A Comparative Analysis of Machine Learning Models for Housing Price Prediction in Nairobi Metropolitan Area

Osoro, Faith Mirriam

150683

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Data Science and Analytics Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June 2025

Declaration

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

Name: **Osoro Faith Mirriam**

Signature: Faith Osoro

Date: May 19, 2025

Approval

Dr. Evans Otieno Omondi

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Prof. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

Background: The Nairobi Metropolitan Area has undergone significant transformation in recent years, driven by urbanization and economic opportunities. As the nation's capital and primary economic hub, it has attracted a growing population in need of housing solutions. This population influx has led to a dynamic housing market with fluctuating property prices, prompting the need for accurate housing price predictions. These predictions are crucial for various stakeholders, including homebuyers, developers, investors, and policymakers, as they guide decisions related to property investment, affordability, and policy formulation.

Methodology: The study employed advanced machine learning techniques to predict housing prices in the Nairobi Metropolitan Area. Various models, including spatial random forest and spatial autoregressive models, were utilized to develop predictive frameworks. Evaluation of these models was conducted using key metrics such as coefficient of determination (R-squared), root mean square error (RMSE), and mean absolute error (MAE).

Results: The results of the study revealed that the spatial random forest model emerged as the most effective tool for predicting housing prices in the Nairobi Metropolitan Area. The analysis also identified several key features that significantly influenced property values, including property size, number of bedrooms, and type of area. These findings underscored the importance of both structural attributes and neighborhood characteristics in determining house prices in the dynamic urban environment of Nairobi.

Conclusion: The study underscores the critical role of accurate housing price predictions in guiding decision-making for various stakeholders in the Nairobi Metropolitan Area. Leveraging advanced machine learning techniques, particularly the spatial random forest model, provides valuable insights into the dynamics of the local real estate market.

Keywords: *Housing prices, Nairobi Metropolitan Area, Machine learning, Predictive modeling, Spatial analysis*

Table of contents

List of figures

Acknowledgement

I express my heartfelt gratitude to my supervisor, Dr.Evans Otieno Omondi, for the support and guidance throughout the development of this dissertation. The expertise and encouragement have provided the necessary direction and inspiration for this study.

Secondly, my gratitude to the knowledge experts and advisors who have provided essential guidance and insights as I formulated the foundation of this proposal. Your expertise has been invaluable.

I am also thankful to my colleagues and peers for their insightful feedback and discussions, which significantly contributed to refining the focus of this study.

The stakeholders, including prospective homebuyers, real estate developers, investors, and government policymakers, the continued interest and input in shaping the research questions and objectives has been very instrumental.

Lastly, I also wish to acknowledge my family for their unwavering encouragement and understanding as I pursue this endeavor. Their support has been a constant source of motivation.

Dedication

I dedicate this research to my children, nieces, and nephews, with the hope that it serves as a testament to the power of knowledge and perseverance. May this work inspire you to embrace the world of learning and exploration and encourage you to read, question, and discover. Remember that with curiosity and dedication, you can unravel the mysteries of any field and contribute meaningfully to the world.

With love and inspiration

Chapter 1

Introduction

1.1 Background

The Nairobi Metropolitan Area has radically transformed over the past few years (Myers, 2015). As the nation's capital city and its primary economic epicenter, Nairobi and its adjoining metropolitan expanse have become magnets for a steadily growing population (Scott, 2019) in search of housing solutions. This population influx, driven by factors ranging from urbanization to economic opportunities, has given rise to a dynamic and ever-evolving housing market. It is within this intricate web of supply and demand dynamics that property prices in the Nairobi Metropolitan Area have experienced pronounced fluctuations (Vuluku et al., 2014), thus elevating the significance of this domain as a focal point of scrutiny and concern for both residents and stakeholders within the real estate sector.

The inherent volatility of the housing market, coupled with the crucial role housing plays in people's lives and the broader economy, underscores the vital importance of accurate housing price predictions (Arundel, 2017). These predictions hold profound implications for a wide spectrum of stakeholders. For prospective homebuyers, they represent a compass for informed decision-making, guiding them in selecting properties that align with their financial means and aspirations. Real estate developers and investors rely on these forecasts to strategically position themselves within the market, optimizing their investments for sustainable returns (Khanna et al., 2015). Government policymakers, cognizant of the housing sector's role in the broader economy and society, seek reliable price predictions to formulate effective housing policies that foster affordability, stability, and equitable access to housing.

In this context, the emergence of machine learning (ML) as a potent tool for predictive analysis assumes paramount significance (Hinterwimmer et al., 2022). Machine learning

techniques, underpinned by their capacity to harness vast datasets and intricate algorithms (Badini et al., 2023), have demonstrated remarkable potential in addressing the multifaceted challenge of predicting housing prices. By leveraging data-driven approaches, Machine Learning models can encapsulate the complex interplay of variables and market dynamics, offering a more nuanced and adaptable framework for modeling and forecasting (Veldkamp and Chung, 2019) housing prices. As such, the intersection of machine learning and the Nairobi Metropolitan Area's housing market presents a promising avenue for enhancing the accuracy of price predictions and revolutionizing how stakeholders navigate the intricate landscape of property transactions and investments in this dynamic urban environment.

This dissertation embarks on a comprehensive exploration of the prediction of housing prices within the Nairobi Metropolitan Area, utilizing advanced machine learning techniques. It endeavors to illuminate the path towards more precise and insightful housing price predictions by systematically investigating various facets of the housing market, coupled with the judicious application of machine learning algorithms. The research aims to empower individuals, developers, investors, and policymakers with valuable insights and tools to navigate and contribute to the sustainable development of the housing sector in Nairobi's bustling metropolis.

The outcomes of this study hold significant implications for various stakeholders, including real estate professionals, policymakers, and urban planners. The predictive models will provide accurate estimations of housing prices in the Nairobi Metropolitan Area, empowering prospective buyers and sellers with valuable information for decision-making. Real estate professionals can use the insights to optimize pricing strategies, while policymakers and urban planners can leverage the findings to inform housing policies, urban development initiatives, and infrastructure planning.

1.2 Problem statement

The rapid urbanization and economic growth in the Nairobi Metropolitan Area have resulted in a surging demand for housing, leading to a dynamic and volatile housing market. This has

created a pressing challenge for stakeholders, including home buyers, real estate developers, investors, and government policymakers. Prospective home buyers face uncertainty when purchasing property due to unreliable housing price predictions. Similarly, real estate developers and investors struggle to make strategic investment decisions in a market characterized by price volatility, potentially resulting in financial losses or missed profit opportunities. Government policymakers struggle to formulate effective housing policies without precise housing price forecasts, which may exacerbate housing affordability issues and economic imbalances.

This research leveraged machine learning techniques to address the problem of inadequate housing price predictions within the Nairobi Metropolitan Area. While machine learning has shown promise in accurately forecasting housing prices by harnessing vast datasets and complex algorithms, its specific application in Nairobi's housing market remains underexplored. By investigating various facets of the housing market and employing advanced machine learning algorithms, this study developed predictive models that empower stakeholders with the insights and tools to navigate the dynamic housing market effectively, enabling them to make informed decisions and contribute to sustainable housing development in Nairobi's metropolitan region.

Efforts to tackle the challenge of inadequate housing price predictions in the Nairobi Metropolitan Area have witnessed strides in historical data analysis, traditional statistical modeling, data collection initiatives, and policy interventions. Researchers and analysts have delved into historical market trends, providing valuable insights, and policymakers have implemented interventions to stabilize the housing market. Extensive datasets related to the housing market have been collected and organized, forming the basis for various analyses. There are no comprehensive machine learning applications adapted to Nairobi's particular market characteristics. The development of localized predictive models considering neighborhood-specific factors, and the establishment of real-time prediction systems have not been adequately explored. Addressing these unexplored territories by integrating comprehensive machine learning approaches, leveraging diverse data sources, focusing on localized models, and enabling real-time predictions is imperative. Such advancements

would not only bridge existing gaps but also equip stakeholders with the precise tools needed to navigate Nairobi's dynamic housing market effectively, making informed decisions and contributing to sustainable urban development.

1.3 Research objectives

1.3.1 General objective

This study aimed to develop and apply a machine learning-based predictive model for estimating housing prices in the Nairobi Metropolitan Area. The goal is to support buyers, real estate developers, and policymakers in making more informed investment and planning decisions through data-driven insights.

1.3.2 Specific objective

1. To identify and evaluate the key structural, locational, and amenity-related variables influencing residential housing prices in the Nairobi Metropolitan Area.
2. To explore and apply multiple machine learning algorithms, including spatial and non-spatial models, for predicting housing prices.
3. To conduct a comparative analysis of the performance of selected machine learning models using appropriate evaluation metrics such as RMSE and R-squared.
4. To interpret the relative importance of predictive variables in influencing housing prices, based on model outputs.
5. To develop and deploy a user-friendly web application that provides real-time housing price predictions for stakeholders in the Nairobi Metropolitan Area.

Figure 1.1: Map of Nairobi Metropolitan Area

1.4 Scope of study

The primary geographic area of concentration is the Nairobi Metropolitan Area, comprising Nairobi County, Kiambu County, Kajiado County, and select portions of Machakos County and Murang'a County as shown in Figure 1.1 This strategic selection allows for an in-depth exploration of diverse housing markets within the broader Nairobi Metropolitan region, each marked by unique characteristics and dynamics.

At the core of this research lies the development and evaluation of machine learning predictive models for housing prices. These models will be carefully constructed to harness the power of data-driven approaches and leverage various machine learning algorithms, including regression models, ensemble methods, and deep learning techniques, to furnish highly accurate predictions of housing prices within the chosen geographical scope. A comprehensive approach to data collection is envisaged, focusing on diverse datasets that encompass critical information on housing attributes, economic indicators, demographic data, and market trends.

Data will be meticulously collected from various sources, including government records, real estate databases, surveys, and publicly available data. Robust preprocessing techniques will be applied to ensure data quality and consistency.

This study is committed to the identification and analysis of pivotal variables that exert influence on housing prices. These variables will encompass location-specific characteristics, intrinsic property features, proximity to essential amenities, and market trends. Selection criteria for these variables will be grounded in their relevance to the unique dynamics of the Nairobi Metropolitan Area's housing market. The research will rigorously investigate and implement various regression machine learning algorithms, including Decision Trees, Random Forests and Ridge Regression, to ascertain their effectiveness in delivering precise and reliable housing price predictions within the defined research context.

The developed predictive models were tested and evaluated, employing appropriate evaluation metrics. One of the principal objectives of this research was to provide insights and recommendations that are directly applicable to practical real estate scenarios within the Nairobi Metropolitan Area. This encompassed addressing the information requirements of various stakeholders, including prospective homebuyers, real estate developers, investors, and government policymakers.

Beyond predictive modeling, this study investigated the broader social and economic implications of accurate housing price predictions in Nairobi. It explored how improved predictive accuracy can potentially influence housing affordability, investment strategies, and the formulation of housing-related policies. While historical data formed the basis of this research, it also factored in the temporal dynamics inherent to the housing market. This entailed accounting for evolving economic conditions, shifting market trends, and their consequential impact on housing prices.

The study was undertaken with an awareness of its limitations, which included constraints related to data availability, potential biases, and the inherently dynamic nature of the housing market. These limitations were transparently acknowledged and addressed in the research's findings and discussions.

1.5 Significance of study

The significance of this study in predicting housing prices in the Nairobi Metropolitan Area is multifaceted and impactful. The complexity of Nairobi's real estate market, marked by varied neighborhood dynamics, rapid urban development, and informal settlements, necessitates a deep understanding of the intricate web of variables influencing house prices. The study contributes substantially to informed decision-making within the real estate sector by unraveling this complexity. Accurate housing price predictions are fundamental for buyers and sellers alike. Potential homeowners rely on precise estimations to make investments aligned with market values, while sellers need reliable pricing strategies to optimize their returns. This study directly empowers individuals and businesses by developing a machine learning model capable of accurate predictions, fostering a more informed and efficient real estate market in Nairobi.

1.6 Research assumptions

The study assumed that market movements within the defined geographic area were homogeneous, allowing for broad generalizations about housing price dynamics. This simplifying assumption allowed a more manageable study, but it might not have fully conveyed the various and complicated structure of the metropolitan region's neighborhoods. Economic conditions were considered stable during the study period, suggesting that large economic upheavals or financial crises were not expected. This assumption was critical because it provided a consistent economic backdrop against which housing market movements could be appropriately analyzed. Furthermore, the study presupposed the availability of precise and reliable data from a variety of sources, such as government records and real estate databases, in order to allow the development of reliable predictive models. However, the inherent limitations of data collection were acknowledged, including potential biases or gaps, emphasizing the importance of robust validation techniques.

The study also assumed that important market variables such as housing demand, supply, and pricing remained stable across the study period. While this simplification helped the modeling process, it might not have adequately reflected the real-world scenario in which these parameters frequently fluctuate over time. Furthermore, the study assumed a certain level of uniformity in the social and cultural elements impacting housing choices in the Nairobi Metropolitan Area. While this assumption simplified the analysis, it might have oversimplified the many cultural elements that influence homebuyers' decisions. Finally, the analysis assumed that no major policy changes occurred throughout the research period, such as significant changes in housing rules or taxation policies, which could have had a large impact on the housing market. Researchers were aware of these assumptions and their potential impact on the study's findings, and they frequently used sensitivity analyses to examine the robustness of the results in the face of differences in these assumptions.

1.7 Limitation of the study

The scope of this work was precisely defined, focusing on constructing machine learning predicting models for housing prices in the Nairobi Metropolitan Area. However, it was critical to recognize the limits of this research project. Because of the various market dynamics, the regional focus on the Nairobi Metropolitan Area limited the direct applicability of findings to other regions. Second, the study's accuracy depended on data availability and quality, with potential biases being a substantial difficulty. Furthermore, the study's bounds were shaped by the temporal nature of the housing market, social and economic effects, ethical questions connected to biases in machine learning, and the demand for stakeholder participation.

To address these limitations, the research used rigorous validation methods, reported data constraints transparently, incorporated temporal analysis, collaborated with experts from various fields, and actively engaged stakeholders to ensure the relevance and ethical integrity of the study's findings.

Chapter 2

Literature Review

2.1 Overview

The chapter is structured around our key objectives, determining key variables for predicting housing prices, developing machine learning models for predicting house prices, and assessing their performance using appropriate metrics. For each objective, it delves into the theoretical underpinnings that guide the study, examining established concepts in housing market dynamics and predictive modeling. This research conducts an empirical review, analyzing previous research endeavors globally and within Africa.

2.2 Theoretical Framework

The theoretical foundation of this study integrates economic, behavioral, and spatial theories that collectively explain the determinants of housing prices and inform the construction of predictive models. These theories not only guide the selection of variables but also support the justification for using machine learning techniques to capture complex, nonlinear relationships in real estate data.

2.2.1 Hedonic Pricing Model

The Hedonic Pricing Model (HPM) serves as a primary theoretical basis for modeling house prices. It posits that the price of a house is a function of its various characteristics, such as location, size, number of rooms, amenities, and neighborhood attributes Heyman and Sommervoll (2019). According to Jia et al. (2023), attributes like proximity to parks, low

crime rates, and cultural centers are capitalized into property values. This model is essential for identifying the key structural, locational, and environmental variables that affect real estate valuation.

2.2.2 Prospect Theory

Prospect Theory, developed within the field of behavioral economics, provides insight into how homebuyers make decisions under risk and uncertainty. It suggests that individuals evaluate potential gains and losses relative to a reference point rather than in absolute terms Barberis (2013). Empirical studies such as Antonides and Ranyard (2017) have demonstrated how mental accounting, loss aversion, and framing effects influence housing preferences, particularly in contexts involving aesthetics and perceived safety. This theory informs the inclusion of subjective or perceived property characteristics in predictive models.

2.2.3 Spatial Autoregressive Models

The spatial dimension of real estate markets is captured through Spatial Autoregressive Models (SAR), which recognize that housing prices are influenced by the values of nearby properties Zhang et al. (2021). As demonstrated by Feng and Humphreys (2018), neighborhood effects such as school quality and crime levels create spatial dependencies that must be considered in predictive modeling. Incorporating spatial autocorrelation into the framework enhances the model's ability to detect localized patterns in housing price distribution.

2.2.4 Capitalization Theory

Capitalization Theory explains how public goods and local government services are embedded into property values. Research by Oates and Schwab has shown that factors such as tax rates, school quality, and infrastructure are capitalized into housing prices Gallagher et al. (2013). This theory supports the inclusion of public policy and governance indicators as variables in the prediction model.

2.2.5 Theory of Generalization in Machine Learning

Machine learning models must be trained not only to fit historical data but also to generalize to unseen examples. The Theory of Generalization, particularly through the Vapnik-Chervonenkis (VC) dimension, provides a mathematical basis for understanding the trade-off between model complexity and generalization error Jakubovitz et al. (2019); Kawaguchi et al. (2018). As Oono and Suzuki (2020) demonstrates, models with appropriate regularization and complexity control are more likely to perform well in real-world prediction tasks. This theory underpins the use of machine learning algorithms in this study and justifies their application to housing price prediction.

2.2.6 Synthesis of Theories

These theoretical perspectives collectively shape the methodological approach of this study. While the HPM and Capitalization Theory inform variable selection, Prospect Theory and Spatial Autoregression guide contextual understanding. The Theory of Generalization justifies the machine learning approach, ensuring the selected models are robust, interpretable, and adaptable to the Nairobi Metropolitan Area's dynamic housing market.

2.3 Empirical Review

This section reviews empirical studies that have investigated the determinants of housing prices and the application of machine learning algorithms in real estate valuation. The review is structured to highlight findings from diverse contexts, emphasizing similarities and contrasts across geographic regions and methodological approaches. These studies provide valuable insights into variable selection, model performance, and contextual considerations relevant to the Nairobi Metropolitan Area.

2.3.1 Empirical Studies on Housing Price Determinants

Numerous studies across different countries have explored factors influencing housing prices. In Malaysia, Ngu Ngu (2017) identified location, building costs, government policy, and interest rates as key drivers of real estate values. Similar results were found in Kosovo by Prishtina Hoxha et al. (2021), where factors such as size, road access, and building quality played significant roles.

In the United States, Dong Dong (2020) analyzed transaction data from New York City and discovered that the impact of housing features varied across market cycles. During periods of price appreciation, land area was a dominant factor, while structure and location became more important during recovery phases. In the Chinese context, a study in Hefei City Xu and Zhang (2021) found that proximity to schools and transport infrastructure had the most significant influence on house prices.

In New Zealand, macroeconomic factors such as interest rates, migration, and income levels have been shown to substantially impact housing markets Chancellor et al. (2016). Related studies in Japan, Norway, and South Africa highlighted the correlation between GDP and house prices Lewandowska et al. (2023), emphasizing the broader economic context in which property markets operate. In Nigeria, Alkali et al. Habanabakize and Dickason (2022) found that inflation, interest rates, exchange rates, and oil prices significantly influenced real estate trends.

Closer to the focus area of this study, a Kenyan study by Ndegwa (2018) found that apartment prices in Nairobi are significantly affected by proximity to malls, the central business district, schools, swimming pools, and land value. These findings provide a foundational understanding of the factors that must be considered when developing predictive models for the Nairobi Metropolitan Area.

2.3.2 Empirical Studies on Machine Learning in Housing Price Prediction

Several empirical studies have demonstrated the utility of machine learning models in improving the accuracy of housing price predictions. A study conducted in Seoul, South Korea by Kim et al. (2022) compared neural networks, random forests, and spatial interpolation methods. Random Forests and Neural Networks outperformed traditional techniques, with Random Forest showing superior accuracy.

In Saudi Arabia, Alshammari (2023) confirmed the effectiveness of Random Forest in real estate prediction tasks. Likewise, a case study in King County, USA by Zhang (2022) highlighted strong correlations between housing prices and features such as living space, location, and building grade. Similarly, in Hong Kong, a comparative analysis of Gradient Boosting Machines (GBM), Random Forests, and Support Vector Machines revealed that ensemble methods generally yield better predictive performance Choy and Ho (2023).

In Brazil, Afonso et al. de Aquino Afonso et al. (2019) analyzed over 12 million housing records using Recurrent Neural Networks and Random Forests. Their findings demonstrated that hybrid models combining multiple ML algorithms and rich datasets can significantly enhance valuation accuracy.

2.3.3 Summary of Empirical Insights

The reviewed studies confirm that both macroeconomic and micro-level housing attributes play vital roles in shaping property values. Moreover, machine learning algorithms—especially ensemble methods—have consistently demonstrated superior predictive power across various datasets and contexts. These empirical findings validate the approach taken in this study, which aims to leverage machine learning techniques to model housing prices in the Nairobi Metropolitan Area while drawing on a diverse set of predictive features.

2.4 Conceptual Framework

The conceptual framework for this study is informed by the theories and empirical findings discussed in the preceding sections. It illustrates the hypothesized relationships between key predictors of housing prices and the predicted outcome (house price), with machine learning serving as the analytical mechanism for modeling these relationships.

Drawing from the Hedonic Pricing Model, property characteristics such as size, number of bedrooms, and amenities are treated as intrinsic value components. Spatial Autoregressive Theory suggests that neighborhood characteristics (e.g., proximity to schools, shopping centers, crime levels) influence property values through spatial spillovers. Capitalization Theory supports the inclusion of public service quality and local government factors, while Prospect Theory introduces behavioral aspects such as perceived safety or aesthetics.

Machine learning algorithms, guided by the Theory of Generalization, are used to model the complex, nonlinear interactions among these variables. The framework reflects a data-driven, integrative approach to house price prediction in the Nairobi Metropolitan Area.

2.5 Research Gap

A comprehensive review of the literature reveals that while numerous studies have explored the determinants of housing prices and the application of machine learning (ML) in real estate valuation, several important gaps remain particularly in the context of developing countries such as Kenya.

The majority of empirical studies on Machine Learning based house price prediction have been conducted in developed countries such as the United States, China, South Korea, and Brazil Choy and Ho (2023); Kim et al. (2022); Zhang (2022). These contexts differ significantly from Nairobi in terms of data availability, urban infrastructure, regulatory environment, and housing market dynamics. As a result, findings from these studies cannot be directly generalized to the Nairobi Metropolitan Area.

Although a few studies within Africa and Kenya have examined housing price determinants Ndegwa (2018), most rely on traditional statistical methods and do not leverage the predictive power of machine learning. There is limited evidence on how various ML algorithms perform within the Kenyan housing market, particularly with regard to comparative evaluation across multiple models.

Also while existing research has identified key predictive variables (e.g., location, property features, amenities), few studies have provided a robust framework for integrating these heterogeneous data sources into a machine learning pipeline tailored for price prediction in Nairobi. The absence of such integration limits the practical application of these models for stakeholders such as developers, investors, and policymakers.

Finally, few studies incorporate spatial and behavioral theories (e.g., Spatial Autoregressive Models, Prospect Theory) alongside ML approaches to offer a holistic understanding of house price drivers. This presents an opportunity to bridge theory with data science for a more comprehensive modeling approach.

In response to these gaps, this study applies and compares multiple machine learning algorithms for housing price prediction using real estate data from the Nairobi Metropolitan Area. It integrates spatial, behavioral, and economic variables within a theoretically grounded ML framework to develop a robust and context-sensitive predictive model. By doing so, the study aims to provide a practical, data-driven tool for real estate valuation that supports more informed decision-making by housing sector stakeholders in Kenya.

Chapter 3

Methodology

3.1 Introduction

This dissertation to utilized emerging technologies such as machine learning techniques to develop a predictive price model to determine the price trends of house prices in Nairobi Metropolitan. The data is on residential houses in Nairobi Metropolitan Area.

3.2 Research design

The Cross Industry Standard Process for Data Mining (CRISP-DM) model was utilized in the research, which was a well-established methodology for completing data mining projects. The CRISP-DM as presented in Figure 3.1 paradigm gave an organized, systematic way to dealing with diverse data mining activities, allowing for a more in-depth knowledge of the underlying data. It was divided into six main phases. The first phase was Business Understanding, which outlined the project's objectives and requirements from a business standpoint. The second stage was Data Understanding, focused on data capture and exploration. Gathering the appropriate datasets, assessing their structure and content, and finding any potential flaws or discrepancies were all part of this process. This stage gave useful insights that were used to inform the succeeding data preparation and modeling processes.

The third phase, data Preparation, was a thorough procedure that involved cleansing, transforming, and integrating the acquired data. Data cleaning entailed dealing with missing numbers, outliers, and errors to ensure the dataset's integrity. Transformation techniques were used to normalize data distributions and suitably scale variables. Data integration was

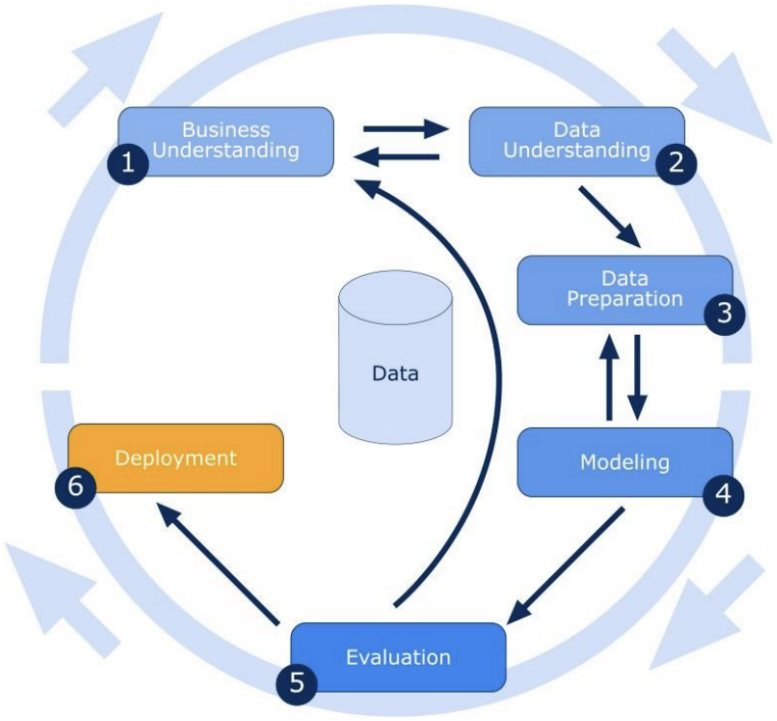


Figure 3.1: CrispDM model

the process of merging disparate information and enriching them with extra variables or features that improved the predictive potential of the models. This phase was critical since the accuracy and dependability of the resulting prediction models were heavily influenced by the quality of the input data.

Following data preparation, machine learning methods were used to develop prediction models in the Modeling step. Given the broad and intricate nature of Nairobi's real estate market, a variety of algorithms, including Support Vector regression, Gradient Boosting, Decision Trees, and Random Forests, were investigated. These algorithms were fine-tuned to the Nairobi Metropolitan Area's specific characteristics, based on both worldwide best practices and localized empirical investigations. To identify the most accurate and trustworthy predictive algorithms for home price projections, iterative experimentation and evaluation of these models against predetermined metrics were carried out.

The model's performance was evaluated during the evaluation phase. In accordance with the global and local empirical research evaluated, evaluation metrics such as Mean Absolute

Percentage Error (MAPE), Root Mean Square Error (RMSE), and R-squared were used. These measures provided quantitative insights into the models' accuracy, precision, and reliability. The models' robustness in reflecting the complexity of Nairobi's housing market was thoroughly tested against historical data. Iterative feedback loops were implemented, allowing model improvements based on evaluation results.

Finally, the deployment step signified the move from theoretical to operational constructions. Predictive models that had been developed and validated were integrated into a user-friendly web application. This application provided real-time, on-demand home price projections for Nairobi. Users, including potential buyers, sellers, and real estate experts, were able to receive reliable and data-driven predictions via a streamlined interface. User feedback was collected and analyzed in real time, allowing for iterative changes to the program and assuring its effectiveness and usability in the real world.The flow chart in Figure 3.2 shows the flow of processes.

This research attempted to not only predict housing prices in Nairobi but also to build a methodological framework that was scalable and transferrable to other complicated real estate markets by applying the CRISP-DM model and diligently navigating through its organized phases. The study aimed to deliver actionable insights, improve decision-making processes, and empower stakeholders in the Nairobi Metropolitan Area's housing market through this holistic methodology.

3.3 Data sources and collection method

The data for this study was obtained online from real estate listing companies and KNBS. The study aimed to focus only on secondary data. The methodology involved the utilization of Python scripts, incorporating the Beautiful Soup library, to extract structured data from HTML web pages. Beautiful Soup, a popular Python library, provided tools for web scraping HTML and XML documents, allowing for seamless extraction of specific elements such as property details, location information, amenities, pricing, and listing dates. The research nav-

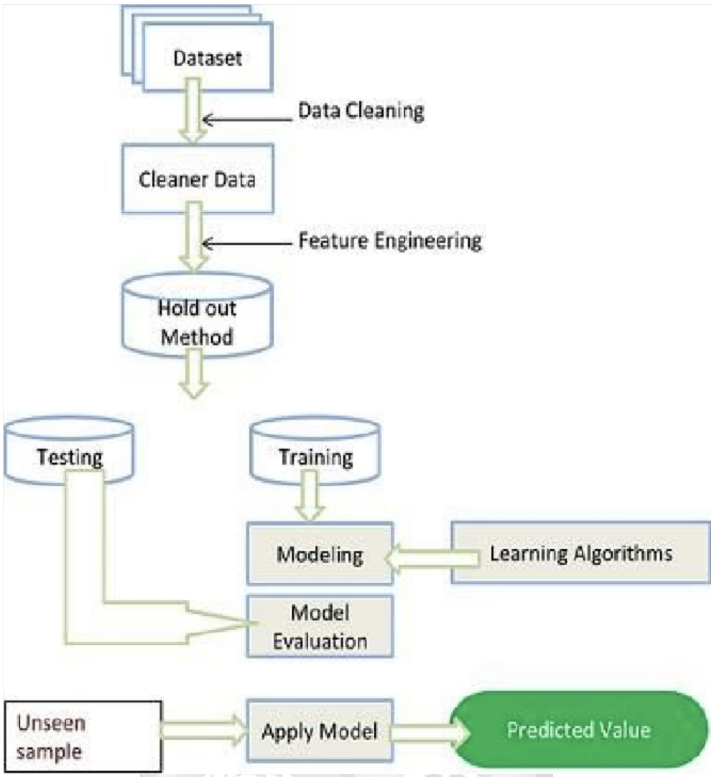


Figure 3.2: Flow chart

igated through the complexities of various real estate websites through meticulous scripting, capturing relevant data points essential for the predictive models.

3.3.1 Description of Key Variables

The dataset consists of various variables that are theoretically and practically relevant to housing price prediction. These variables were selected based on their prevalence in real estate valuation literature, as well as their availability in the Nairobi Metropolitan Area dataset. The features fall under four main categories:

Structural property features: *House type*, *number of bedrooms*, *number of bathrooms*, and *property size (square footage)*. These features directly describe the physical attributes of the property and are known to be significant drivers of price.

Locational features: *Area type* and geographic coordinates. These capture spatial characteristics that influence demand, including accessibility and proximity to commercial centers.

Amenities and services: Presence of features such as a *swimming pool*, *internet fibre*, *borehole*, *staff quarters*, *generator*, and *gym*. These factors add value to a property and reflect lifestyle preferences of buyers in upper and middle-income segments.

Target variable: *Price* of the property, which is the dependent variable for prediction using machine learning algorithms.

3.4 Data pre-processing

The next step in the research was to pre-process and transform the data. Once the raw data was scraped from real estate websites, it underwent a crucial process of data cleaning and integration. This step ensured the gathered information was accurate, consistent, and ready for analysis. Here's how we refined the collected data.

3.4.1 Data cleaning

Data cleaning involved identifying and correcting errors or inconsistencies in the scraped data. This included handling missing values, correcting formatting issues, and removing duplicates. By meticulously examining each data point, the study ensured the dataset was free from inaccuracies that could impact the reliability of our predictions.

Data inspection and exploration: Understanding the complexities of the scraped real estate dataset was based on data exploration and inspection. This thorough procedure went beyond a glance and delved deeply into the dataset's details to reveal hidden patterns and potentially insightful information essential for making accurate predictions.

Identifying data entry errors and noise sources: Real estate datasets, especially those scraped from online platforms, were susceptible to data entry errors and noise. These errors included typos in property descriptions, inaccurate pricing information, or discrepancies in amenities listed. Through meticulous scrutiny, potential data entry inaccuracies were spotted. Additionally, noise sources, such as inconsistent formatting or mismatched property details,

were meticulously identified. Addressing these errors and sources of noise was essential to ensure the dataset's reliability. Techniques like data cleansing, where erroneous entries were corrected, and noise reduction, where inconsistencies were normalized, were applied to enhance data quality.

Exploring relationships: Examining the connections between variables was a further extension of data exploration. Correlation analysis was performed to reveal relationships between traits, providing insightful information on how different features interact with one another. For instance, understanding the relationship between property size and price provided insight into the market's valuation standards. Visualizing tools like scatter plots and heat maps represented these links visually, making complex patterns more understandable. Investigating these correlations helped choose the most relevant features for the prediction models.

Handling missing data: In addressing missing data, particularly in the size column, the MICE (Multivariate Imputation by Chained Equations) library in R was utilized as part of a comprehensive data preprocessing strategy. The MICE approach facilitated the imputation process by considering the interrelationships between the size variable and other relevant features within the dataset. Through iterative iterations, MICE imputed missing values for the size column based on information derived from other variables, thereby capturing the underlying patterns and dependencies present in the data. By leveraging this multivariate imputation technique, the integrity of the size variable was preserved, enabling subsequent analyses to proceed with a more complete and representative dataset.

3.5 Data transformation

3.5.1 Categorical and variable and encoding

Categorical variables, including location, property type, and security level, were represented as labels that could not be directly utilized as inputs in machine learning algorithms. They were encoded into numerical values using the One-Hot Encoding method to enable their

integration into the computational framework. This technique transformed each distinct category within the original categorical variables into a separate binary variable, denoted as 0 or 1, thus creating a structured numerical representation suitable for machine learning analysis.

3.5.2 Feature scaling

The next stage in the methodology involved feature scaling, which aimed to standardize the range of features to a common scale, typically within a predefined range such as [0, 1] or [-1, 1]. Notably, the target variable underwent scaling by taking the natural logarithm, a common practice to address skewed distributions and stabilize variance. This transformation was applied to ensure that the target variable conformed to the same scaling standards as the input features.

The dataset underwent scaling to address discrepancies in feature scales, which could potentially impact the performance of machine learning algorithms. Specifically, features with differing scales were standardized to have zero mean and unit variance, aligning with the study's customary scaling methodology.

3.5.3 Feature selection

In the process of feature selection, statistical tests were employed to assess the significance of variables concerning the target variable, price. For numerical variables, the Spearman correlation test was utilized to explore potential monotonic relationships with price. Additionally, for categorical variables with more than three levels, the chi-square test was employed to evaluate associations between categories and price. Conversely, the Wilcoxon rank-sum test was applied to categorical variables with only two levels, enabling comparison of distributions between groups. By employing these tests based on variable types and levels, a comprehensive assessment of their significance in relation to price was achieved, aiding in the identification of informative features for subsequent modeling efforts. Variables with

p-values lower than 0.05 were prioritized during selection, ensuring stronger evidence of association with price and enhancing the robustness of the predictive model.

3.6 Modelling

The study, we employed a range of modelling techniques to analyze and predict property prices, utilizing both spatial and non-spatial approaches. The primary aim of the modelling strategy was to accurately capture the relationships between property prices and various influential factors, including both intrinsic property characteristics and spatial dependencies. To achieve this, the study applied non-spatial and spatial Random Forest (RF) models, along with Geographically Weighted Regression (GWR) and Spatial Autoregressive (SAR) models. This multi-faceted approach allows us to leverage the strengths of different methodologies to provide a comprehensive understanding of the factors driving property prices and to account for spatial heterogeneity and autocorrelation in our predictions. The following sections detail the specific modelling techniques employed, the processes for fitting each model, and the methods used to evaluate their performance.

3.6.1 Spatial autoregression(SAR)

The Spatial autoregressive (SAR) model was designed to account for spatial dependencies in the data by incorporating the spatial relationships among observations directly into the regression framework. This model was particularly useful when the residuals of a standard regression model exhibited spatial autocorrelation, indicating that the assumption of independence among observations was violated.

Model specification

The general form of the SAR model was expressed as:

$$y = \rho W y + X\beta + \varepsilon \tag{3.1}$$

where:

- y was the dependent variable (property prices in our case).
- ρ was the spatial autoregressive parameter that captured the strength of the spatial dependence.
- W was the spatial weights matrix, which defined the spatial structure of the data (i.e., which observations were considered neighbors and how strongly they influenced each other).
- Wy was the spatial lag of the dependent variable.
- X was the matrix of explanatory variables (predictors such as swimming pool, bedrooms, etc.).
- β was the vector of coefficients for the explanatory variables.
- ε was the error term, assumed to be normally distributed with mean zero and constant variance.

Spatial weights matrix

The spatial weights matrix W was a crucial component of the SAR model. It captured the spatial structure of the data by defining how each observation was related to others in space. There were several ways to construct W, such as using distance-based criteria or contiguity-based criteria. For this study, we used a distance-based weights matrix, where neighbors within a certain distance threshold were assigned non-zero weights.

Model estimation

The SAR model parameters were typically estimated using Maximum Likelihood Estimation (MLE). The log-likelihood function for the SAR model was given by:

$$\log L(\beta, \rho, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) + \log|I - \rho W| - \frac{1}{2\sigma^2}(y - \rho Wy - X\beta)'(y - \rho Wy - X\beta) \tag{3.2}$$

where:

- n was the number of observations.
- I was the identity matrix.
- $|I - \rho W|$ was the determinant of the matrix $I - \rho W$.

This approach allowed us to effectively account for spatial dependencies in property prices, leading to more accurate and reliable model estimates.

3.6.2 Geographically weighted regression (GWR)

Geographically weighted regression (GWR) was employed to explore spatial heterogeneity by allowing the relationships between the dependent and independent variables to vary over space. Unlike traditional regression models that assume a global relationship, GWR captured local variations, providing a more nuanced understanding of spatial data.

Model specification

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \tag{3.3}$$

where:

- y_i was the dependent variable at location i.

- $\beta_0(u_i, v_i)$ was the intercept at location (u_i, v_i).
- $\beta_k(u_i, v_i)$ were the local regression coefficients for the k-th predictor variable at location (u_i, v_i).
- x_{ik} were the predictor variables at location i.
- ε_i was the error term at location i.

Spatial weights

The local coefficients $\beta_k(u_i, v_i)$ were estimated using spatial weights that decreased with distance from the location i. A kernel function was used to define these weights, ensuring that closer observations had more influence on the estimation of local parameters. Common kernel functions included Gaussian and bi-square.

Bandwidth selection

The bandwidth, which determined the range of influence of the spatial weights, was a critical parameter in GWR. It controlled the trade-off between bias and variance. A smaller bandwidth captured finer spatial variations but increased variance, while a larger bandwidth smoothed out local variations but could introduce bias. The optimal bandwidth was typically selected using cross-validation or an information criterion such as AIC.

Model estimation

The parameters of the GWR model were estimated by solving a set of weighted least squares problems, one for each location i. The weighted least squares estimator was given by:

$$\hat{\beta}(u_i, v_i) = \left(X^T W_i X\right)^{-1} X^T W_i y \tag{3.4}$$

where:

- X was the matrix of predictor variables.
- y was the vector of observed values.
- W_i was the diagonal matrix of spatial weights for location i.

This methodology allowed us to account for spatial heterogeneity, providing insights into how the relationships between property prices and their predictors varied across different locations.

3.6.3 Spatial random forest

Spatial random forest (SRF) extends the traditional Random Forest algorithm to handle spatially correlated data. It leverages the power of ensemble learning to model complex spatial relationships while incorporating spatial autocorrelation.

Model specification

The spatial random forest model extends the standard Random Forest model by incorporating spatial coordinates or neighborhood information as additional features. The general form of the model remains similar to the traditional Random Forest:

$$y = f(x) + \varepsilon \tag{3.5}$$

where:

- y is the dependent variable (prices).
- $f(x)$ is the function learned by the model.
- x is the vector of predictor variables (including spatial coordinates or neighborhood information).
- ε is the error term.

Spatial weights

In spatial random forest, spatial weights are introduced to account for spatial autocorrelation. These weights capture the spatial relationships between observations, allowing the model to learn from neighboring data points. Various methods can be used to define spatial weights, such as distance-based weights or contiguity-based weights.

Model training

The training process of spatial random forest involves growing an ensemble of decision trees using the training dataset. At each node of the tree, a random subset of predictor variables is considered for splitting, leading to diverse trees. The splitting criterion, often based on measures like Gini impurity or information gain, is chosen to maximize predictive accuracy.

Prediction

During prediction, each tree in the ensemble provides a prediction for the target variable. The final prediction is then obtained by aggregating the predictions of all trees. For regression tasks, this aggregation is typically done by averaging the predictions, while for classification tasks, it may involve voting or averaging class probabilities.

This methodology allows for the creation of accurate and interpretable models for spatially correlated data, making it suitable for various applications in spatial analysis and environmental modeling.

3.6.4 Non-spatial random forest methodology

Non-spatial random forest (RF) is a powerful ensemble learning technique widely used for regression and classification tasks. It constructs multiple decision trees during training and combines their predictions to improve accuracy and generalization.

Model specification

The non-spatial random forest model aims to predict a target variable y based on a set of predictor variables x. The general form of the model is:

$$y = f(x) + \varepsilon \tag{3.6}$$

where:

- $f(x)$ is the function learned by the model, represented by an ensemble of decision trees.
- ε is the error term.

Model training

The training process involves constructing an ensemble of decision trees. Each tree is built using a bootstrap sample of the training data and a random subset of predictor variables at each split. The splitting criterion, often based on measures like Gini impurity or information gain, is chosen to maximize predictive accuracy.

Prediction

During prediction, each tree in the ensemble provides a prediction for the target variable. The final prediction is obtained by aggregating the predictions of all trees. For regression tasks, this aggregation is typically done by averaging the predictions, while for classification tasks, it may involve voting or averaging class probabilities.

3.7 Evaluation metrics

Root mean square error (RMSE) is a machine learning evaluation metric used to assess model performance. RMSE, on the other hand, is similar to the mean square error (MAE). Whereas all errors in MAE have the same weight, RMSE penalizes variation, giving greater weight to errors with large absolute values than those with small absolute values. As a result, when RMSE and MAE are computed, RMSE is always greater than MAE. Because RMSE is more sensitive to errors than MAE, it is preferable for measuring performance. RMSE is determined as the square root of the sum of squared errors

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.7}$$

The equation shows that when the total of squared errors approaches zero, the RMSE approaches zero. As a result, when RMSE is 0, there are no mistakes between the actual value y and the anticipated value y. R-squared, also known as coefficient determination, is a statistical metric used in machine learning to determine how near the data is to the fitted regression line. R-squared is assigned a value between 0 and 1. Where 1 denotes perfect, and 0 denotes imperfection. Furthermore, R-squared is determined by dividing the deviations of the observations from their anticipated values by the variances of the observations from their mean.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.8}$$

3.8 Hyperparameter tuning

A systematic approach was adopted to optimize the model's performance in the hyperparameter tuning for the Random Forest Regressor. Initially, a parameter grid was defined, encompassing a range of hyperparameters known to significantly influence the Random Forest's predictive ability. This grid included parameters such as the number of estimators

(trees) in the forest, maximum depth of the trees, criteria for splitting nodes, and minimum samples required to split internal nodes or form leaf nodes. This exhaustive grid aimed to explore various combinations of hyperparameters to identify the configuration that yields the best model performance. Subsequently, GridSearchCV, a cross-validated grid search, was employed to systematically evaluate each combination of hyperparameters using a 5-fold cross-validation strategy. The objective was to find the combination that minimizes the negative mean squared error, serving as the scoring metric. By iterating through the parameter grid and evaluating each configuration's performance, GridSearchCV determined the optimal hyperparameters for the Random Forest Regressor. Following the hyperparameter tuning process, the best set of hyperparameters was identified, representing the configuration that yielded the highest model performance. These optimal hyperparameters were then utilized to initialize a new Random Forest Regressor instance. Subsequently, the model was trained on the training dataset using the best hyperparameters to leverage the entire dataset for model fitting

3.9 Deployment

The deployment process of a machine learning model in a web interface involved several sequential steps to ensure smooth integration and user accessibility. Initially, the trained machine learning model was prepared and saved in a compatible format. Subsequently, a web application file was created to host the model. Within this file, the model was loaded into memory. User input components were then designed to enable users to input data for predictions, employing features like text inputs, sliders, or dropdown menus. Upon receiving user input, the model performed predictions based on the provided data. The results were then presented back to the user within the web interface.

Chapter 4

Analysis, results and interpretations

4.1 Introduction

This chapter presents an empirical investigation into predicting residential property values(price) within the Nairobi metropolitan area. The methodology involved robust data acquisition and preprocessing procedures to ensure dataset integrity and reliability. Machine learning algorithms were then employed to construct predictive models capable of accurately estimating house prices. Key determinants, including structural attributes like bedrooms and bathrooms, spatial coordinates, and amenities such as swimming pools or generators, were scrutinized for their impact on property valuations. The section unveils findings from empirical analyses, clarifying patterns and correlations in Nairobi's real estate market dynamics through examination and statistical inference. Through thorough examination and statistical inference, employing tests such as the Wilcoxon rank-sum test, this section unveils insights into patterns and correlations within Nairobi's real estate market dynamics, shedding light on factors influencing house prices.

4.2 Characteristics of predictor variables for property housing

In this analysis, the study investigated the significance of various variables to the target variable, the price of properties. The Table4.1 summarizes the significance of both numerical and categorical variables, categorizing them based on their levels.

Table 4.1: Characteristics of predictor variables (covariates) and the house prices

Variable	Test-value	Df	SE	P-value
Chi squared test				
House type	283.54	5	32.23	<0.001
Area type	155.95	2	24.22	<0.001
Wilcoxon rank-sum test				
Gym	1181610	-	27257.97	0.175
Swimming pool	926442.5	-	25318.07	<0.001
Internet fibre	1479892.0	-	34514.03	<0.001
Generator	921531.5	-	32157.33	0.162
Borehole	495693.0	-	26806.87	<0.001
Staff quarters	309234.5	-	21344.49	<0.001
Spearman's test				
Size	0.6127391	-	0.01447372	<0.001
Bedrooms	0.6135400	-	0.01194080	<0.001
Bathrooms	0.5726622	-	0.01138617	<0.001

In the analysis of predictor variables for property housing, each numeric variable's correlation with the target was carefully examined using the Speakman coefficient, standard errors (SE), and p-values. Size emerged as a significant predictor, demonstrating a strong positive correlation with property prices (Speakman coefficient = 0.6127391). This indicates that larger-sized properties tend to command higher prices. The precise correlation estimate for size, reflected by a low standard error of 0.01447372, underscores the reliability of this relationship. Moreover, the extremely low p-value (<0.001) emphasizes the significant association between property size and price, highlighting its importance as a determinant of property value.

Similarly, bedrooms exhibited a high positive correlation with the target variable, with a Speakman coefficient of 0.6135400. Properties with more bedrooms were found to command higher prices. The low standard error (0.01194080) suggests high precision in the correlation estimate, while the exceptionally low p-value (<0.001) underscores the crucial role of the number of bedrooms in determining property prices. Likewise, bathrooms showed a positive correlation with the target, with a Speakman coefficient of 0.5726622, indicating that properties with more bathrooms are associated with higher prices. The relatively precise correlation estimate for bathrooms (standard error = 0.01138617) and the low p-value (<0.001) further support the significant relationship between the number of bathrooms and property prices.

These findings indicate that size, bedrooms, and bathrooms are strongly correlated with property prices, with low standard errors and extremely low p-values, suggesting their influential roles as determinants of property prices.

The presence of amenities like swimming pools, boreholes, and staff quarters demonstrated substantial associations with property prices, supported by their high Wilcoxon statistics and low p-values (<0.001). However, gyms and generators did not show statistically significant relationships with property prices, despite some association.

In assessing categorical variables with three or more levels, the chi-squared test was employed to determine their association with the target variable (price). Both the type of house and type of neighbourhood exhibited notable associations with property prices. House type displayed a strong association, with a high chi-squared statistic (283.5417), low standard error (32.76150), and an extremely low p-value (3.451704e-59). Similarly, area type showed a significant association, with a substantial chi-squared statistic (155.9456), low standard error (24.19431), and an extremely low p-value (1.370360e-34). These results underscore the importance of both type of house and area type in property pricing models, with variables having p-values less than 0.05 being included in the predictive model due to their significant associations with property prices.

4.3 Modelling

In this section, the study delved into the application of various predictive modelling techniques to estimate house prices. After identifying significant predictor variables in the preceding analysis, including both numerical attributes as well as categorical features, we proceeded to construct predictive models using advanced statistical and machine learning methodologies.

Before applying these models, the study assessed the presence of spatial autocorrelation in the data, as property prices are often influenced by their geographic location. To evaluate this, Moran's I test, which measures spatial autocorrelation was conducted. The test yielded a Moran's I statistic of 0.4514027 with a p-value of 0.001, indicating a significant level of spatial dependence among property prices in the dataset. This result justified the use of spatial modelling techniques to capture the inherent spatial relationships. The aim was to explore the effectiveness of different modelling approaches in capturing the complex dynamics of Nairobi's real estate market and accurately predicting property prices. A range of techniques tailored to handle spatial dependencies and heterogeneity inherent in real estate data were applied. These techniques included Spatial Autoregression, which incorporates information from neighboring properties into the predictive process; Geographically Weighted Regression (GWR), which allows model parameters to vary spatially to reflect localized relationships; Spatial Random Forest, which extends the traditional random forest algorithm by incorporating spatial information; and Non-Spatial Random Forest, serving as a baseline model to compare the impact of accounting for spatial dependencies. Plots for the training set are shown in Figure A.3

4.4 Spatial autoregression

We fitted a Spatial Autoregression (SAR) model using a distance-based spatial weights matrix to account for spatial dependencies in property prices. This matrix was created using the `dnearneigh` function, specifying a distance range of 0 to 5 units, which ensures that only

neighbors within this range are considered, excluding self-neighbors. The resulting spatial weights were converted to binary weights. The Table 4.2 shows the summary of the SAR model.

Table 4.2: Regression results for house price prediction

Variable	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)	9320242.5	1828880.0	5.0961	<0.001
swimming_pool	4373200.1	347179.2	12.5964	<0.001
bedrooms	4292530.3	358611.6	11.9699	<0.001
borehole	-1088260.9	576256.8	-1.8885	0.06
staff_quarters	6081957.9	801499.8	7.5882	<0.001
bathrooms	851842.1	225387.6	3.7795	<0.001
house_type_encoded	-1115998.2	206035.0	-5.4165	<0.001
size	32484.1	2902.5	11.1916	<0.001
area_type_encoded	-4479697.8	575540.8	-7.7835	<0.001

The SAR model indicates several significant predictors of property prices within the Nairobi metropolitan area. The presence of a swimming pool and the number of bedrooms both show a strong positive association with property prices, suggesting that properties with these features tend to be more expensive. Similarly, the availability of staff quarters significantly increases property prices. The number of bathrooms and property size also positively correlate with higher prices. Different house types and certain area types significantly negatively impact property prices. Although the presence of a borehole shows a negative estimate, it is not statistically significant at the 5% level.

The significant negative lambda value ($\lambda = -0.94617$, p-value = 6.0908e-07) indicates strong spatial error dependence, further validating the use of spatial autoregression in capturing the spatial dynamics of property prices. The model fit statistics, including a lower AIC (111770 compared to 111800 for a linear model), suggest that the SAR model provides a better fit for the data by accounting for spatial dependencies.

Overall, the SAR model effectively captures the influence of various predictors on property prices in Nairobi, accounting for both spatial dependencies and significant variables identified in the analysis.

4.5 Geographically weighted regression

Table 4.3: Summary of GWR coefficient estimates

Variable	Min.	1st Qu.	Median	3rd Qu.	Max.
Intercept	-55452519.0	-4588953.0	636993.4	10770845.6	81441414
swimming_pool	1637679.5	2266760.1	2509406.8	2750026.3	8786428
bedrooms	737944.1	1594630.5	2146628.0	2380931.8	6063696
borehole	-5443811.8	510161.5	2095289.3	2405707.4	3496178
staff_quarters	-7963498.9	-1330450.7	961726.2	3591603.2	9766558
bathrooms	36008.6	163558.4	249632.5	1329295.5	3602471
house_type_encoded	-1742335.5	511973.5	2407044.2	3147826.8	5195998
size	-2608.3	43928.3	53859.4	59739.6	68211
area_type_encoded	-29636903.3	-4525676.3	-903264.4	282464.8	19145751

The Table 4.3 presents a summary of GWR model. The Akaike Information Criterion (AIC) value for the model is 109480, indicating its goodness of fit relative to its complexity. Lower AIC values suggest better model fit. The Bayesian Information Criterion (BIC) value is 106698.9, providing another measure of model fit and complexity. As with AIC, lower BIC values indicate better model fit. The residual sum of squares, a measure of the discrepancy between the observed and predicted values, is extremely high at 1.259942×10^{17}, indicating a significant amount of unexplained variance in the model. The R-squared value of 0.7221171 suggests that approximately 72.2% of the variance in the target variable is explained by the predictor variables in the model. The adjusted R-squared value, which adjusts for the number of predictor variables in the model, is 0.71515. This value is slightly lower than the R-squared value, indicating a penalty for the inclusion of additional predictors.

4.6 Spatial random forest

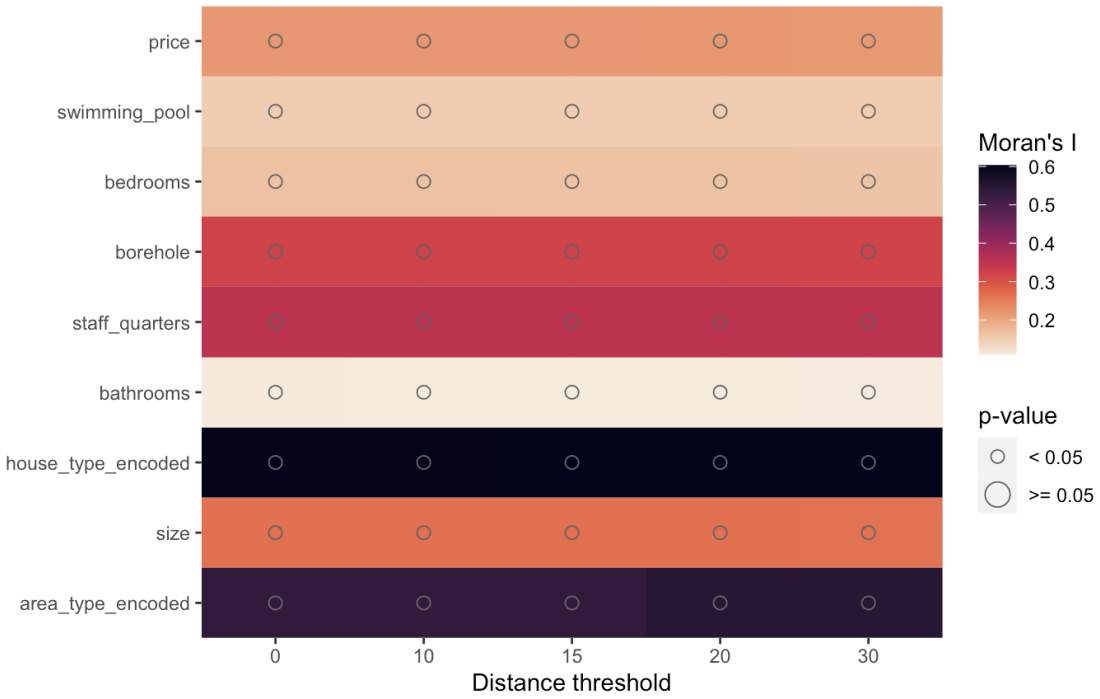


Figure 4.1: Spatial autocorrelation of the response variable and predictors across distance

Low Moran's I and p-values equal or lower than 0.05 indicate that there is spatial autocorrelation for the given variable and distance threshold as shown in Figure 4.1

The spatial random forest model's performance is evaluated using several metrics, the results are presented in Table 4.4 The out-of-bag (OOB) R-squared value is 0.6025854, which indicates the proportion of variance explained by the model on the OOB samples (samples not used in training). An OOB R-squared of 0.603 suggests that the model explains approximately 60.3% of the variability in the response variable on the unseen OOB data. Additionally, the R-squared value derived from the square of the correlation coefficient between the observed and predicted values is 0.8199942. This implies that about 82.0% of the variability in the observed data is captured by the model's predictions. The pseudo R-squared, the correlation between observed and predicted values (not squared), is 0.8902243. A value of 0.9055353 signifies a strong positive correlation, indicating that the model's

predictions are highly correlated with the actual observed values. The residuals for this model are shown in Figure A.2

Table 4.4: Spatial random forest model performance metrics

Metric	Value
R squared (OOB)	0.6025854
R squared (cor(obs, pred)^2)	0.8199942
Pseudo R squared (cor(obs, pred))	0.9055353
RMSE (OOB)	7499288
RMSE	5284978
Normalized RMSE	0.4065368

4.7 Non spatial random forest

The model's performance is evaluated using several metrics. The out-of-bag (OOB) R-squared value is 0.6980193, which indicates the proportion of variance explained by the model on the OOB samples (samples not used in training). An OOB R-squared of 0.698 suggests that the model explains approximately 69.8% of the variability in the response variable on the unseen OOB data. Additionally, the R-squared value derived from the square of the correlation coefficient between the observed and predicted values is 0.7924993. This implies that about 79.25% of the variability in the observed data is captured by the model's predictions. Furthermore, the Pseudo R-squared, which is the correlation between observed and predicted values (not squared), is 0.8902243. A value of 0.8902 signifies a strong positive correlation, indicating that the model's predictions are highly correlated with the actual observed values. This summary is illustrated in Table 4.5

Table 4.5: Non SpatialRF performance metrics

Metric	Value
R squared (OOB)	0.6980193
R squared (cor(obs, pred)^2)	0.7924993
Pseudo R squared (cor(obs, pred))	0.8902243
RMSE (OOB)	7499288
RMSE	5284978
Normalized RMSE	0.4065368

The residual diagnostic for non spatial random forest are shown in Figure A.1

4.8 Models comparison

The comparison of model performances shows that the spatial random forest model outperforms the nonspatial random forest model, GWR, and SAR models based on several key metrics. Below is the summary of the performance metrics for each model:

Table 4.6: Comparison of random forest models performance Metrics

Metric	Nonspatial RF	Spatial RF
R squared (OOB)	0.6980193	0.6025854
R squared (cor(obs, pred)^2)	0.7924993	0.8199942
Pseudo R squared (cor(obs, pred))	0.8902243	0.9055353
RMSE (OOB)	6537139	7499288
RMSE	5550987	5284978
Normalized RMSE	0.426999	0.4065368

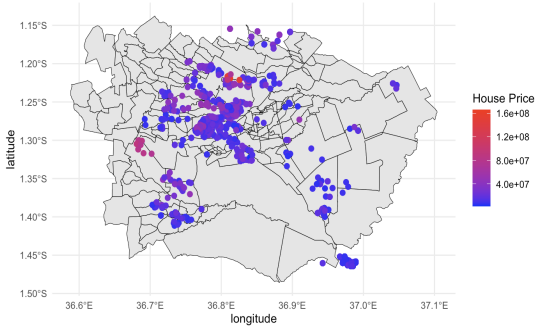
For the Geographically Weighted Regression (GWR) model, the performance metrics indicate a decent fit to the data. The R-squared value is 0.7221171, which means that approximately

72.2% of the variance in the response variable is explained by the model. The adjusted R-squared value is slightly lower at 0.7151545, reflecting the proportion of variance explained after accounting for the number of predictors in the model. These metrics suggest that the GWR model provides a reasonable explanation of the variability in the data, though it is slightly less effective than the spatial random forest model in terms of the R-squared measure.

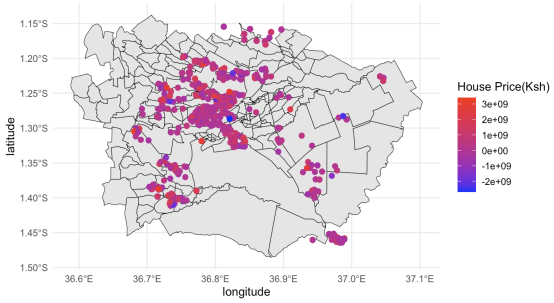
For the Spatial Autoregressive (SAR) model, several key statistics are reported. The Wald statistic is extremely high at 1249800 with a p-value less than 2.22e-16, indicating that the model parameters are highly significant. The log likelihood for the error model is -55876.25, which is a measure of the model's fit to the data, with higher (less negative) values indicating a better fit. The Akaike Information Criterion (AIC) is 111770, which is lower than the AIC for a comparable linear model (111800), suggesting that the SAR model provides a better balance between model complexity and goodness of fit. The model also reports an approximate standard error of 0.00084636 for the numerical Hessian, with a z-value of -1117.9 and a p-value less than 2.22e-16, further supporting the significance of the model. The maximum likelihood residual variance (sigma squared) is 2.3765e+11, with a sigma value of 487490, indicating the spread of residuals.

Overall, while both the GWR and SAR models show strong performance with significant parameters and good fit measures, the spatial random forest model still appears to be the best performer among the compared models based on the combined criteria of R-squared values and RMSE.

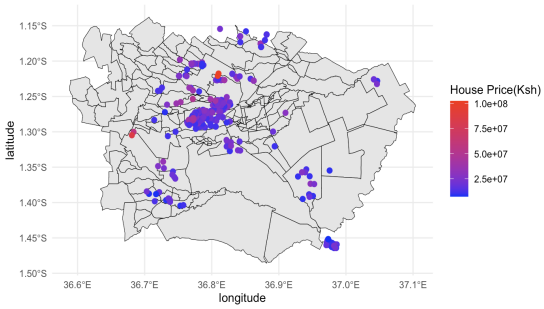
Based on the metrics on Table 4.6, the spatial random forest model demonstrates superior performance, with the highest Pseudo R-squared and the lowest RMSE and Normalized RMSE values, making it the best model among the compared options.



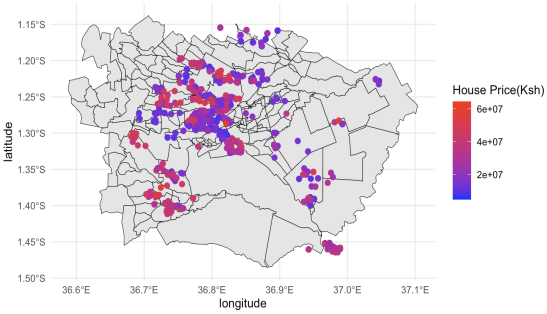
(a) Actual price mapping



(b) SAR predictions



(c) Random Forest predictions



(d) GWR predictions

Figure 4.2: House price plot for actual and model distribution

Figure 4.2 shows house prices ploted in the location of the properties. Panel (a) is a plot of the actual price of houses, panel (b) is plot on predictions made by SAR, model, the plots are very different from the actual, GWR model prediction in panel (d) the plot is an improvement from the SAR model but predicted prices are higher as compared to the actual plot. Panel (c) are predictions from the spatial random forest model, the plots are almost similar to the actual price map.

4.9 Important features

Based on the top five important predictors identified for both the non-spatial and spatial random forest models illustrated in Figure 4.3, it is evident that certain features carry significant weight in predicting the target variable. In the non-spatial model, the most influential predictors include the size of the property, the number of bedrooms, the area type,

the type of house, and the number of bathrooms. These features align with conventional wisdom in real estate, where factors such as property size, bedroom count, and bathroom amenities often play crucial roles in determining property value and desirability.

On the other hand, the spatial model highlights similar predictors such as property size, bedroom count, house type, and bathrooms, with the addition of staff quarters as a notable predictor. The inclusion of staff quarters in the spatial model suggests that factors related to auxiliary accommodation or additional amenities may have spatial dependencies that influence the target variable. This underscores the importance of considering spatial relationships and neighborhood characteristics when predicting property values or other related outcomes.

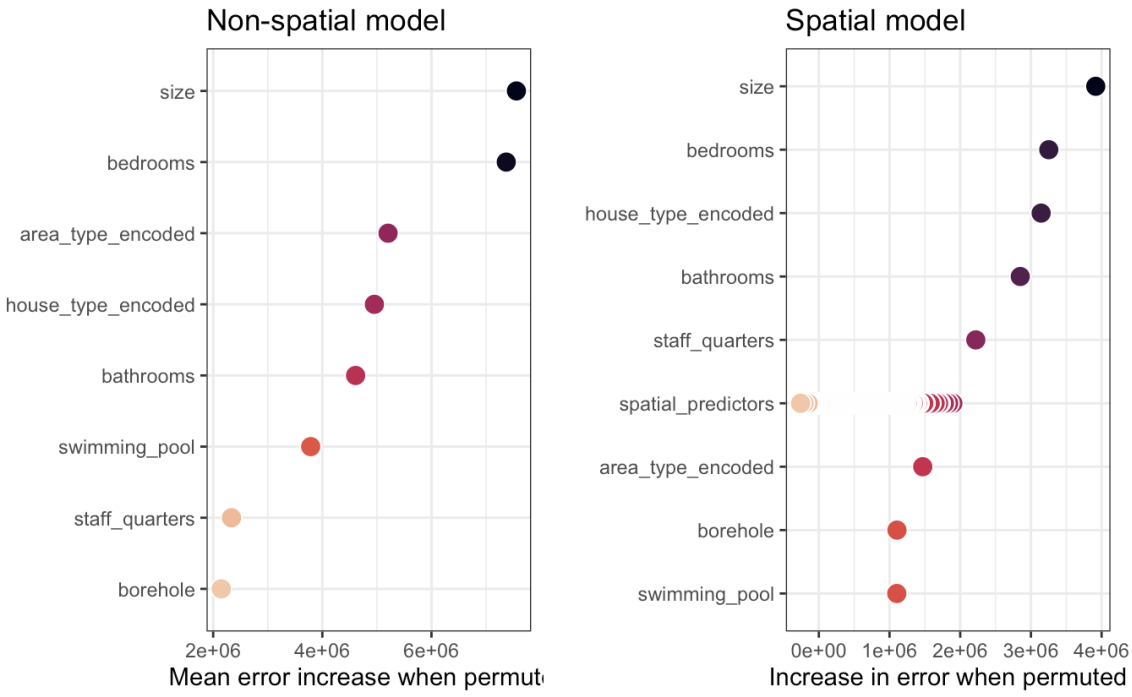


Figure 4.3: Important features

The Geographically weighted regression (GWR) model reveals insightful patterns regarding the factors influencing property prices. Among the top predictors identified, the presence of a swimming pool emerges as the most influential, with properties boasting this amenity experiencing a substantial increase in price. Following closely behind, the number of bedrooms and the inclusion of staff quarters also significantly contribute to higher property

valuations. Additionally, the size of the property emerges as a crucial determinant, with larger properties commanding higher prices.

The identification of these top predictors underscores the value of both conventional property attributes and spatially dependent features in predicting real estate outcomes. By incorporating such insights into predictive models, stakeholders can gain a deeper understanding of the underlying factors driving property values and make more informed decisions in the real estate domain.

4.10 Model deployment

4.10.1 Deployment architecture

The deployment of the random forest (RF) model is implemented using RShiny, which allows for creating interactive and user-friendly web applications. The architecture consists of the following components illustrated in flowchart in Figure4.4

1. **RShiny server**: Hosts the RShiny application, handling user interactions and executing the RF model.
2. **Web application (RShiny UI)**: Provides an interactive interface for users to upload data, view predictions, monitor model performance, and provide feedback.
3. **Database**: PostgresSQL stores input data, prediction results, and user interactions.

The architecture is designed to ensure scalability, reliability, and ease of use.

4.10.2 User interfaces

The deployment includes several user interfaces to enable different types of interactions tailored to the needs of policymakers, real estate investors, and home buyers:

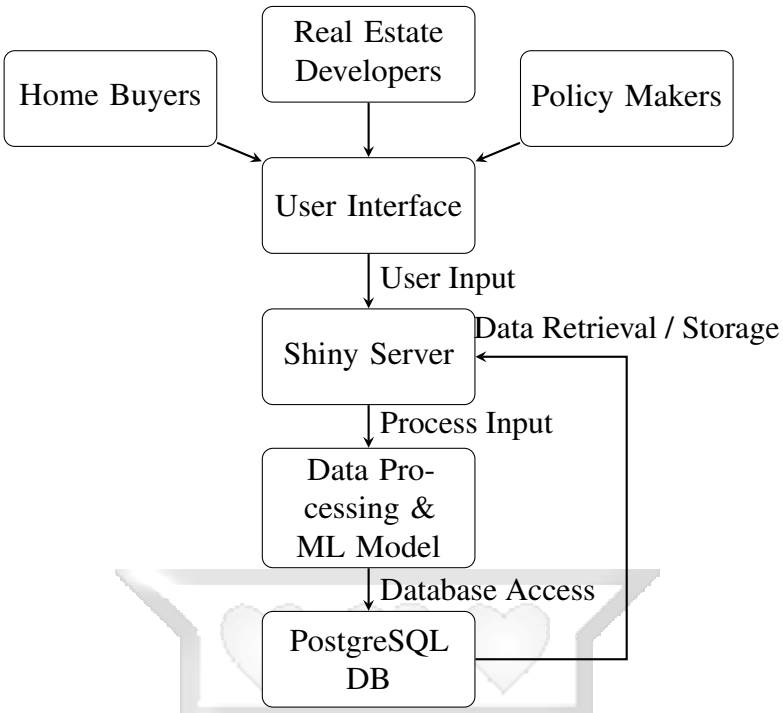


Figure 4.4: Flowchart illustrating the interaction between different components of the system.

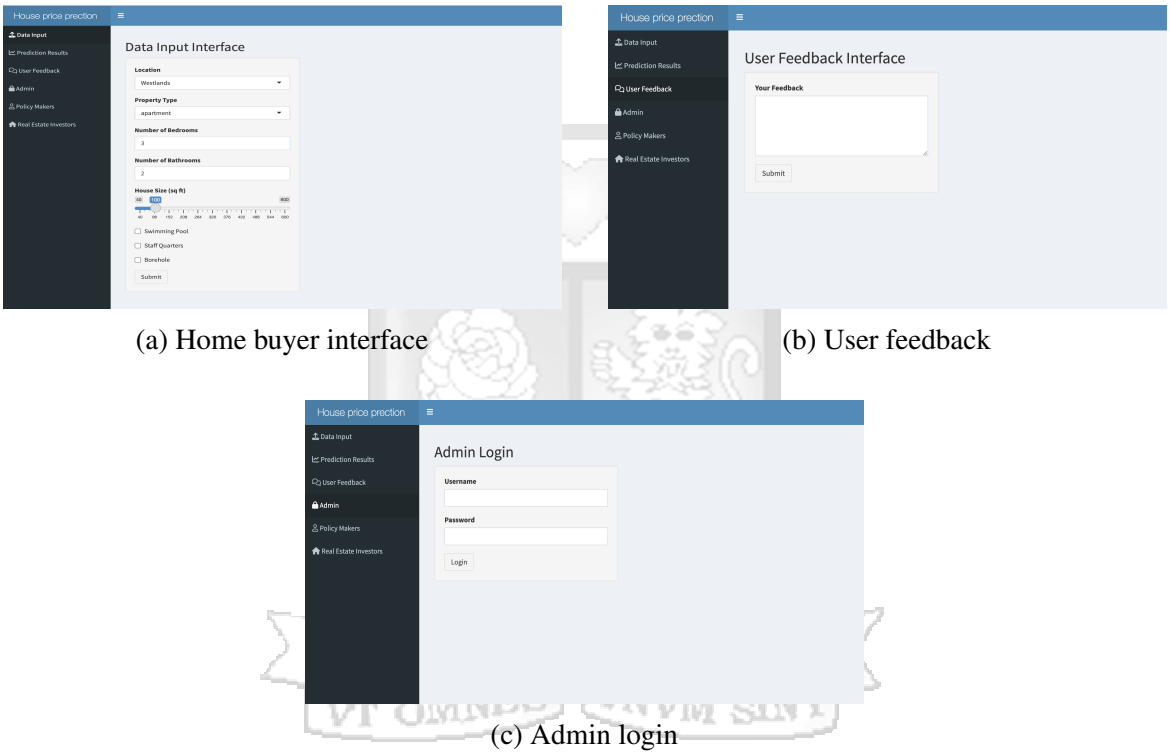
Data Input interface: This interface allows users to input spatial data for prediction by selecting locations, amenities, and house features directly from the application. Users can specify various parameters through interactive forms. Policymakers can input information on urban planning and infrastructure projects. Real Estate Investors can use the data input by Home Buyers to identify where more people are looking to buy houses and the features most sought after. Home Buyers enter their preferences for desired home features and neighborhoods to find suitable properties.

Prediction results interface: The Prediction results interface displays the predictions made by the spatial RF model, including the predicted price for properties at around 10 different locations represented by coordinates. Additionally, a map is provided to illustrate the predicted prices at these locations.

User feedback interface: The User feedback interface allows users to provide feedback on the predictions generated by the model. Policymakers can utilize this feature to gather public input on the impacts of policies and community needs. Real estate investors benefit by collecting insights on market demand and preferences from potential buyers and renters.

Home buyers can share also feedback on property predictions, contributing to improving model accuracy.

Admin section (Model monitoring interface): Restricted to internal use, this section provides insights into model performance and usage statistics. It is initially hidden and only accessible to authenticated users. this section provides insights into model performance and ensures data integrity across user interactions. Admin ensures the reliability and accuracy of predictions for all users, maintaining trust in the system.

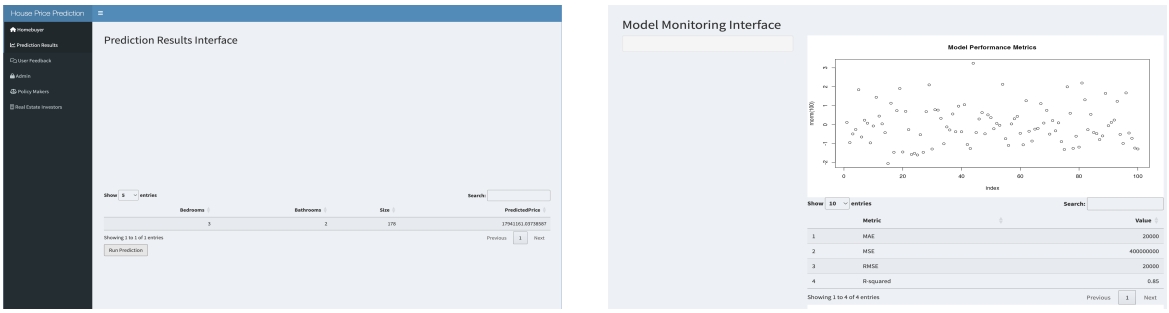


(a) Home buyer interface (b) User feedback

(c) Admin login

Figure 4.5: Data input interfaces

Figure 4.5 shows data input interfaces, panel (a) represents home buyer interface, here the buyer enters features of a house they want to buy. Panel (b) is where the users can give feedback. The last panel (c) is admin only, they need to have login to access model performance interface.



(a) Prediction results interface

(b) Model monitoring

Figure 4.6: Prediction and model monitoring interfaces

Figure 4.6 shows output interfaces. Panel (a) represents the prediction interface. Based on user input, the model gives predicted house price based on user input. Panel (b) represents an interface of model performance that is only accessible to the admin, it displays evaluation metrics of predicted prices.

Chapter 5

Discussions, conclusions and recommendations

5.1 Introduction

The objective of this research was to develop predictive models for house pricing in the Nairobi Metropolitan area, utilizing both spatial and non-spatial approaches. Additionally, the study aimed to assess the performance of these models in comparison to spatial autoregression (SAR) and geographically weighted regression (GWR) techniques. The research sought to identify key variables influencing house prices and determine whether spatial relationships play a significant role in pricing dynamics. To achieve these objectives, a comprehensive dataset was curated, comprising property characteristics, location attributes, and socio-economic factors. The data underwent rigorous preprocessing, including spatial aggregation and feature engineering, to enhance model performance. Machine learning techniques, such as Random Forest and Spatial Random Forest, were applied to develop predictive models, alongside traditional spatial regression methods like SAR and GWR. Through this study, insights were sought into the effectiveness of spatial modeling techniques in capturing local variations and improving predictive accuracy for house pricing in the Nairobi Metropolitan area.

5.2 Discussions

In our study focused on predicting house prices in the Nairobi Metropolitan Area, we employed a range of machine learning models to construct predictive frameworks aimed

at capturing the intricate dynamics of the local real estate market. Among these models, the spatial random forest emerged as the most efficacious tool for evaluation, consistently outperforming alternative methodologies. By prioritizing spatial dependencies and localized patterns, the spatial random forest model exhibited superior performance, particularly in terms of the coefficient of determination (R-squared), a pivotal metric for gauging predictive accuracy (Feng and Humphreys, 2018). Our analysis unveiled that the spatial random forest model attained higher R-squared values compared to competing approaches, underscoring its robustness in encapsulating the variability in house prices across diverse neighborhoods within the metropolitan area.

Venturing deeper into the determinants of house prices, our study identified several key features exerting substantial influence on property values in the Nairobi Metropolitan Area. Through a comprehensive examination of the dataset, we discerned that factors such as property size, number of bedrooms, and area type wielded significant impacts on house prices. These findings align with extant literature on real estate economics, which underscores the significance of both structural attributes and neighborhood characteristics in shaping property values (Feng and Humphreys, 2018). Furthermore, our study shed light on the spatial dimension of these relationships, emphasizing the necessity to consider not only the intrinsic attributes of a property but also its external environment when predicting house prices accurately.

Our study's findings resonate strongly with the spatial perspective articulated by the theory of Spatial Autoregressive Models, as discussed in existing literature. By implementing spatial models, our study acknowledges and quantifies the influence of proximity and neighborhood characteristics on housing prices (Alshammari, 2023). The identification of spatially dependent features in our predictive models accentuates the localized price patterns observed in the Nairobi Metropolitan Area. Moreover, our study underscores the importance of incorporating spatial dependencies into predictive models, providing empirical evidence that bolsters the theoretical understanding of housing price dynamics proposed by Spatial Autoregressive Models.

Furthermore, our study not only affirmed the significance of proximity and neighborhood characteristics but also pinpointed several key features pivotal in predicting house prices. Among these, the number of bedrooms, property size, and area type emerged as fundamental determinants of house pricing. This concurs with findings from analogous studies conducted in Seoul, South Korea, and Binhu New District, Hefei City, Anhui Province, China (**?**). Bringing the analysis closer to home, a study conducted by Ndegwa in Nairobi's business district identified additional factors such as the presence of a swimming pool, apartment size, and location as crucial features influencing house prices (Ndegwa, 2018). These findings not only reinforce the significance of key features identified in other studies but also furnish valuable insights into the specific context of the Nairobi Metropolitan Area's real estate market.

5.3 Limitations

One significant limitation of our study was the challenge of obtaining comprehensive data from all estates within the Nairobi Metropolitan Area. This limitation restricted the representativeness of our dataset and may have introduced bias into our analysis. Consequently, our findings may not fully capture the diversity of housing market dynamics across the entire metropolitan area.

5.4 Recommendations

In conclusion, our study underscores the need for further exploration of the factors influencing house prices in the Nairobi Metropolitan Area. To enhance the depth of our understanding, future research should consider incorporating additional variables such as proximity to amenities, transportation infrastructure, and environmental quality. By expanding the scope of analysis, researchers can provide a more comprehensive picture of the housing market dynamics in the region.

Moreover, we recommend the adoption of primary data collection methods to ensure the quality and reliability of the dataset. Gathering firsthand information from residents, real estate professionals, and local authorities can offer unique insights into housing preferences, market trends, and neighborhood dynamics. This approach will not only enrich the dataset but also provide a deeper understanding of the nuanced factors driving property values.

Furthermore, integrating advanced data collection techniques such as surveys, interviews, and on-the-ground observations can offer qualitative context to complement quantitative analysis. By combining both quantitative and qualitative approaches, researchers can uncover hidden patterns, identify emerging trends, and provide more robust recommendations for stakeholders in the real estate sector.

5.5 Conclusions

In conclusion, our study delved into the intricate realm of predicting house prices in the Nairobi Metropolitan Area, employing various machine learning models to construct predictive frameworks. Through rigorous analysis, we found that the spatial random forest model emerged as the most effective tool for evaluation, showcasing superior performance in capturing the variability in house prices across diverse neighborhoods within the metropolitan area. By prioritizing spatial dependencies and localized patterns, this model provided robust insights into the dynamics of the real estate market.

The study identified several key features that significantly influence property values in the Nairobi Metropolitan Area, including property size, number of bedrooms, and area type. These findings underscored the importance of both structural attributes and neighborhood characteristics in shaping property values, aligning with existing literature on real estate economics.

Importantly, our study highlighted the spatial dimension of these relationships, emphasizing the need to consider not only the intrinsic attributes of a property but also its external environment when predicting house prices accurately. By incorporating spatial dependen-

cies into predictive models, we provided empirical evidence that supported the theoretical understanding of housing price dynamics proposed by Spatial aspects.

References

Alshammari, T. O. (2023). Evaluating machine learning algorithms for predicting house prices in saudi arabia. *2023 International Conference on Smart Computing and Application (ICSCA)*, pages 1–5.

Antonides, G. and Ranyard, R. (2017). Mental accounting and economic behaviour. *Economic psychology*, pages 123–138.

Arundel, R. (2017). Equity inequity: Housing wealth inequality, inter and intra-generational divergences, and the rise of private landlordism. *Housing, Theory and Society*, 34(2):176–200.

Badini, S., Regondi, S., and Pugliese, R. (2023). Unleashing the power of artificial intelligence in materials design. *Materials*, 16(17):5927.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *Journal of economic perspectives*, 27(1):173–196.

Chancellor, W. J., Abbott, M., and Carson, C. (2016). A study of the factors influencing residential house prices in auckland and new zealand. *Journal of Applied Business Research*, 14:55.

Choy, L. H. T. and Ho, W. K. (2023). The use of machine learning in real estate research. *Land*.

de Aquino Afonso, B. K., Melo, L. C., Oliveira, W., da Silva Sousa, S. B., and Berton, L. (2019). Housing prices prediction with a deep learning and random forest ensemble. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2019)*.

Dong, D. (2020). Factors influencing housing prices: An empirical analysis from new york city. In *Proceedings of the 2020 3rd International Conference on E-Business, Information Management and Computer Science*, pages 106–113.

Feng, X. and Humphreys, B. (2018). Assessing the economic impact of sports facilities on residential property values: A spatial hedonic approach. *Journal of Sports Economics*, 19(2):188–210.

Gallagher, R. M., Kurban, H., and Persky, J. J. (2013). Small homes, public schools, and property tax capitalization. *Regional Science and Urban Economics*, 43(2):422–428.

Habanabakize, T. and Dickason, Z. (2022). Political risk and macroeconomic effect of housing prices in south africa. *Cogent Economics & Finance*, 10.

Heyman, A. V. and Sommervoll, D. E. (2019). House prices and relative location. *Cities*, 95:102373.

Hinterwimmer, F., Lazic, I., Suren, C., Hirschmann, M. T., Pohlig, F., Rueckert, D., Burgkart, R., and von Eisenhart-Rothe, R. (2022). Machine learning in knee arthroplasty: specific data are key—a systematic review. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30(2):376–388.

Hoxha, V., Hoxha, D., and Hoxha, J. (2021). Study of factors influencing apartment prices in prishtina, kosovo. *International Journal of Housing Markets and Analysis*.

Jakubovitz, D., Giryes, R., and Rodrigues, M. R. (2019). Generalization error in deep learning. In *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*, pages 153–193. Springer.

Jia, N., Molloy, R., Smith, C., and Wozniak, A. (2023). The economics of internal migration: Advances and policy questions. *Journal of economic literature*, 61(1):144–180.

Kawaguchi, K., Bengio, Y., Verma, V., and Kaelbling, L. P. (2018). Generalization in machine learning via analytical learning theory. *arXiv preprint arXiv:1802.07426*.

Khanna, T., Palepu, K. G., and Sinha, J. (2015). Strategies that fit emerging markets. In *International Business Strategy*, pages 615–631. Routledge.

Kim, J.-U., Lee, Y., Lee, M., and Hong, S.-Y. (2022). A comparative study of machine learning and spatial interpolation methods for predicting house prices. *Sustainability*.

Lewandowska, G., Taracha, M., and Maciuk, K. (2023). Socio-economic factors associated with house prices. evidence based on key macroeconomic aggregates globally. *Budownictwo i Architektura*.

Myers, G. (2015). A world-class city-region? envisioning the nairobi of 2030. *American Behavioral Scientist*, 59(3):328–346.

Ndegwa, J. N. (2018). Determinants of apartment prices within housing estates of nairobi metropolitan area. *International journal of economics and finance*, 10:104.

Ngu, T. X. (2017). Factors influencing the housing price from developers' perspectives in sibu, sarawak.

Oono, K. and Suzuki, T. (2020). Optimization and generalization analysis of transduction through gradient boosting and application to multi-scale graph neural networks. *Advances in Neural Information Processing Systems*, 33:18917–18930.

Scott, A. J. (2019). City-regions reconsidered. *Environment and Planning A: Economy and Space*, 51(3):554–580.

Veldkamp, L. and Chung, C. (2019). Data and the aggregate economy. *Journal of Economic Literature*.

Vuluku, G., Gachanja, J., et al. (2014). Supply side aspects of residential housing for low income earners in kenya. *Research in Applied Economics*, 6(3):271–286.

Xu, S. and Zhang, Z. (2021). Spatial differentiation and influencing factors of second-hand housing prices: A case study of binhu new district, hefei city, anhui province, china. *Journal of Mathematics*.

Zhang, J. (2022). Research on the prediction and influencing factors of house price based on regression analysis model. *BCP Business & Management*.

Zhang, Y., Zhang, D., and Miller, E. J. (2021). Spatial autoregressive analysis and modeling of housing prices in city of toronto. *Journal of Urban Planning and Development*, 147(1):05021003.

Appendix A

Additional results

A.1 Residual Diagnostics

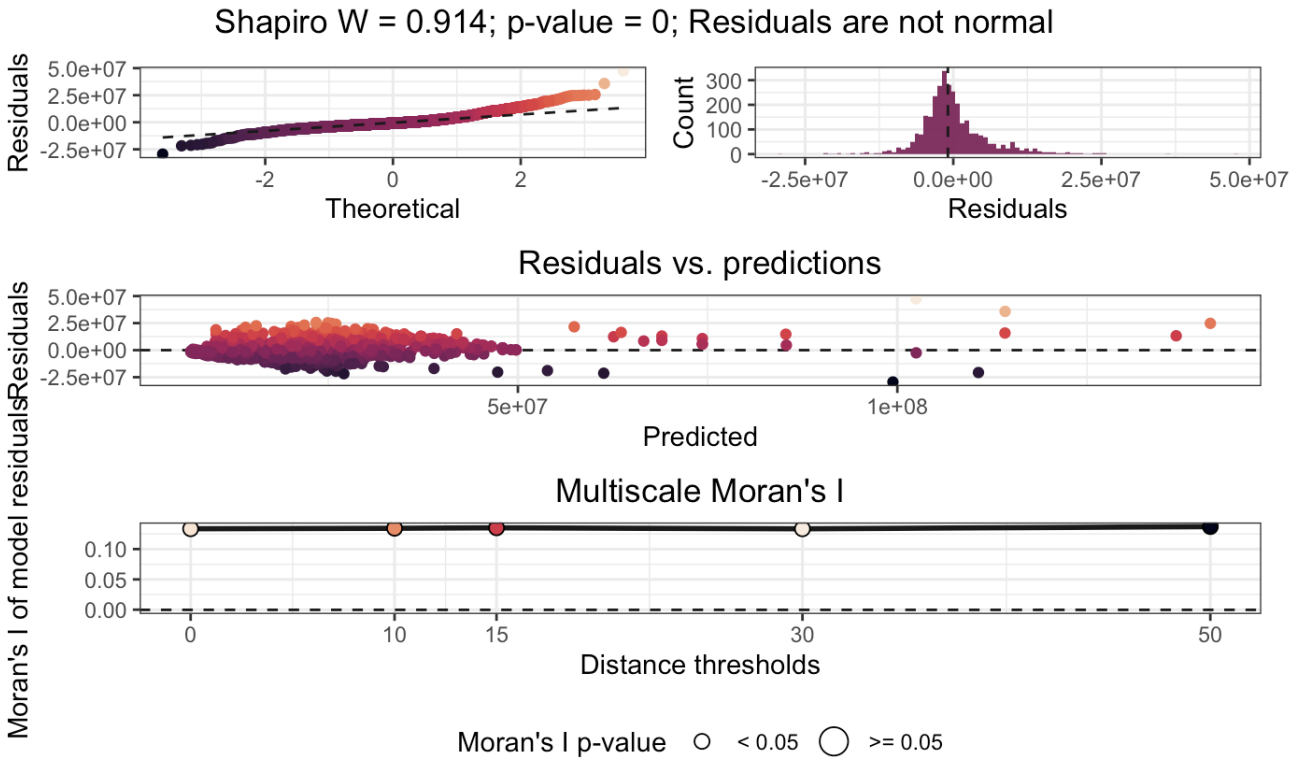


Figure A.1: NonSpatial residual diagnostic

A.2 Residual Diagnostics

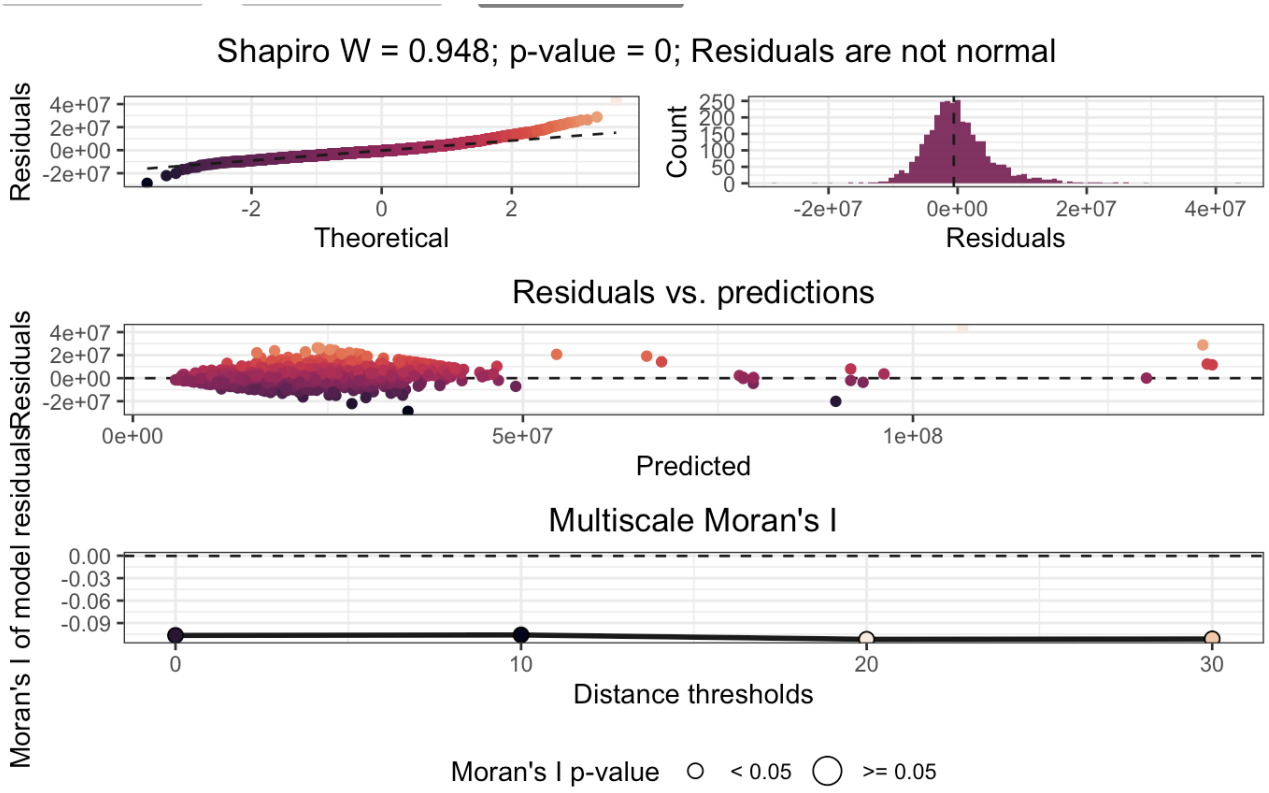


Figure A.2: Spatial residual diagnostic

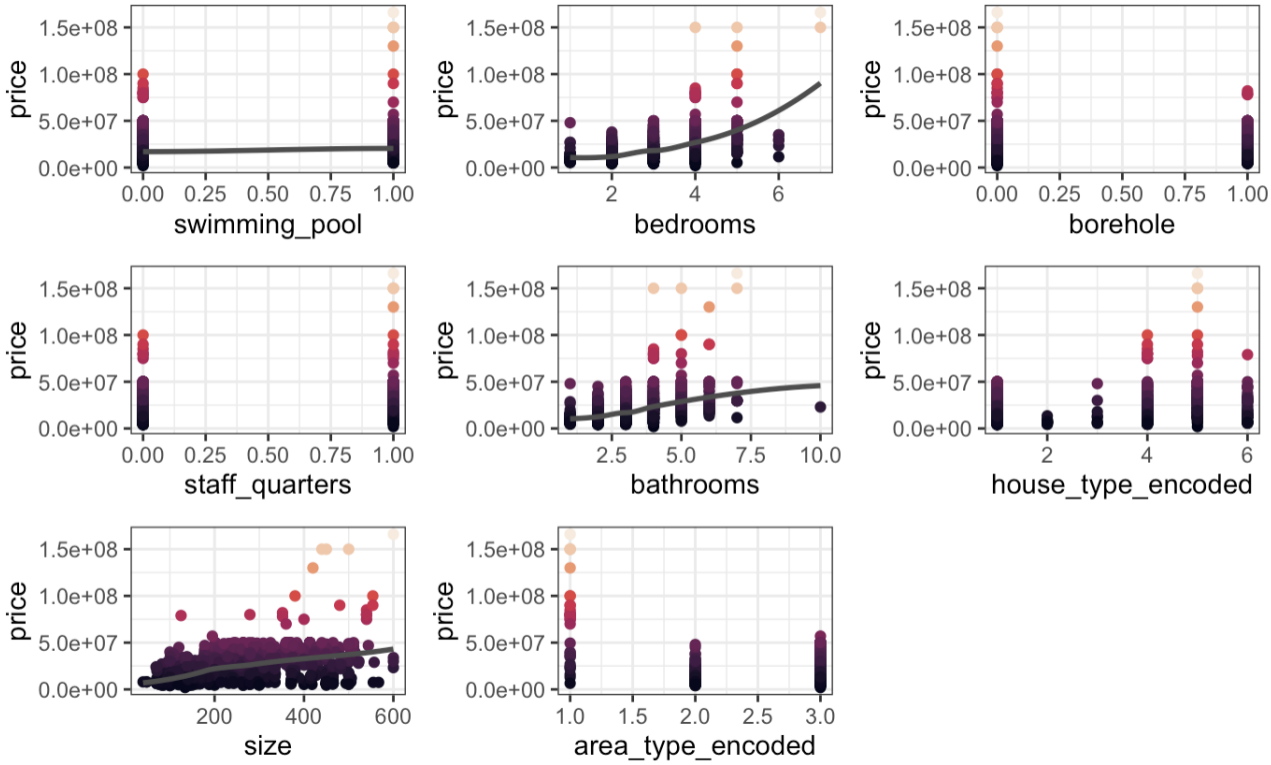


Figure A.3: Training dataset plot

VT OMNES
VNVM SINT

Appendix B

Letter of ethical review

Strathmore UNIVERSITY

12th February 2024

Ms Osoro Faith,
faith.mirriam@strathmore.edu

Dear Ms Osoro,

RE: Prediction of Housing Prices in Nairobi Metropolitan Area using Machine Learning

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC1999/24.** The approval period is from **12th February 2024 to 11th February 2025.**
This approval is subject to compliance with the following requirements:

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

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

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

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

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

Appendix C

Code

You can access the complete code for this project on GitHub using the following link:

GitHub Repository: House Price Prediction

Appendix D

Plagiarism report

Turnitin Originality Report

Processed on: 23-Feb-2025 9:01 PM EAT
ID: 2596099074
Word Count: 14599
Submitted: 1

FAITH OSORO DISSERTATION.pdf By Faith Osoro

Similarity Index	Similarity by Source	
17%	Internet Sources:	10%
	Publications:	12%
	Student Papers:	14%

6% match (student papers from 16-Oct-2023)
Class: DSA 8403 Dissertation Seminars 2 (Moodle PP)
Assignment: DS 2 CAT 1 Submission (Moodle PP)
Paper ID: 2197548103

1% match (student papers from 09-May-2023)
Submitted to Strathmore University on 2023-05-09

< 1% match (student papers from 13-May-2023)
Submitted to Strathmore University on 2023-05-13

< 1% match (student papers from 10-Apr-2023)
Submitted to Strathmore University on 2023-04-10

< 1% match (student papers from 23-Oct-2022)
Submitted to Monash University on 2022-10-23

< 1% match ("Inventive Communication and Computational Technologies", Springer Science and Business Media LLC, 2023)
"Inventive Communication and Computational Technologies", Springer Science and Business Media LLC, 2023