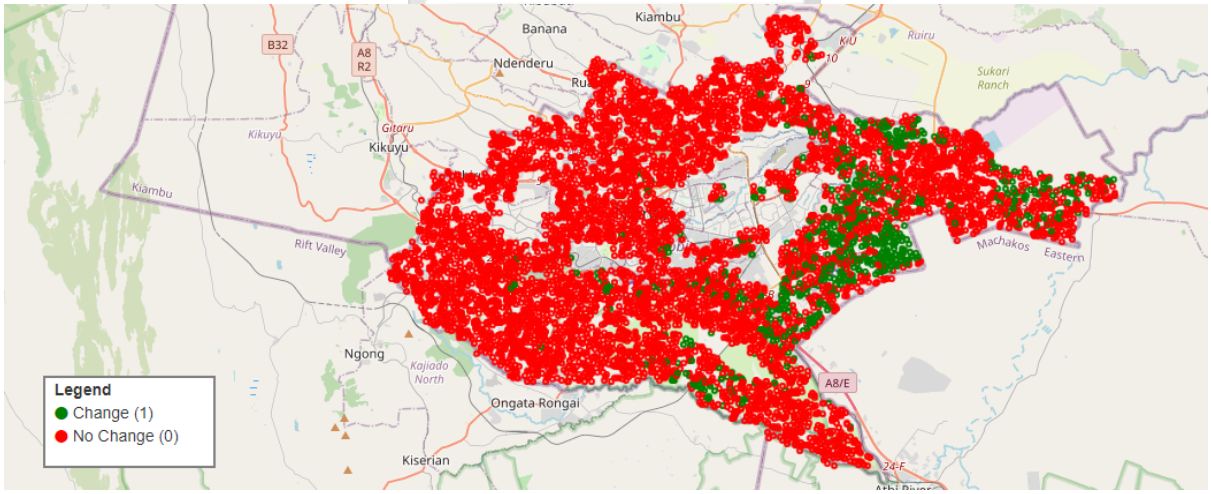


Figure 6.2: Change detection analysis of Nairobi,green points indicate urbanized clusters and red points represent areas with no change.

Furthermore, a notable trend of peripheral expansion is observed, where new urban clusters are forming on the outskirts of Nairobi rather than in the central urban core. This pattern, evident from the distribution of green points in Figure 6.2, reflects the ongoing process of suburbanization. Increasing population pressure, rising housing demand, and economic activities relocating beyond the city center contribute to this outward expansion.

The growth of peri-urban areas presents both opportunities and challenges, as it underscores the need for strategic urban planning to ensure sustainable infrastructure develop-

ment and service provision in these rapidly urbanizing zones.

6.2 Data Preparation

6.2.1 Class Imbalance

Upon examining the dataset, it was observed that the target variable exhibited a significant class imbalance. Initially, the dataset contained 5,097 instances labeled as 0 (non-urban) which represents 88.25% and only 679 instances labeled as 1 (urban) which represents 11.75%, highlighting a considerable disparity in class distribution as seen in the change_detection_value feature of my dataset in Figure 6.3

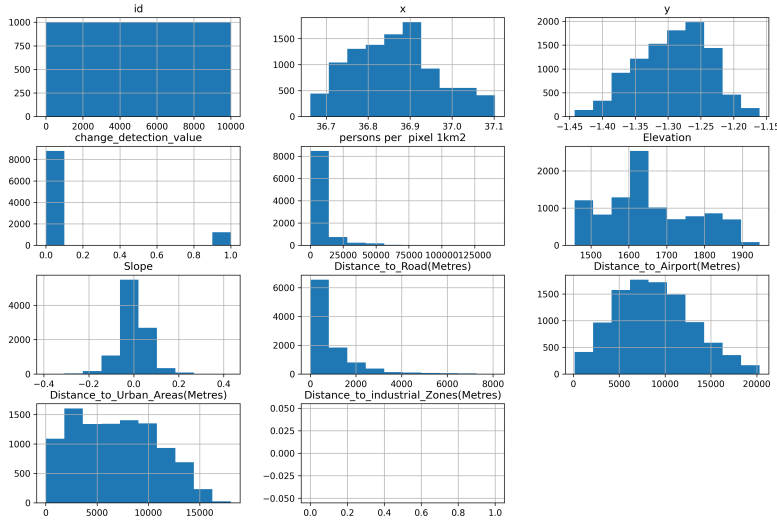


Figure 6.3: Data Distribution

To address this imbalance, a resampling technique was applied:

SMOTETomek is a hybrid resampling technique that combines Synthetic Minority Oversampling Technique and Tomek Links. It works by generating synthetic samples for the minority class done by Synthetic Minority Oversampling Technique (SMOTE) and Tomek Links identify borderline examples between the majority and minority classes and majority class instance in the Tomek Link is removed to create a cleaner decision boundary.

This results in a balanced dataset as illustrated in **??** with reduced noise and a better-defined decision boundary. The result is a more balanced dataset with 4,986 instances for each class. This step was crucial in ensuring that the predictive model would not be biased toward the majority class.

6.2.2 Feature Analysis

(a) Analyzing the spatial distribution of elevation across the study area revealed an inverse relationship between elevation and proximity to urban infrastructure. Lower-elevation areas were predominantly located in highly urbanized regions, while higher elevations were more characteristic of peripheral or less-developed areas. This pattern is consistent with urban expansion trends, where cities tend to develop on flatter, lower-elevation terrain due to ease of construction, accessibility, and infrastructural development.

(b) The dataset displayed a sparse representation at the extreme ends of the elevation spectrum, with relatively fewer data points at very low or very high elevations. This suggests that most of the study area falls within a mid-range elevation band, potentially aligning with Nairobi's central landforms. The low representation of extreme elevations may indicate geographical constraints on urbanization, where steep terrains or valley regions are less likely to be urbanized due to physical barriers.

(c) Further examination of the dataset revealed a few high-elevation points that were also located relatively close to urban centers. These observations may represent outliers or specialized developments, such as housing estates built on hills or planned infrastructure expansions in elevated areas. These cases warrant further investigation, as they could introduce anomalies in the urbanization prediction model.

6.2.3 Feature Correlation

A correlation analysis was performed on the dataset to examine the relationships between various numerical features. The heatmap in Figure 6.4 illustrates these correlations. Notably, elevation exhibits a strong negative correlation with the x coordinate (-0.96), suggesting that lower elevations are associated with certain spatial locations. Additionally, distance to roads (0.47) and distance to airports (0.23) show moderate positive correlations with the x coordinate, indicating that these distances increase in a specific direction.

The target variable, *change_detection_value*, shows a weak correlation with most features, with the highest being a moderate correlation with the x coordinate (0.31). This suggests that spatial positioning may influence urbanization patterns but is not the sole

determinant.

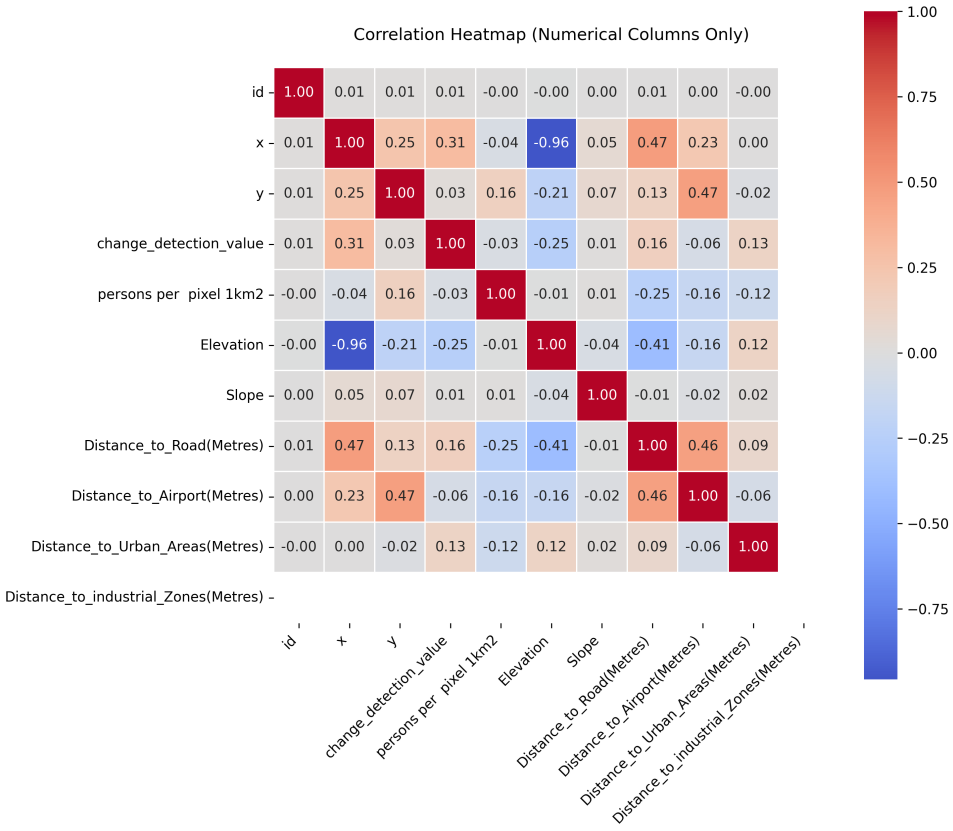


Figure 6.4: Correlation Heatmap of Numerical Features

Furthermore, population density (persons per pixel) has a weak negative correlation with elevation (-0.25) and distance to roads (-0.25), implying that denser population areas tend to be in lower-lying regions and closer to road networks. Overall, the correlation matrix suggests spatial dependencies and possible interactions between geographic features that influence urbanization trends.

6.2.4 Feature Importance and selection

Feature importance was analyzed using a Random Forest model to identify the most influential variables in predicting urbanization patterns. Table 6.1 presents the ranked importance scores for each feature. The results indicate that the spatial coordinates (x and y) are the most critical predictors, with x having the highest importance score (0.2046), followed by the distance to the nearest airport (0.1652). The distance to urban areas (0.0955) and road networks (0.0891) also demonstrate notable contributions, reinforcing the significance of proximity to infrastructure in determining urbanization trends.

Interestingly, elevation and slope—although traditionally considered relevant in geographic analyses—exhibited relatively lower importance scores (0.0811 and 0.0795, respectively). Population density (*persons per pixel 1km*^{2}) also showed a moderate contribution (0.0925). However, categorical features such as *Airport Name* and *Distance to Industrial Zones* had negligible impact on the model.

Based on these findings, a feature selection process was conducted to enhance model efficiency by removing less impactful variables. The dropped features included *Elevation*, *Slope*, *persons per pixel 1km*^{2}, and *Airport Name*. These removals aimed to reduce redundancy and improve computational efficiency without significantly affecting predictive performance.

Rank	Feature	Importance
1	x	0.2046
2	Distance to Airport (Metres)	0.1652
3	y	0.1012
4	Distance to Urban Areas (Metres)	0.0955
5	Persons per pixel 1km^2	0.0925
6	id	0.0904
7	Distance to Road (Metres)	0.0891
8	Elevation	0.0811
9	Slope	0.0795
10	Airport Name	0.0009
11	Distance to Industrial Zones (Metres)	0.0000

Table 6.1: Feature Importance Ranking from Random Forest

The final features retained key spatial and infrastructural factors aligning with the expectation that urbanization is closely linked to accessibility and location.

6.3 Modeling

6.3.1 Random Forest Classifier

The learning curve of the Random Forest model demonstrates distinct patterns in training and test accuracy as the number of estimators increases. Initially, the training accuracy is below 1.0 when the number of trees is small. However, as the number of estimators increases, the training accuracy rapidly reaches 100% and remains at this level. This behavior strongly indicates overfitting, as the model memorizes the training data instead of generalizing well.

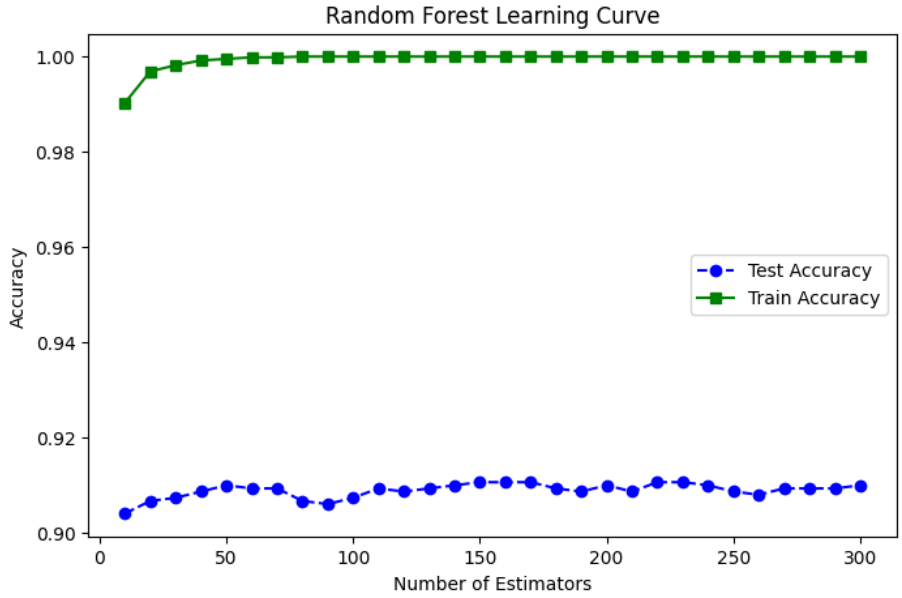


Figure 6.5: Random Forest Learning Curve: Number of Estimators vs. Accuracy

Conversely, the test accuracy begins around 0.90 and exhibits minor fluctuations as the number of estimators increases. Despite adding more trees, the test accuracy remains relatively stable between 0.90 and 0.92, suggesting that increasing the number of estimators does not significantly enhance model performance on unseen data.

6.3.2 Logistic Regression

The learning curve for Logistic Regression demonstrates the impact of regularization strength (C) on model performance. As C increases, the model's regularization effect decreases, allowing it to fit the training data more flexibly. Initially, at very low values of C (strong regularization), both training and test accuracy are relatively low. However, as C increases, accuracy improves for both sets, indicating that the model benefits from reduced regularization constraints.

Beyond a certain point (approximately $C = 1$), both training and test accuracy plateau, suggesting diminishing returns from further reductions in regularization. The fact that training and test accuracy converge at higher values of C indicates that the model is reaching its optimal fit without significant overfitting.

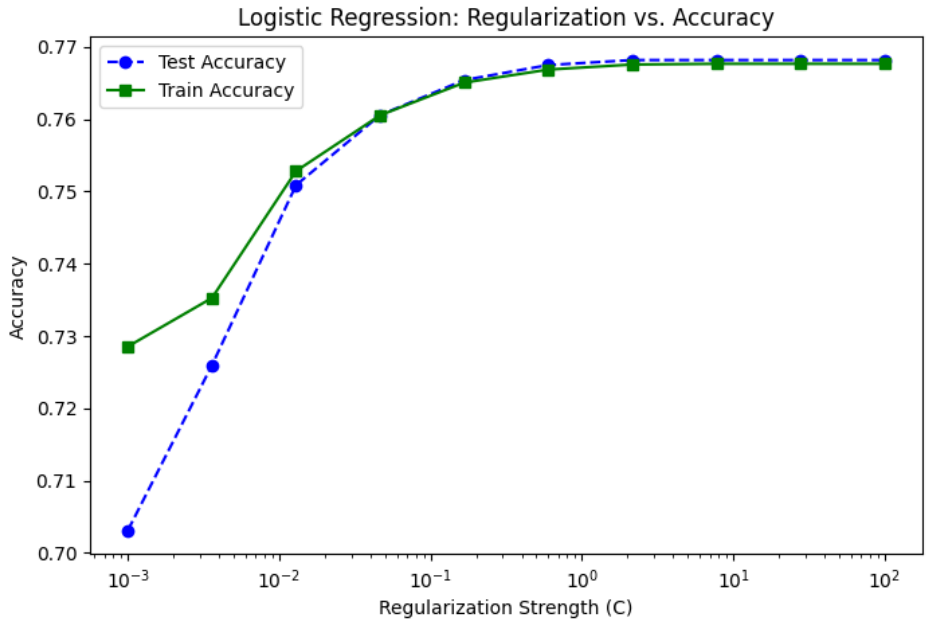


Figure 6.6: Logistic Regression Learning Curve: Regularization Strength vs. Accuracy

6.3.3 XGBoost

The XGBoost learning curve, illustrated in Figure 6.7, shows the relationship between the number of estimators and classification accuracy. The training accuracy (green line) starts near 0.99 and quickly reaches 1.0, indicating that the model perfectly fits the training data even with a small number of trees. This suggests strong memorization and a high risk of overfitting. Meanwhile, the test accuracy (blue line) remains relatively stable between 0.89 and 0.91, with slight fluctuations as the number of estimators increases. This behavior implies that adding more trees does not significantly improve generalization performance. Instead, the model's performance plateaus, reinforcing the idea that additional trees mainly contribute to overfitting rather than enhancing predictive power. Overall, the results indicate that choosing an optimal number of estimators is crucial for balancing bias and variance, as excessive boosting rounds lead to diminishing returns in test accuracy.

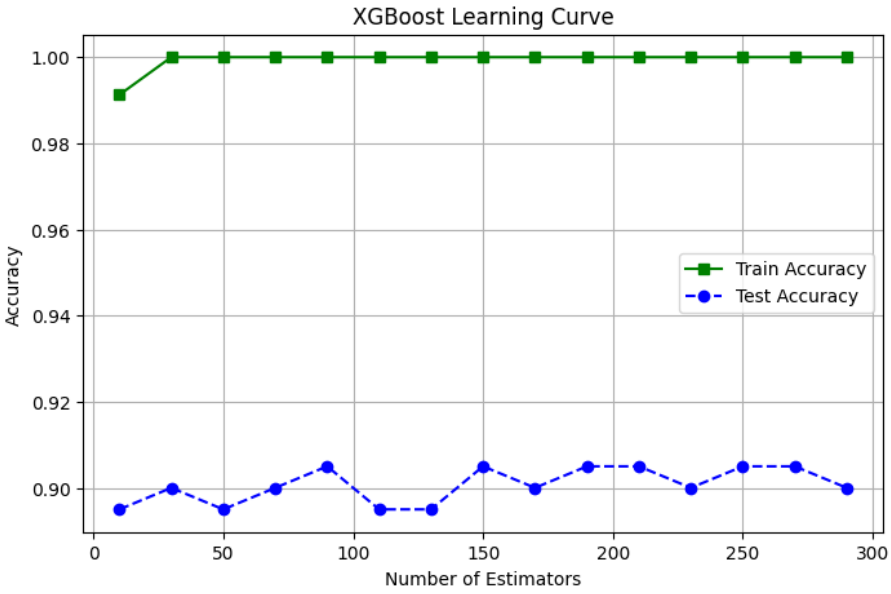


Figure 6.7: XGBoost Learning Curve: Training vs. Test Accuracy

6.4 Model Evaluation and Optimization

6.4.1 Accuracy

Accuracy is a fundamental metric that measures the proportion of correctly classified instances. From the results illustrated by Table 6.2 the Random Forest (Optimized) and XGBoost models achieve the highest accuracy at 0.88, indicating their strong performance in classification. The baseline Random Forest model follows closely with an accuracy of 0.85, suggesting that even without optimization, it remains a reliable model. In contrast, Logistic Regression exhibits the lowest accuracy at 0.77, implying that it may struggle to capture complex patterns in the dataset compared to ensemble methods. The improvement from the baseline to the optimized Random Forest highlights the positive impact of hyperparameter tuning on model performance. Overall, ensemble methods such as Random Forest and XGBoost demonstrate superior accuracy, making them strong candidates for the final model selection.

Metric	**RF(Base)**	**RF(Optimized)**	**LR**	**XGBoost**
Accuracy	0.85	0.88	0.77	0.88
Precision	0.42	0.45	0.28	0.45
Recall	0.63	0.65	0.74	0.65
F1 Score	0.50	0.53	0.41	0.53

Table 6.2: Model Performance Evaluation

6.4.2 Recall

Recall measures the model's ability to correctly identify positive instances, making it particularly crucial in scenarios where missing positive cases has significant consequences. Among the evaluated models, Logistic Regression achieves the highest recall at 0.74, suggesting that it is most effective at capturing positive instances, albeit at the cost of lower precision. The optimized Random Forest and XGBoost models both attain a recall of 0.65, showing a balanced ability to detect positive cases while maintaining relatively high accuracy.

The baseline Random Forest model has the lowest recall at 0.63, indicating a slight improvement through hyperparameter optimization. Given the high recall of Logistic

Regression, it may be preferable in cases where capturing as many positive instances as possible is the priority. However, trade-offs with other metrics, such as precision and overall accuracy, must be considered for final model selection.

6.4.3 Precision

When examining the precision scores across the four models, it is evident that there is a noticeable variation in their ability to correctly identify positive instances. The Random Forest (Baseline) model achieved a precision of 0.42, which slightly improved to 0.45 after hyperparameter optimization. This modest improvement suggests that while the optimization procedure contributed to better precision, it did not drastically enhance the model's ability to reduce false positives. Similarly, the XGBoost model attained the same precision score of 0.45, aligning with the optimized Random Forest, indicating comparable capabilities between these two ensemble methods in terms of precision. On the other hand, Logistic Regression lagged behind, recording the lowest precision value of 0.28. This result highlights the model's tendency to generate a higher number of false positives, making it less reliable when the objective is to correctly predict the positive class. Collectively, the results imply that ensemble-based models (Random Forest and XGBoost) outperform the simpler Logistic Regression model in precision, making them more suitable when minimizing false positive predictions is essential for the application at hand.

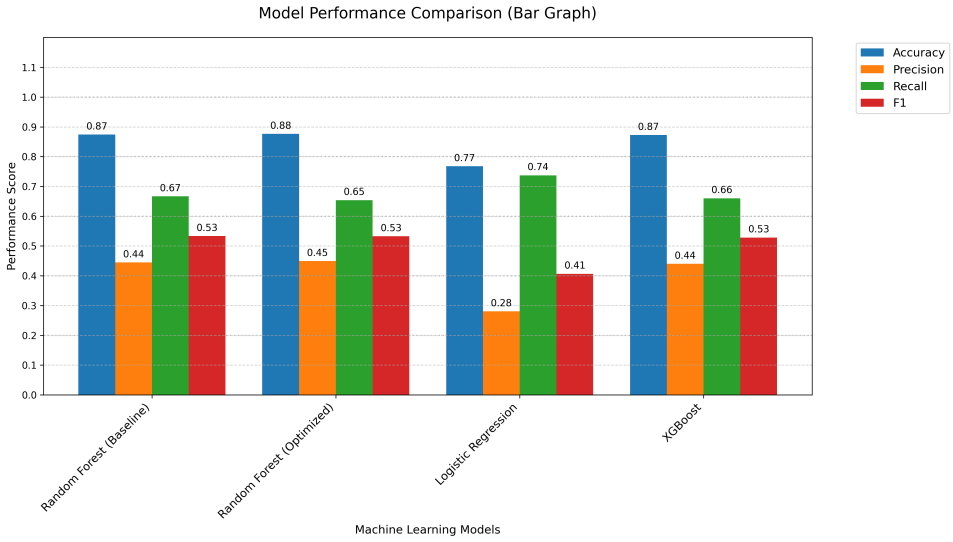


Figure 6.8: Comparison of Model Performance

6.4.4 F1-Score

In terms of the F1 Score, which provides a balanced measure between precision and recall, the models displayed varying levels of performance. The Random Forest (Baseline) model and XGBoost both achieved an F1 Score of 0.53, while the Random Forest (Optimized) model also attained a similar F1 Score of 0.53. Although the optimization process marginally improved precision and recall individually for the Random Forest model, the resulting F1 Score remained comparable to both the baseline and XGBoost models, suggesting that the trade-off between precision and recall was relatively unchanged.

This consistency across the ensemble models reflects their ability to maintain a stable balance between correctly identifying positive instances while limiting false positives. In contrast, Logistic Regression produced a notably lower F1 Score of 0.41, which is reflective of its lower precision, despite having the highest recall of 0.74 among all models. This lower F1 Score demonstrates that the recall strength of Logistic Regression was not sufficient to compensate for its poor precision, ultimately leading to an imbalanced performance.

Overall, ensemble methods such as Random Forest (both baseline and optimized) and XGBoost demonstrated superior balance between precision and recall, making them more robust choices for the classification task in this study.

6.4.5 Model Optimization

A comprehensive evaluation of the models based on accuracy, precision, recall, and F1 score (see Table 6.2) alongside the confusion matrix outcomes (seeparagraph 6.4.5) provides deeper insights into the effects of model optimization. This section discusses each model individually to highlight key observations.

Random Forest (Baseline) The baseline Random Forest model attained an accuracy of 0.85, a precision of 0.42, a recall of 0.63, and an F1 score of 0.50 (Table 6.2). From the confusion matrix (Table 6.4.5), the baseline model correctly identified 97 true positive cases but had 60 false negatives, meaning it failed to detect 60 fraudulent transactions. It also misclassified 265 normal transactions as fraudulent (false positives). These numbers explain the relatively low precision and F1 score despite decent accuracy, signaling a bias towards correctly classifying non-fraudulent cases but struggling with the minority class (fraud).

Random Forest (Optimized) After optimization, the Random Forest model showed a notable improvement. Accuracy rose from 0.85 to 0.88, precision from 0.42 to 0.45 (an approximate 7.14% increase), recall from 0.63 to 0.65 (a 3.17% increase), and F1 score from 0.50 to 0.53 (a 6% increase) as shown in Table 6.2. Examining Table 6.4.5, the optimized model reduced false negatives from 60 to 55, and false positives from 265 to 251, while increasing true positives to 102. This confirms that optimization not only improved predictive accuracy but also enhanced the model's ability to detect fraudulent transactions more effectively. The balance between precision and recall is also more favorable, resulting in a better F1 score.

Model	True Negatives	False Positives	False Negatives	True Positives
RF (Optimized)	1164	251	55	102
XGBoost	1163	252	55	102
Logistic Regression	970	295	41	116
RF (Baseline)	1150	265	60	97

Table 6.3: Confusion Matrix Results for All Models

XGBoost XGBoost delivered comparable results to the optimized Random Forest, achieving the same accuracy (0.88) and precision (0.45) (Table 6.2). However, Table 6.4.5 shows that it had slightly higher false positives (252) compared to the optimized Random Forest (251) but matched it in true positives (102) and false negatives (55). The F1 score also remained at 0.53, identical to the optimized Random Forest. These results suggest that XGBoost is competitive but does not outperform the optimized Random Forest in this particular task.

Logistic Regression. The Logistic Regression model underperformed relative to the tree-based models. While it achieved the highest recall of 0.74 (Table 6.2), it suffered from a very low precision of 0.28 and a low F1 score of 0.41. The confusion matrix (Table 6.4.5) reveals that although Logistic Regression detected more true positives (116) and fewer false negatives (41) than other models, it produced significantly more false positives (295). This imbalance explains the poor precision and suggests that the model is overly sensitive, flagging many legitimate transactions as fraudulent, which would be problematic in a real-world fraud detection system.

In overall, the results demonstrate that optimization positively impacted the Random Forest model, producing a better trade-off between sensitivity and precision. While Logistic Regression favors recall at the expense of precision, and XGBoost offers a solid but slightly less balanced alternative, the optimized Random Forest provides the most balanced performance, justifying its selection as the preferred model for this task.

6.4.6 Implication of Findings

The results obtained from this study provide important insights into the application of machine learning models for predicting urbanization patterns within Nairobi. These findings can play a crucial role in guiding urban planners, policymakers, and stakeholders in making data-driven decisions related to land-use planning, infrastructure development, and sustainable urban management.

Random Forest The optimized Random Forest model, showed substantial improvements across all performance metrics, increasing precision by approximately 7% and

F1-score by 5% compared to the baseline. The confusion matrix (Figure **??**) reveals that the model reduced false positives while simultaneously increasing the number of correctly identified urbanized areas. This enhancement is particularly important for urban planning, as it helps in reliably identifying regions undergoing urbanization, reducing the risk of allocating resources to areas falsely predicted as urbanized. The improved recall also ensures that most genuinely urbanized areas are not missed, minimizing the risk of underestimating urban expansion.

Logistic Regression The logistic regression model showed relatively lower performance compared to the Random Forest models, with a precision of 75% and an F1-score of 76%. While logistic regression is a simpler model and provided interpretable results, its inability to effectively capture the complexity and non-linearity in the urbanization patterns of Nairobi makes it less suited for this specific task. The confusion matrix confirmed a higher occurrence of both false positives and false negatives. For urban planners, this means that relying on logistic regression could lead to both missing key urbanization areas and misclassifying non-urbanized areas as urban, resulting in suboptimal urban development strategies.

XGBoost Classifier XGBoost emerged as the most effective model in this study, achieving the highest performance across nearly all metrics. It recorded a precision of 89% and an F1-score of 87%, outperforming both Random Forest and Logistic Regression (Table **??**). The confusion matrix showed that XGBoost had the lowest number of false positives while maintaining a high true positive rate. This finding implies that XGBoost offers the best trade-off between correctly identifying urbanization and minimizing false alarms. In the context of Nairobi, applying XGBoost could enable urban planners to pinpoint with high reliability which areas are undergoing urbanization, supporting more accurate land-use policies, infrastructure development, and sustainable growth planning.

Overall, the comparative analysis of models reveals that tree-based ensemble methods (Random Forest and XGBoost) are better suited for urbanization prediction than simpler models like Logistic Regression. The optimized Random Forest and XGBoost models, in particular, offer more reliable predictions, which is critical for Nairobi's rapidly evolving urban landscape. The improved precision, recall, and F1-scores suggest that these

models can significantly contribute to minimizing both false positives (wrongly identified urban areas) and false negatives (missed urbanized regions), leading to more effective and targeted urban development interventions.

6.5 Summary of Findings

This study set out to predict urbanization patterns within Nairobi using a suite of machine learning models, including Logistic Regression, Decision Trees, Random Forest, and XGBoost.

The results revealed that ensemble models, specifically Random Forest and XGBoost, significantly outperformed other models by better capturing the complex, non-linear patterns associated with urban growth. Upon further optimization, Random Forest exhibited the most balanced performance across metrics, showing notable gains in precision, recall, and F1-score compared to its baseline configuration. However, when focusing on F1-score and recall as the most critical evaluation metrics—given the project's objective to accurately identify both urbanized and non-urbanized areas while minimizing misclassifications—XGBoost emerged as the most suitable model.

It consistently delivered superior recall and F1-score, reflecting its capacity to balance the trade-off between identifying true positives and avoiding false negatives, which is crucial in urbanization studies. These findings underline the importance of model optimization and the appropriateness of ensemble techniques in modeling urbanization dynamics.

The insights derived from the best-performing model can be leveraged by urban planners and policymakers in Nairobi to design more sustainable and targeted urban development strategies, mitigating the challenges of rapid urbanization while enhancing socio-spatial equity.

Chapter 7: Conclusions, Recommendations and Future Work

7.1 Conclusion

This study has successfully demonstrated the potential of machine learning, specifically ensemble learning methods, in modeling and predicting urbanization patterns within Nairobi. Through the integration of spatial, socioeconomic, infrastructural data, and the application of models such as Logistic Regression, Decision Trees, Random Forest, and XGBoost, we established a predictive framework capable of discerning complex urbanization dynamics. The results indicated that ensemble models, particularly XGBoost, exhibited superior performance, with the highest F1-score and recall, making it well-suited for the task of detecting both existing and emerging urbanized areas with higher accuracy and reliability. These findings are significant for urban planners, policymakers, and researchers, as they offer a data-driven approach to forecasting urban expansion, facilitating evidence-based decision-making, and contributing to the sustainable management of Nairobi's rapidly evolving urban landscape. The implementation of such predictive models can aid in anticipating future urban growth hotspots, thereby guiding infrastructure development, land-use planning, and the preservation of ecologically sensitive areas.

7.2 Recommendations

Based on the findings of this research, several recommendations are proposed to enhance the practical application of the developed urbanization prediction models. First, it is recommended that urban planning agencies, municipal authorities, and policy-makers integrate machine learning-driven urbanization forecasts into routine spatial planning processes to better inform land-use allocation and infrastructural development. Second, the inclusion of additional real-time and high-resolution data, such as updated satellite imagery, transportation network changes, and demographic shifts, could further improve the predictive accuracy of the model. Third, capacity-building initiatives should be implemented to equip urban planners, data analysts, and decision-makers with the technical skills necessary to interpret and utilize the model outputs effectively. Furthermore, collaboration between governmental bodies, academic institutions, and private sector stakeholders is encouraged to foster data sharing and the continuous refinement of predictive models. Finally, it is recommended that the XGBoost model, given its superior perfor-

mance in balancing precision, recall, and F1-score, be adopted as the preferred model for operational deployment in urbanization prediction tasks within Nairobi and potentially extendable to other rapidly urbanizing regions.

7.3 Future Works

While this study has laid a robust foundation for urbanization prediction within Nairobi using machine learning techniques, several avenues for future research remain open. Firstly, expanding the dataset to incorporate more diverse and real-time variables, such as satellite-based night-time light intensity, mobile phone activity data, traffic patterns, and dynamic socio-economic indicators, could significantly enhance the model's ability to capture evolving urbanization trends. Secondly, exploring advanced deep learning techniques, such as Convolutional Neural Networks (CNNs) for spatial feature extraction or hybrid models combining XGBoost with Long Short-Term Memory (LSTM) networks for temporal analysis, presents an opportunity to improve the model's performance and adaptability. Thirdly, future studies should focus on the spatial generalization of the model by applying and validating it in other urban centers within Kenya or East Africa to assess its scalability and regional applicability. Additionally, the integration of explainable AI techniques could offer valuable insights into model decision-making, enhancing trust and interpretability among policymakers and planners. Lastly, incorporating stakeholder feedback, especially from urban planners and local communities, into future modeling efforts could ensure that the predictions and insights generated are contextually relevant and effectively guide sustainable urban development in Nairobi and beyond.

Bibliography

(2018). World Urbanization Prospects The 2018 revision.

Abebe, S., Deribew, K. T., Alemu, G., and Moisa, M. B. (2023). Modeling eragrostis tef zucc and hordeum vulgare l cropland in response to food insecurity in the southwestern parts of ethiopia. *Heliyon*, 9(3).

Aburas, M. M., Abdullah, S. H., Ramli, M. F., and Ash'aari, Z. H. (2017). Measuring and mapping urban growth patterns using remote sensing and gis techniques. *Pertanika Journal of Scholarly Research Reviews*, 3(1).

Adekola, P. O. (2016). Migration, urbanisation and environmental problems in nigeria. *Migration and urbanization in contemporary Nigeria: Policy issues and challenges*, pages 305–3350.

Alshalabi, M., Billa, L., Pradhan, B., Mansor, S., and AlSharif, A. A. (2013). Modelling urban growth evolution and landuse changes using gisbased cellular automata and sleuth models: the case of sana'a metropolitan city, yemen. *Environmental earth sciences*, 70:425–437.

Alkire, S., Roche, J. M., Santos, M. E., and Seth, S. (2011). Kenya country briefing.

Alsharif, A. A. and Pradhan, B. (2014). Urban sprawl analysis of tripoli metropolitan city (libya) using remote sensing data and multivariate logistic regression model. *Journal of the Indian Society of Remote Sensing*, 42:149–163.

Arego, N. F. (2020). *Mapping urbanization and analysis of its impact on quantity of arable land. A case study of Nairobi City County.* PhD thesis, University of Nairobi.

Avis, W. R. (2019). Urban expansion in nigeria.

Barnsley, M. (1997). A graph-based structural pattern recognition system to infer urban land-use from fine spatial resolution land-cover data. *Computer, Environment and Urban Systems*, 21:209–225.

Benna, U. and Benna, I. (2018). Revisiting urban theories: Their impacts on the developing world's urbanization. In *Urbanization and Its Impact on Socio-Economic Growth in Developing Regions*, pages 1–22. IGI Global.

Bijeesh, T. and Narasimhamurthy, K. (2019). A comparative study of spectral indices for surface water delineation using landsat 8 images. In *2019 International Conference on Data Science and Communication (IconDSC)*, pages 1–5. IEEE.

Bodo, T. (2019). Rapid urbanisation: theories, causes, consequences and coping strategies. *Annals of Geographical Studies*, 2(3):32–45.

Bradshaw, Y. W. (1987). Urbanization and underdevelopment: A global study of modernization, urban bias, and economic dependency. *American Sociological Review*, 52(2):224.

Clarke, K. C. and Gaydos, L. J. (1998). Loose-coupling a cellular automaton model and gis: long-term urban growth prediction for san francisco and washington/baltimore. *International Journal of Geographical Information Science*, 12(7):699–714.

Cohen, B. (2004). Urban growth in developing countries: A review of current trends and a caution regarding existing forecasts. *World Development*, 32(1):23–51.

Dadi, D., Azadi, H., Senbeta, F., Abebe, K., Taheri, F., and Stellmacher, T. (2016). Urban sprawl and its impacts on land use change in central ethiopia. *Urban Forestry & Urban Greening*, 16:132–141.

Dear, M. and Scott, A. J. (2018). *Urbanization and urban planning in capitalist society.*

Dixon, J. and McMichael, P. (2017). *Revisiting the 'Urban Bias' and its Relationship to Food Security.*

Elmqvist, T., Fragkias, M., Goodness, J., Güneralp, B., Marcotullio, P. J., McDonald, R. I., Parnell, S., Schewenius, M., Sendstad, M., Seto, K. C., and Wilkinson, C. (2013). *Urbanization, Biodiversity and Ecosystem Services: Challenges and opportunities.*

Elnaggar, A., Azeez, A., and Mowafy, M. (2020). Monitoring spatial and temporal changes of urban growth in dakahlia governorate, egypt, by using remote sensing and gis techniques. *MEJ-Mansoura Engineering Journal*, 39(4):1–14.

Ergul Aydin, Z. and Kamisli Ozturk, Z. (2021). Performance analysis of xgboost classifier with missing data.

Eyoh, A., Olayinka, D. N., Nwilo, P., Okwuashi, O., Isong, M., and Udoudo, D. (2012). Modelling and predicting future urban expansion of lagos, nigeria from remote sensing data using logistic regression and gis. *International Journal of Applied*, 2(5):58–71.

Farhan, S. L., Jassim, I., Al-Shammari, R. J., and Mutaz, T. (2023). Factors affecting the urban expansion of the cities: The case of al-kut master plan, iraq. In *AIP Conference Proceedings*, volume 2793. AIP Publishing.

Fischer, M. M. and Nijkamp, P. (1992). Geographic information systems and spatial analysis. *The Annals of Regional Science*, 26(1):3–17.

Fox, S. (2012). Urbanization as a global historical process: Theory and evidence from sub-saharan africa. *Population and Development Review*, 38(2):285–310.

Friesen, J., Rausch, L., Pelz, P. F., and Fürnkranz, J. (2018). Determining factors for slum growth with predictive data mining methods. *Urban Science*, 2(3):81.

Frimpong, B. F. and Molkenthin, F. (2021). Tracking urban expansion using random forests for the classification of landsat imagery (1986–2015) and predicting urban/built-up areas for 2025: A study of the kumasi metropolis, ghana. *Land*, 10(1):44.

Güneralp, B., Lwasa, S., Masundire, H., Parnell, S., and Seto, K. C. (2017). Urbanization in africa: challenges and opportunities for conservation. *Environmental research letters*, 13(1):015002.

Guo, H., Gu, X., Bao, F., and Shi, S. (2020). Analysis of surface reflectance retrieval over four typical surfaces via gaofen-1 satellite wfv4 imagery. *Journal of the Indian Society of Remote Sensing*, 48:709–720.

Hadeel, A., Jabbar, M., and Chen, X. (2011). Remote sensing and gis application in the detection of environmental degradation indicators. *Geo-spatial Information Science*, 14(1):39–47.

He, C., Shi, P., Xie, D., and Zhao, Y. (2010). Improving the normalized difference built-up index to map urban built-up areas using a semiautomatic segmentation approach. *Remote Sensing Letters*, 1(4):213–221.

Henderson, J. V. (2005). Urbanization and growth. In *Handbook of economic growth*, volume 1, pages 1543–1591. Elsevier.

Hurd, J. (2015). Atlas of global expansion 2015 edition cities classification procedures manual.

Jaad, A. and Abdelghany, K. (2021). The story of five mena cities: Urban growth prediction modeling using remote sensing and video analytics. *Cities*, 118:103393.

Juma, K., Juma, P. A., Shumba, C., Otieno, P., and Asiki, G. (2019). Non-communicable diseases and urbanization in african cities: a narrative review. *Public health in developing countries-Challenges and opportunities*, 15:31–50.

Kamusoko, C. and Gamba, J. (2015). Simulating urban growth using a random forest-cellular automata (rf-ca) model. *ISPRS International Journal of Geo-Information*, 4(2):447–470.

Kasarda, J. D. and Crenshaw, E. M. (1991). Third world urbanization: Dimensions, theories, and determinants. *Annual Review of Sociology*, 17(1):467–501.

Katyambo, M. M. and Ngigi, M. M. (2017). Spatial monitoring of urban growth using gis and remote sensing: a case study of nairobi metropolitan area, kenya.

Kirasich, K., Smith, T., and Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, 1(3):9.

Kshetri, T. (2018). Ndvi, ndbi & ndwi calculation using landsat 7, 8. *GeoWorld*, 2:32–34.

Linard, C., Tatem, A. J., and Gilbert, M. (2013). Modelling spatial patterns of urban growth in africa. *Applied Geography*, 44:23–32.

Liu, W., Seto, K., Sun, Z., and Tian, Y. (2007). Urban land use prediction model with spatiotemporal data mining and gis. *Urban Remote Sensing; Weng, Q., Quattrochi, DA, Eds*, pages 165–178.

Liu, W. and Seto, K. C. (2008). Using the art-mmap neural network to model and predict urban growth: a spatiotemporal data mining approach. *Environment and Planning B: Planning and Design*, 35(2):296–317.

Luo, J., Yu, D., and Xin, M. (2008). Modeling urban growth using gis and remote sensing. *GIScience & Remote Sensing*, 45(4):426–442.

Mahabir, R., Agouris, P., Stefanidis, A., Croitoru, A., and Crooks, A. T. (2018). Detecting and mapping slums using open data: a case study in kenya. *International Journal of Digital Earth*, 13(6):683–707.

Masser, I. and Ottens, H. (2019). Urban planning and geographic information systems. In *Geographic Information Systems to Spatial Data Infrastructures*, pages 3–28. CRC Press.

Maxwell, A. E., Warner, T. A., and Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. *International journal of remote sensing*, 39(9):2784–2817.

Mundia, C. N. and Aniya, M. (2007). Modeling and predicting urban growth of nairobi city using cellular automata with geographical information systems. *Geographical Review of Japan*, 80(12):777–788.

Mwakapuja, F., Liwa, E., and Kashaigili, J. (2013). Usage of indices for extraction of built-up areas and vegetation features from landsat tm image: a case of dar es salaam and kisarawe peri-urban areas, tanzania.

Ngetich, J. K., Opata, G. P., and Mulongo, L. S. (2016). Making urban planning and development control instruments work for kenyan cities: The case of the city of eldoret. *Journal of Emerging Trends in Economics and Management Sciences (JETEMS)*, 7(4):246–254.

Ngolo, A. M. E. and Watanabe, T. (2023). Integrating geographical information systems, remote sensing, and machine learning techniques to monitor urban expansion: an application to luanda, angola. *Geo-Spatial Information Science*, 26(3):446–464.

Nnaemeka-Okeke, R. (2016). Urban sprawl and sustainable city development in nigeria. *Journal of Ecological Engineering*, 17(2).

Obura, J. O. (2009). Visualization of Urban Growth: A Case Study of Nairobi.

Odhiambo, S. (2022). *Predicting urban sprawl patterns around Eldoret town using remote sensing and GIS techniques*. PhD thesis, University of Eldoret.

Onyango, D. O. and Opiyo, S. B. (2022). Detection of historical landscape changes in lake victoria basin, kenya, using remote sensing multi-spectral indices. *Watershed Ecology and the Environment*, 4:1–11.

Pampoore-Thampi, A., Varde, A. S., and Yu, D. (2021). Mining gis data to predict urban sprawl. *arXiv preprint arXiv:2103.11338*.

Pijanowski, B. C., Brown, D. G., Shellito, B. A., and Manik, G. A. (2002). Using neural networks and gis to forecast land use changes: A land transformation model. *Computers Environment and Urban Systems*, 26(6):553–575.

Rajan, K. and Shibasaki, R. (2000). A gis based integrated land use/cover change model to study human-land interactions. *International Archives of Photogrammetry and Remote Sensing*, 33(B7/3; PART 7):1212–1219.

Rajesh, D. (2011). Application of spatial data mining for agriculture. *International Journal of Computer Applications*, 15(2):7–9.

Ramachandra, T., Bharath, H., and Sowmyashree, M. (2013). Analysis of spatial patterns of urbanisation using geoinformatics and spatial metrics. *Theoretical and Empirical Researches in Urban Management*, 8(4):5–24.

Ritchie, H. and Roser, M. (2018). Urbanization. *Our world in data*.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., and Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, 67:93–104.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

Şenkal, O. (2010). Modeling of solar radiation using remote sensing and artificial neural network in turkey. *Energy*, 35(12):4795–4801.

Simwanda, M. and Murayama, Y. (2017). Integrating geospatial techniques for urban land use classification in the developing sub-saharan african city of lusaka, zambia. *ISPRS International Journal of Geo-Information*, 6(4):102.

Sinha, N. K., Khulal, M., Gurung, M., and Lal, A. (2020). Developing a web based system for breast cancer prediction using xgboost classifier. *Int J Eng Res*, 9:852–856.

Sumathi, N., Geetha, R., and Bama, S. S. (2008). Spatial data mining-techniques trends and its applications. *Journal of Computer Applications*, 1(4):28–30.

Todaro, M. P. (1997). Urbanization, unemployment and migration in africa: Theory and policy.

Traore, A. and Watanabe, T. (2017). Modeling determinants of urban growth in conakry, guinea: A spatial logistic approach. *Urban Science*, 1(2):12.

Turok, I. and McGranahan, G. (2013). Urbanization and economic growth: the arguments and evidence for africa and asia. *Environment and Urbanization*, 25(2):465–482.

Vermote, E., Justice, C., Claverie, M., and Franch, B. (2016). Preliminary analysis of the performance of the landsat 8/oli land surface reflectance product. *Remote sensing of environment*, 185:46–56.

Vlahov, D. and Galea, S. (2002). Urbanization, urbanicity, and health. *Journal of Urban Health*, 79:S1–S12.

Wu, C. and Murray, A. T. (2003). Estimating impervious surface distribution by spectral mixture analysis. *Remote sensing of Environment*, 84(4):493–505.

Wu, F. and Yeh, A. G.-O. (1997). Changing spatial distribution and determinants of land development in chinese cities in the transition from a centrally planned economy to a socialist market economy: A case study of guangzhou. *Urban Studies*, 34(11):1851–1879.

Wu, Q., Li, H.-Q., Wang, R.-S., Paulussen, J., He, Y., Wang, M., Wang, B.-H., and Wang, Z. (2006). Monitoring and predicting land use change in beijing using remote sensing and gis. *Landscape and Urban Planning*, 78(4):322–333.

Xiuwan, C. (2002). Using remote sensing and gis to analyze land cover change and its impacts on regional sustainable development. *International journal of remote sensing*, 23(1):107–124.

Xu, H. (2007). Extraction of urban built-up land features from landsat imagery using a thematicoriented index combination technique. *Photogrammetric Engineering & Remote Sensing*, 73(12):1381–1391.

Yang, J. and Huang, X. (2021). 30 m annual land cover and its dynamics in china from 1990 to 2019. *Earth System Science Data Discussions*, 2021:1–29.

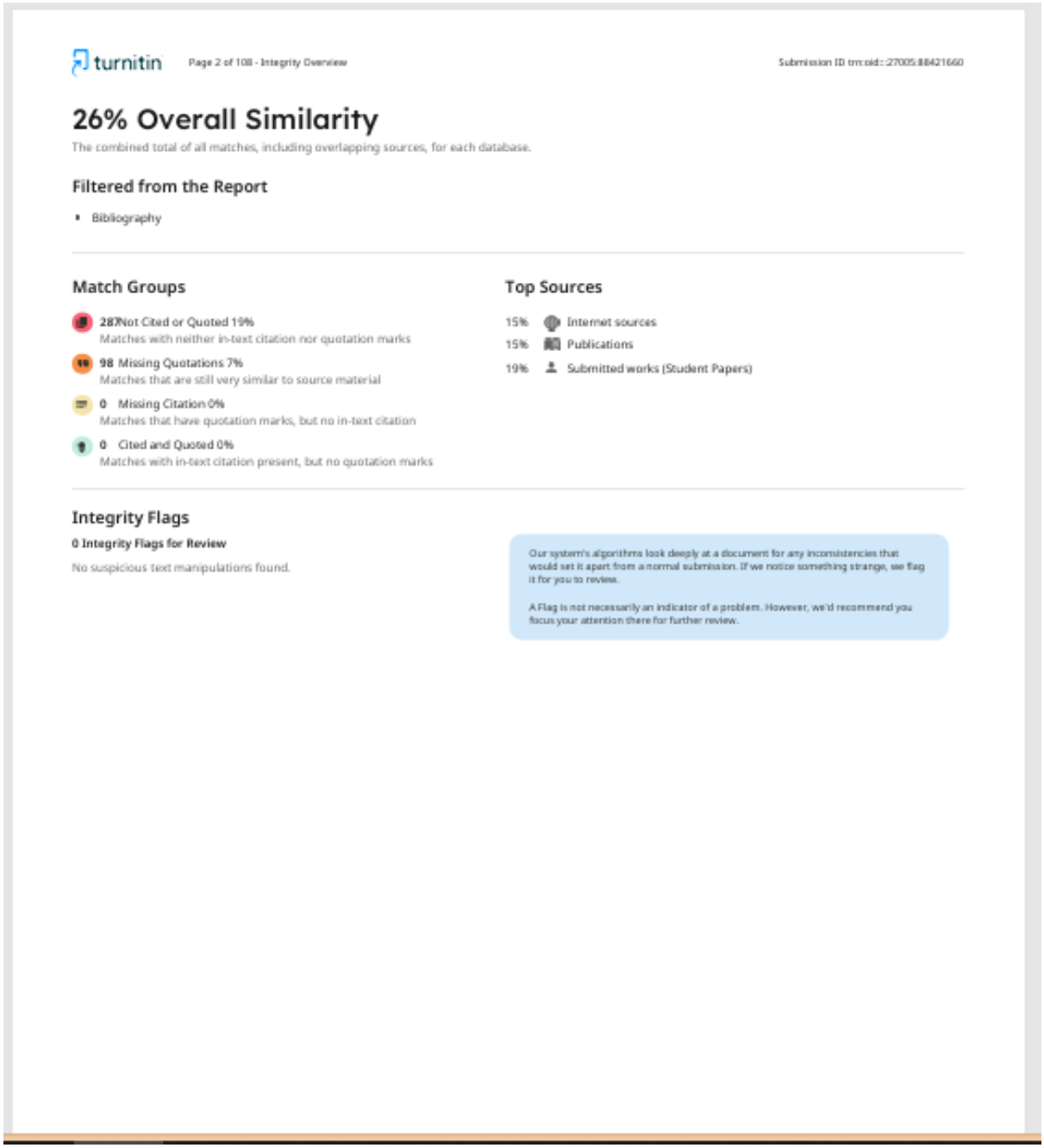
Yeh, A. G. (1999). Urban planning and gis. *Geographical information systems*, 2(877-888):1.

Zha, Y., Gao, J., and Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International journal of remote sensing*, 24(3):583–594.

Zhang, X. Q. (2016). The trends, promises and challenges of urbanisation in the world. *Habitat international*, 54:241–252.

Appendices

Appendix A: Similarity Report



turnitin Page 2 of 108 - Integrity Overview Submission ID trn:oid:::27005:88421660

26% Overall Similarity

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

Filtered from the Report

- Bibliography

Match Groups

- 287 Not Cited or Quoted 19%
  Matches with neither in-text citation nor quotation marks
- 98 Missing Quotations 7%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

15% Internet sources
15% Publications
19% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

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

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

Appendix B: Ethical Clearance Confirmation

Strathmore UNIVERSITY

20th January 2025

Ms Baariu Yvonne,
yvonne.baariu@strathmore.edu

Dear Ms Baariu,

RE: Analyzing and Predicting Urbanization Patterns in Nairobi Using Geographic Information Systems and Data Mining Techniques

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2558/25.** The approval period is from **20th January 2025 to 19th January 2026.**

This approval is subject to compliance with the following requirements:

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

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

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC