LEVERAGING MACHINE LEARNING IN HOUSING PRICE PREDICTION IN NAIROBI COUNTY.

JENNIFER WAMBUI NDUATI
MPPM

ADMISSION NUMBER 135508

A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS OF PUBLIC POLICY MANAGEMENT OF STRATHMORE UNIVERSITY BUSINESS SCHOOL.

STRATHMORE BUSINESS SCHOOL
STRATHMORE UNIVERSITY
NAIROBI, KENYA

2023

DECLARATION

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

Name of Candidate: **Jennifer Wambui Nduati**

Approval

The dissertation of **Jennifer Wambui Nduati** was approved by the following:

Name of Supervisor: **Dr. John Olukuru**

School/Institute/Faculty: **Strathmore Business School**

Dr. Ceaser Mwangi
Executive Dean
Strathmore University Business School.

Dr. Bernard Shibwabo
Director, Office of Graduate Studies

DEDICATION

I dedicate this dissertation to my husband **Tony** and our two sons **Chris** and **Ian** for their unwavering support, prayers, love and encouragement during this journey.

ACKNOWLEDGEMENT

To my supervisor **Dr. Olukuru** who always inspired me to trudge on and always availed himself. Your guidance, support and encouragement saw me through this journey. To **Tim**, **Chakaya** and **Janice** for aiding with data collection and analysis

To my mums **Mary** and **Eunice** for their prayers and cheering me on. To **Sandra** and **Reuben** for their support and always having our little ones when I needed a break.

I thank the Almighty **God** for guidance, strength, protection and sustaining me.

ABSTRACT

Housing prices have in the recent years dominated social and economic discussions in both developed and developing countries such as Kenya. Literature shows that modelling housing pricing in Kenya is predominantly based on conventional statistical theory and methodologies that are increasingly becoming sub-optimal in the wake of big data and technological advancements. To fill this gap, this research leveraged Machine Learning (ML) in housing price modelling and prediction in Nairobi County. The project further sought to achieve four objectives, namely: 1) to determine the features that significantly influence housing prices in Nairobi County; 2) to compare different ML models and techniques used to predict housing prices in Nairobi County; 3) to predict housing prices in Nairobi County using trained ML models based on unseen data in production; and, 4) to make policy recommendations on determination of housing prices. Primary data was collected from homeowners and potential homeowners from Shauri Moyo and Kibra. Further, Key Informant Interviews (KII) were undertaken with respondents from the State Department for Housing. The survey achieved a response rate of 85.3% and established that participants strongly agreed that; presence of tarmacked road within 5 Kilometres (Km), electricity connection/generator, a hospital within 5Km, a school within 5 Km, internet connectivity, age of the house *et cetera* significantly determined housing prices. Secondary data on the other hand was retrieved from Property24.co.ke. Fourteen ML/ Deep Learning models were trained, optimized and tested based on the evaluation metrics; Root Mean Squared Error (RMSE) and R-Squared. Insights from the secondary data showed that; number of bedrooms, bathrooms, parking lots, location and type of the house accounted for at least 88% of variations in the predicted house sale price in half of the ML models. The best ML candidate was the Light-Gradient Boosted Machine (Light GBM) with a RMSE of 11.21635 and an R-Squared score of 88.65%. The least performing was the Elastic Net model with a RMSE of 15.405066 and an R-Squared score of 78.59%. Four models with the best predictive accuracy were used to predict housing prices in Nairobi County based on real world data points from 9 random locations, resulting to predictions that had minimal disparities. The study recommended to property developers and policy makers as stakeholders in the housing sector to: allocate resources towards consistent and efficient collection, storage and sharing of quality data on housing features that significantly influenced housing prices in Nairobi County. Additionally, the study advocated for data-oriented Government policies in the housing sector and the implementation of new-age technologies such as AI/ML for efficient modelling and prediction of housing prices.

Keywords: Machine learning; price prediction; supervised learning; housing features.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AEH	-	Adaptive Expectation Hypothesis
AHP	-	Affordable Housing Program
AI	-	Artificial Intelligence
ARIMA	-	Autoregressive Integrated Moving Average
ANN	-	Artificial Neural Network
CatBoost	-	Category Boosting Machine
DL	-	Deep Learning
DT	-	Decision Trees
E-Commerce	-	Electronic Commerce
EDA	-	Exploratory Data Analysis
ENet	-	Elastic Net Regression
Ext-DT	-	Extra Decision Trees
IQR	-	The Inter-Quartile Range
FCDO	-	Foreign Commonwealth and Development Office
FGDs	-	Focused Group Discussions
GBM	-	Gradient Boosted Machine
GDP	-	Gross Domestic Product
GOK	-	Government of Kenya
Kes	-	Kenya Shillings
KII	-	Key Informant Interviews
KM	-	Kilometres
KNBS	-	Kenya National Bureau of Statistics
K-NN	-	K – Nearest Neighbors
LASSO	-	Least Absolute Shrinkage and Selection Operator
LR	-	Linear Regression
MAE	-	Mean Absolute Error
MAPE	-	Mean Absolute Percentage Error
ML	-	Machine Learning
MLP	-	Multi-Layer Perceptron
MSE	-	Mean Square Error
NACOSTI	-	National Commission for Science, Technology and Innovation
NHC	-	National Housing Corporation

NLP	**-**	Natural Language Processing
OECD	**-**	Organisation for Economic Cooperation Development
PE	**-**	Prediction Error
Q1	**-**	Quarter 1
Q4	**-**	Quarter 4
ReLU	-	Rectified Linear Unit
RF	**-**	Random Forest
RFR	**-**	Random Forest Regression
RMSE	**-**	Root Mean Square Error
SDGs	**-**	Sustainable Development Goals
State Department	**-**	State Department for Housing and Urban Development
SVM	**-**	Support Vector Machine
SVR	**-**	Support Vector Regression
UCI	**-**	University of California Irvine
US	**-**	United States
VAR	**-**	Vector Autoregressive Model
VECM	**-**	Vector Error Correlation Model
XG Boost	**-**	Extreme Gradient Boosting

OPERATIONAL DEFINITION OF TERMS

Big Data: This means a combination of structured, semi-structured, and unstructured data that is extremely large, fast or complex such that it is either very difficult or impossible to store or process it using classical data management tools (Botello, 2021).

Features: In the context of Machine Learning, features are explanatory variables that through data predict a regressor or a classifier (Brown, 2019).

Fundamental Variables: These refer to factors that drive house prices by causing either a shift in demand or supply of houses (Bundhun, 2020).

HPI: This is an abbreviation for Housing Price Index, which refers to a measure of movement of house prices, specifically measuring average price changes in the housing sector (Cabrera, Wang, & Yang, 2011).

House Prices: These refer to asset prices of residential houses and the land associated with (Oust, 2018).

Housing Bubble: This refers to a situation where increase in house prices is not justified by macroeconomic fundamentals or variables and other underlying factors (Oust, 2018).

Production: The designed system in which a trained Machine Learning model consumes real world data for prediction purposes (Mayr, 2020).

Unseen Data: This is real world data that is not used in training a ML model (Mayr, 2020)

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Housing is defined by the World Health Organization as one of the basic human needs (Habitat for Humanity, 2021). Gaining access to housing is crucial for reduction of poverty, advancement of equal opportunities, and bolstering sustainable growth. This is why housing was prioritised as goal Eleven (11) of the Sustainable Development Goals (SDGs) (Habitat for Humanity, 2021). It was subsequently mainstreamed and prioritised in the 'Kenya Kwanza Plan: The Bottom-Up Economic Transformation Agenda 2022-2027' as the Affordable Housing Program (AHP) in the current Kenyan government administration. The AHP is aimed at providing 250,000 affordable homes to be implemented annually from 2022 to 2027 (UDA-Kenya Kwanza , 2022).

Housing should be *physically adequate* and *fit for human habitation* in order for it to be considered as decent. Hence, a house cannot be considered habitable if it is overcrowded and unhealthy (Habitat for Humanity, 2021). As a result, the pricing of a house should consider features such as age, size, number of rooms and location. Additional attributes of a suitable house include availability of and access to ancillary utilities such as water, electricity and ample space (Habitat for Humanity, 2021). Other important factors are proximity to important institutions and facilities including schools and hospitals as they contribute to making a house healthy and fit for human living.

Literature shows that the world economy has suffered at least ten recessions since the Second World War period of 1939-1945 (Plakandaras, Gogas, Gupta & Papadimitriou, 2015) . Of the ten economic recessions, eight were predicted by the housing market, underscoring the importance of housing in stabilising the world economy (Plakandaras, Gogas, Gupta & Papadimitriou, 2015). Relatedly, the global financial crisis in 2008 generated a housing catastrophe in Europe that continues to date (Habitat for Humanity, 2021). These developments show that proper housing policies and practices are closely intertwined with the economic stability of a country or region.

Real estate, is one of the fastest-developing and significant sectors of the economy, thereby requiring a rationalized evaluation and forecasting of housing prices using mathematical modelling and other appropriate scientific tools (Plakandaras, Gogas, Gupta & Papadimitriou, 2015). Indeed, some scholars have sought to study the multivariate dynamics that shape the broader housing ecosystem in various parts of the world. One of these scholars, who has focused on the intersection between housing and market dynamics, attributes various factors to the risen research on undercurrents of housing prices (Otrok, 2005). A critical reason for scholarly interest in understanding price dynamics of houses is because in the current neoliberal society, houses are highly valued assets that are either owned or desired by most households (Plakandaras, Gogas, Gupta & Papadimitriou, 2015). Indeed, costs involved in housing account for a significant share of the total portfolio of financial intermediaries (Chen, Chen, & Huang, 2013). In Kenya and many developing economies with a fledgling middle and working classes - expenditure related to housing, including rent, construction and maintenance, gobbles up a huge chunk of the total expenditure of households (Chen, Chen, & Huang, 2013).

Nonetheless, the spectre of housing taking up the bulk of household incomes is not an aberration unique to Africa. The same state obtains in other parts of the world, as evidenced by statistics attributed to the Organization for Economic Cooperation and Development (OECD, 2021) which show that house-related expenses form the highest household outlays in the European Union and OECD countries. Accordingly, these housing expenditure in Europe and OECD countries averaged 22% of household costs in 2019 (OECD, 2021).

Understanding dynamics that underpin the housing value chain is also critical because generally, housing contributes significant portions of the Gross Domestic Product (GDP) of all countries. Without doubt, therefore, housing is a particularly vital asset class for most people, yet the information we have on prices is inadequate compared to the information we have on foreign exchange rates, share prices and bond prices (OECD, 2021). The gap arises partly because the research on determining housing prices in Kenya using statistical/numerical (traditional) models dominates existing literature.

Despite these traditional tools offering useful information, they are limited in some ways. Among these limitations are those that arise from technological advancements, which generate new challenges. For example, emerging problems include inability to process and analyse vast amounts of data, processing highly unstructured data, as well as high computational costs. Yet,

(Taffese, 2006) argues that efficient prediction of real estate value is critical to key stakeholders namely: prospective investors, homeowners, developers, appraisers, tax assessors, policy makers, mortgage lenders and insurers.

Furthermore, there has been growing demand for more powerful technologies such as Artificial Intelligence (AI) tools – that incorporate Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) – raising the need for different sectors of the economy to adopt actionable insights from data (Taffese, 2006). Worth noting is that AI encapsulates the science of getting computers and machines to impersonates the problem-solving and decision-making capabilities of the human mind (IBM, 2021). ML on its part is one of the trailblazing technologies in AI whose performance improves with more exposure to data, and can be used to process and analyse huge and complex data (Ngiam & Khor, 2019). Lately, the influence of ML has yielded tangible results that have disrupted major industries, including real estate, healthcare, retail, manufacturing, E-commerce, and logistics (Ngiam & Khor, 2019).

Some scholars have shown how developed countries are successfully applying ML models to determine housing prices. For instance, (Ong, 2021) explored five ML algorithms in predicting housing prices in Singapore and established that those additional variables increased model accuracy. This has aided buyers to establish whether they are getting a fair value while purchasing houses. In developing countries such as Kenya, the extent to which ML models have been used in determining housing prices remain uncertain, mainly because little current research exists that can illuminate the gap.

Regrettably, the housing market in Kenya has attracted relatively little empirical research (Moko & Olima, 2014). Further, there is evidence of inadequate empirical literature on the features which determine housing prices in Kenya. Instead, more focus has been placed on house characteristics that determine cost of house construction (Moko & Olima, 2014). Consequently, it is important to conduct more empirical research on aspects of housing pricing in order to determine how current best practice ML models elsewhere can be replicated in Kenya for modelling and ultimately prediction of housing prices based on unseen real data points once the models are deployed in production. This move would boost our understanding of how market and other dynamics that are data-based shape housing pricing.

1.2 Statement of the Problem

This study was aimed at adopting ML in modelling and prediction of housing prices in Nairobi County. The study further highlighted the challenges that are currently in place and inhibit the determination of housing prices in the said County. Additionally, it identified housing features that determine pricing, focusing on different variables in order to guide policy decisions aimed at enhancing affordability of houses.

Existing literature and our preliminary investigations showed that there was a great problem in assessing housing affordability in Kenya through price modelling this mostly relies on classical statistical modelling techniques (Makena, 2011); (Kihumba, 2017); (Kosgei, 2018); (Ungayi, 2019). However, such techniques have become increasingly sub-optimal and undependable especially in light of technological advancements. These technological developments include data storage capacity, increased computing power, and availability of vast amounts of complex data rich in insights. Technological advances have in turn created the need for more powerful and effective tools such as ML for modelling and forecasting housing prices. Resultantly, key stakeholders in the housing sector such as policy makers, home owners, investors, regulators, investors, lenders and investors are affected due to the aforementioned inadequate mechanisms.

Furthermore, available secondary data on properties in Nairobi contained in the various property websites barely listed any properties located in low-income locations in Nairobi County.

In an aim to bridge the gap of available data on low-income location, this study collected data from two low-income areas, namely, Kibra and Shauri Moyo. This study also transcended the gap of use of classical statistical methods by adopting ML in modelling and prediction of housing prices. In so doing, this study examined different variables to enhance price prediction in order to guide policy decisions on affordability. This researcher envisaged that the study would fill the gaps in efficient and robust modelling and prediction of housing prices in ways that align with technological advancement and Big Data. Additionally, the study will contribute to policy advancement and implementation of the housing agenda through enabling stakeholders namely: home owners; policy makers; regulators; lenders and investors. The researcher chose to use Nairobi County as the case study because of the county's highly cosmopolitan nature and representation of a wide range of socio-economic classes.

1.3 General Objective

To examine the efficacy of utilizing ML tools to model and predict housing prices in Nairobi County to aid stakeholders namely: home owners; policy makers; regulators; lenders and investors in making data-driven decisions.

1.4 Specific Objectives

i) To determine the features that significantly influence housing prices in Nairobi County.
ii) To compare different ML models and techniques used to predict housing prices in Nairobi County.
iii) To predict housing prices in Nairobi County using trained ML models based on unseen data in production.
iv) To make policy recommendations on determination of housing prices in Kenya.

1.5 Research Questions

i) What are the features that significantly determine housing prices in Nairobi County?
ii) Which ML models are the best candidate(s) in modelling and predicting-housing prices in Nairobi County?
iii) How reliable are ML models in predicting housing prices in Nairobi County, based on unseen data in production?
iv) What policy recommendations can be put in place to determine housing prices?

1.6 Scope of the Study

The study was restricted to, firstly, determining features that significantly affected housing prices in Nairobi County. Secondly, the study focused on modelling and predicting housing prices of residential houses in Nairobi County, Kenya. The study was based on primary data collected from Kibra and Shauri Moyo, both being low-income regions in Nairobi County, as well data collected from the directors at the State Department for Housing and Urban Development (State Department). This was further supplemented by insights underpinning secondary data collected from Property24.co.ke.

The study also focused on Supervised Learning to develop ML architectures for modelling and predicting housing pricing in Nairobi County. This was aimed at ascertaining the reliability and scalability of the tested ML models in predicting housing pricing using unseen data. The researcher anticipated that actionable insights from ML driven analytics would consequently influence impactful data-driven decisions for property investors and policy makers in Nairobi County specifically and Kenya generally.

1.7 Significance of the Study

This study was important as it would enable all key stakeholders in the housing sector namely: real estate investors; appraisers; tax assessors; mortgage lenders; insurers and policy makers to make evidence-based decisions on housing prices determination through data-intensive house price modelling and forecasting leveraging Machine Learning technology. The study would do so through yielding scalable solutions that inform policies that are data-driven. Lastly, the study would also contribute towards achieving the housing agenda by the GOK alongside SDG 11.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presented a theoretical review, empirical review, the research gaps and the conceptual framework of the study, all in line with the study objectives. The theoretical review focused on theories related to features influencing housing prices, adoption of ML in modelling, and prediction of housing prices. The empirical review, on its part, analysed features determining housing prices and adoption of ML models thereto. The section on research gap highlighted the informational lacunae in the empirical literature with regard to housing price prediction in Nairobi County, while the conceptual framework presented hypotheses that reflect the relations between various features and the target variables.

2.2 Theoretical Framework

A number of theories were used by the researcher to anchor the study and to determine the features influencing housing prices through adoption of ML modelling. These theories were the Adaptive Expectation Hypothesis (AEH), Housing Needs Theory, and Housing Deficit Theory.

2.2.1 The Adaptive Expectation Hypothesis Theory

The Adaptive Expectation Hypothesis Theory is traceable to the mid-20th Century, credited to Cagan (Cagan, 1956) . The hypothesis asserts that the determination of future house prices is dependent on house price trends and past information. In this regard, (Campbell & Shiller, 1988) refer to AEH as extrapolating behaviour, whereas Dipasquale and Wheaton (1996) refer to the same as backward-looking expectation. On their part, Kaiser and Cressie (1997) aver that positive expectations of growth in an economy boost the housing cycle, thereby creating optimism among investors when the housing market prices rise.

According to Riddel (1999), decisions based on demand are made by individuals on the basis housing prices in the past instead of applying key market fundamentals. This has in turn brought about by an over reliance on information contained in previously traded houses, while estimating the current prices in the market Hwang and Quigely (2010). Intriguingly, David and Zhu (2004) posit that as house markets experience a drop in price, home buyers remain

optimistic and are willing to pay higher prices, exacerbating a further rise in house prices. This remains evident in Kenya.

Malpezzi & Wachter (2005) posited that the rise in house prices, even with new stock of houses being put up, has been associated to speculative behaviours. These scholars in turn established that the housing market in the context of AEH does not follow a "random walk" due to the reliance on previous trends where the growth rate path expectations for house prices is associated with former trends in the movement of house price.

This theory provided a much-needed explanation regarding the less predictable increase in housing prices in Nairobi County despite the market changes experienced in the housing market. This researcher complemented the Adaptive Expectation Hypothesis Theory with the Housing Needs Theory, and Housing Deficit Theory as discussed below.

2.2.2 Housing Needs Theory

Similar to the Adaptive Expectation Hypothesis Theory, the Housing Needs Theory also emerged in the Mid-20th Century (Rossi, 1955). This theory was pioneered by Rossi in 1955 when he introduced the notion of housing needs to hypothesize residential satisfaction or lack thereof. In so doing, Rossi averred that changing aspirations and housing needs such as bigger spaces for growing families as households evolve through different life cycle stages lead to households falling out of conformity with their housing and neighbourhood situations (Rossi, 1955).

The gap between their present and desired or aspirational housing needs creates discontentment with their current places of residence, thus resulting in households responding to such dissatisfaction through migration to an environment within its housing needs (Rossi, 1955). Accordingly, households tend to feel dissatisfied if their housing and neighbourhood do not meet their current needs and aspirations. This theory is relevant in this study as it speaks to how housing features such as size or proximity to a school determine house pricing as ascertained in the findings in this study and would therefore be key in predicting housing pricing. However, it only alludes to the relationship between these features and the supply/demand dimension of housing. To address this disconnect, the researcher enlisted the

Housing Deficit Theory, also as a way of addressing the corpus of variables that underpin the housing ecosystem.

2.2.3 Housing Deficit Theory

In 1978, Morris and Winter coined the notion of 'housing deficit' to theorize residential fulfilment or lack thereof as posited by Mohit and Al-Khanbashi (2014). In this model, Morris and Winter argued that individuals assess their housing conditions in accordance with defined cultural and personal norms. Interestingly, both sets of norms shaped different households' own standards for housing (Mohit & Al-Khanbashi, 2014). In Kenya, this theory is applicable bearing in mind, for example, that families with sons tend to have additional units outside the main house for when the sons become of age.

This means that any inconsistency between the actual housing situation and the cultural and or familial housing norms results in a housing deficit, which gives rise to residential dissatisfaction (Mohit & Al-Khanbashi, 2014). Households or families with a housing deficit who are hence dissatisfied are more likely than not to consider some form of housing modifications. These households may attempt to make adjustments to reduce their dissatisfaction by revising their needs and aspirations by improving their housing conditions through remodelling or relocating (Mohit & Al-Khanbashi, 2014). Therefore, the relevance of this theory in this study was that it spoke to societal and familial norms which explain why housing features such as the size of the house determine house pricing and should always be effectively taken into account in determining housing pricing in Nairobi County.

2.3 Theoretical Review on Machine Learning Models

Housing price modelling and prediction using key ML Regression algorithms were evaluated. The choice of these models was aimed at leveraging the different strengths and proficiencies of the models in deriving meaningful insights from the analysis of dependencies between the target variable and features.

2.3.1 Support Vector Regression

Support Vector Regression (SVR) is a supervised learning algorithm based on the Support Vector Machine (SVM) classification algorithm which is credited to (Boser, Guyon, & Vapnik, 1992). Accordingly, SVM generates a hyperplane that separates data points into different classes while SVR produces a hyperplane accommodating the maximum data points. Support

Vector Machines continue to gain popularity for their distinguished features and different approaches to classification and regression tasks for linear and non-linear data. They are also popular for their better empirical prediction efficiency coupled with avoidance of over-fitting (Panigrahi & Mantri, 2016).

2.3.2 Random Forest Regression

Random Forest Regression (RFR) is an ensemble learning algorithm that uses supervised learning approach. The Random Forest (RF) algorithm on which RFR is based, is employed for both classification and regression tasks. Random Forests were invented by Breiman (2001) and are an ensemble of decision trees in a way that each single tree is a function of a random vector independently and identically sampled for all trees within the forest.

During model training phase, RF combines a number of decision trees and averages their output to give a weighted prediction of all decision trees (Cutler, Cutler, & Stevens, 2012). Figure 2.1 below demonstrates a decision tree performing a classification task.

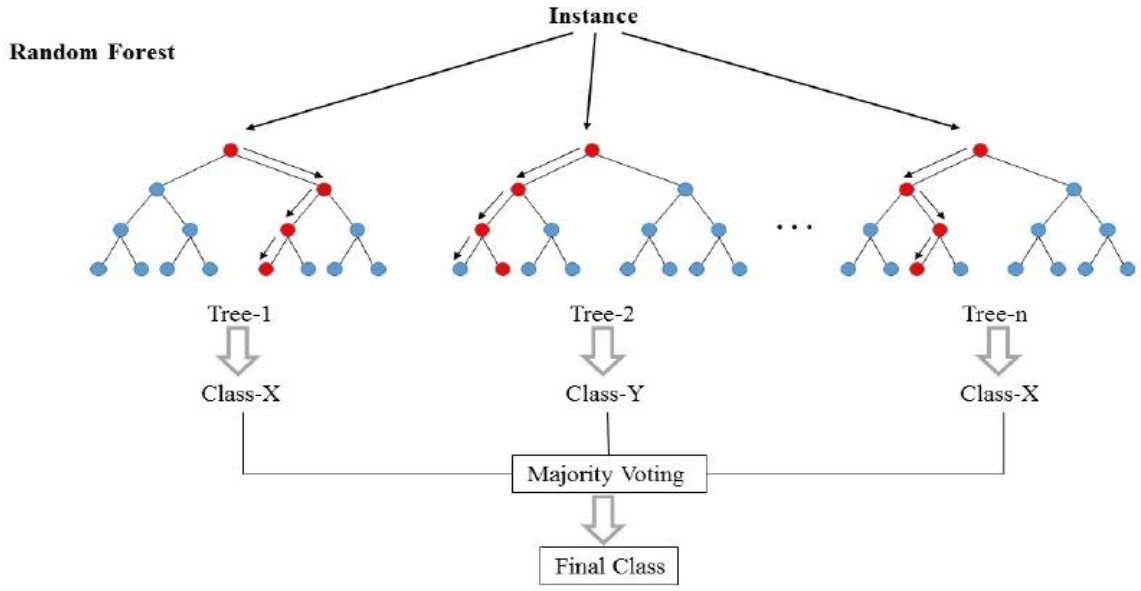


Figure 2.1 Overview of a Tree-Based Random Forest

Source: (Biau & Fr, 2012)

Statistically, RFs are appealing on a number of fronts namely: missing values imputation, measurement of feature importance, differential class weighting, and visualization of results (Ravikumar, 2018).

2.3.3 Multi-Layer Perceptron Regression

The idea of Neural Networks was introduced by McCulloch and Pitts (1943), and it began as a model of how neurons in the human brain function, termed 'connectionism' and used connected circuits to simulate intelligent behaviour. The Multi-Layer Perception Model (MLP) is a Deep Learning architecture based on feed-forward Artificial Neural Networks (ANN) that mimic the functioning of the human brain (Lim, Wang, Wang & Chang, 2016). The MLP comprises of three types of layers: input, hidden, and output. The input layer holds the features and receives the input signal for processing. The actual computation and processing of the input data takes place within the hidden layers whose number is arbitrarily determined. Finally, the predicted output is performed by the output layer. Data flows in a forward direction from the input layer, through the hidden layers, to the output layer as illustrates in Figure 2.2 below.

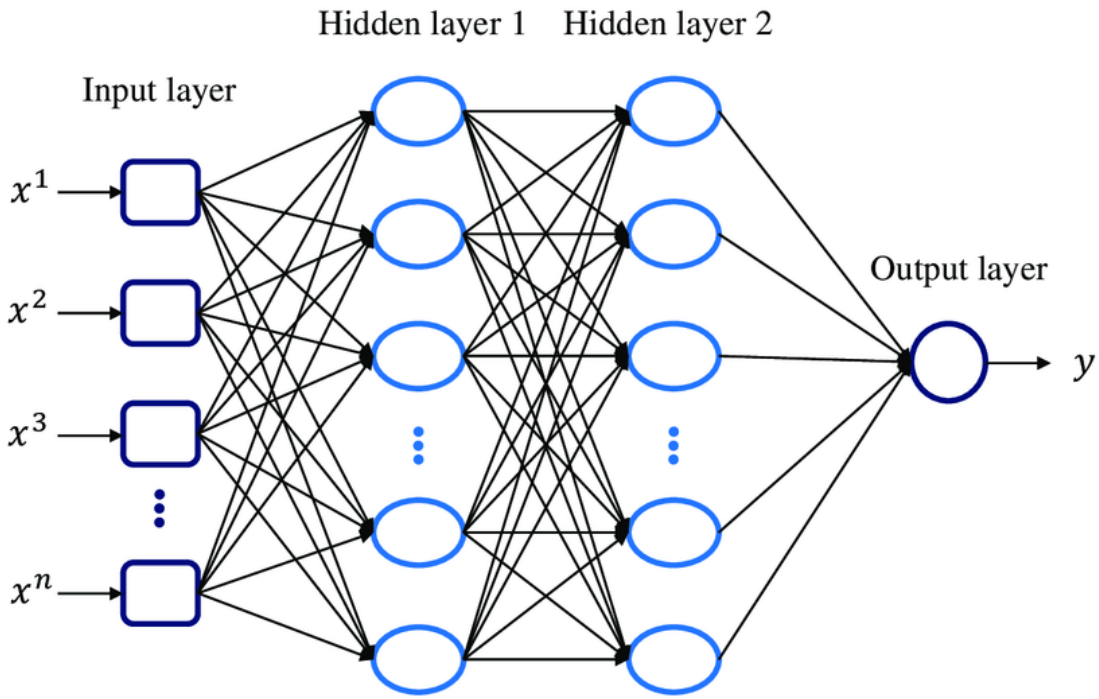


Figure 2.2 Illustration of the Multi-Layer Perceptron

Source: (Lim et al., 2016)

The neurons in the MLP are activated through activation functions and are trained with the back-propagation learning algorithm (Abirami & Chitra, 2020). Generally, MLPs have the capacity to capture deep patterns and dependencies in data as they are designed for the approximation of any continuous function and problems which are not linearly separable as is the case with most real-world data.

2.3.4 Linear Regression

Linear Regression (LR) is an approach used to model the linear relationship between a response variable and one or a set of predictors. The goal of LR is to find the best linear relationship between the dependent variable and the independent variables (Montgomery, Peck, & Vining, 2021).

Linear Regression assumes that both the dependent and independent variables have a linear relationship. This implies that the dependent variable's change is proportional to the independent variables' change. When making predictions about the dependent variable, the method estimates the independent variable's coefficients using the least squares approach. Simple Linear Regression and Multiple Linear Regression are two examples of the many variants of linear regression. The most common application of linear regression occurs in economics, where it is used to examine data on inflation, economic growth, and other economic indicators. The simplicity of linear regression is one of its main advantages. The results are uncomplicated to interpret, and the methodology is simple to comprehend and apply. The technique is both well-known and widely applied, and there is a wealth of literature and resources that may be used to support its application (Hastie, Tinshirani, & Friedman, 2009) .

However, there are several limitations on LR. The approach is susceptible to outliers and the assumption of linearity may not always hold true in practice. Furthermore, the residuals are assumed to be independent and normally distributed with a constant variance, which may not necessarily be the case in data from the actual world. It is important to note that LR is a basic technique in statistics and machine learning, and it has received much study and application across many disciplines.

2.3.5 The XG Boost Regression

The XG Boost is a gradient boosting algorithm commonly used as a regressor or a classifier depending on the ML problem. It is an implementation of Tianqi Chen's Gradient Boosting Framework and works best when used with huge datasets that have a lot of features (Chen & Guestrin, 2016). The approach has been integrated into several well-known machine learning libraries, including scikit-learn and R's caret package.

One of XG Boost's major benefits is its capacity to deal with categorical features and missing values directly, without the requirement for pre-processing. A variety of regularization parameters are also included, including the L1 and L2 regularization terms, which can aid in preventing overfitting. Additionally, XG Boost is renowned for its quick computation and top-notch performance, especially when used on large datasets with a lot of features. The algorithm employs a method known as "stochastic gradient boosting," which is intended to increase the model's speed and accuracy.

On a variety of datasets, including ones with a lot of features and a non-linear connection between the features and the target variable, XG Boost has been demonstrated to perform well in terms of performance. It has been discovered to be more accurate than other well-known algorithms like Random Forest and Support Vector Machine in several instances. The effectiveness of XG Boost and its different parameters have been the subject of numerous studies. For example, in a study by (TongHe, 2016), the authors found that XG Boost performed well on a large dataset of 1.46 million samples and fifty-four (54) features. The study found that XG Boost was able to achieve an accuracy of 99.4% on the test set, which was significantly better than other machine learning algorithms such as Random Forest, SVM, and Logistic Regression.

Another study by (Fang Chen, 2016), compared the performance of XG Boost to several other machine learning algorithms on a dataset of 569 samples and 30 features. The authors found that XG Boost performed better than the other algorithms in terms of both accuracy and computational efficiency. Overall, XG Boost is a powerful and efficient algorithm that is well suited for large datasets with a high number of features. It has been shown to be effective in both classification and regression tasks and has been used in a wide range of real-world applications (Fang Chen, 2016).

2.4 Review of Empirical Literature

This section reviewed and critiqued previous authors and researchers' findings on the application of ML modelling in determining housing price prediction. The sections are presented in the order of: features that determine housing pricing; and the algorithm statistical learning approaches to Machine Learning.

2.4.1 Features Determining Housing Prices

Broadly, housing features refer to the amenities that are associated with particular houses and housing plans. Such features include the presence or lack of wheelchair access, letter boxes for apartment blocks, security and parking slots, as well as corridors and other common areas (Chan & Xiong, 2007). In recent developments, features may also refer to common attributes associated with particular set-ups or socio-cultural and architectural designs that can be found in particular regions. An example of this meaning can be seen in the courtyard housings of Beijing (Chan & Xiong, 2007).

Understanding the concept of housing features, therefore, calls for region-specific tools of interpretation. As an example, traditional data analytics tools, time series and regression modelling techniques namely: Vector Autoregressive Model (VAR), Vector Error Correlation Model (VECM), Multiple Regression, and Hedonic Regression dominate Kenyan literature (Makena, 2011); (Kihumba, 2017); (Waari & Kosgei, 2018); (Ungayi, 2019). These are highly unstructured, manual, unstable and unreliable techniques while dealing with highly unstructured datasets.

Elsewhere, (Park & Bae, 2015) made a valuable contribution to the Pakistan housing market by collecting data and developing the first scientific housing dataset which is now widely used. In so doing, they explored eleven ML algorithms and used them to predict housing prices for the real estate market in Pakistan utilizing Islamabad's city housing data. They considered various features, among them living area size, number of bedrooms, number of bathrooms, presence of a dining room, presence of a lounge, presence of a swimming pool, location of the house, presence of a nearby hospitals, presence of nearby schools, presence of a nearby shopping malls, availability of parking space, and availability of security staff.

The study used three model performance evaluation metrics namely; Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) to compare the models based on their predictive performance on the test set that was set aside prior to training the ML algorithms. Their results were as follows; Linear Regression model had a MAPE of 5627.9369, MAE of 10928.2603, and RMSE of 16658.4158, Bayesian Ridge Regression had a MAPE of 7383.9969, MAE of 10930.6388, and RMSE of 16661.3350, Support Vector Regression had a MAPE of 1918.4957, MAE of 8595.6057 and RMSE of 18209.5558, Stochastic Gradient Descent Regression had a MAPE of 10698.1442, MAE of

13139.1928, and RMSE of17345.1444, Elastic-Net Regression had a MAPE of 7388.1547, MAE of10927.7181, and RMSE of 16658.2267, Gradient Boosting Regression had a MAPE of 5267.4830, MAE of 9563.4324, and RMSE of 16772.3870, Lasso-Lars Regression has a MAPE of 7382.6600 , MAE of 10938.5807, and RMSE of 16670.3489, Random Forest Regression had a MAPE of 7371.0746, MAE of 10902.9762, and RMSE of 17105.2596, Passive Aggressive Regression had a MAPE of 2133.8370, MAE of 8621.9391, and RMSE of 18069.2298 while Theil-Sen Regression had a MAPE of 6031.6336, MAE of 10151.4884 and RMSE of 16754.2930.

Other researchers, (Belke & Keil, 2017), appraised the key determinants of real estate prices of one hundred (100) German cities. They found that the supply-side factors of construction activity and housing stock, as well as the demand-side factors of apartment rents, market size, age structure, local infrastructure, and rental prices, are the important determinants of real estate prices.

On their part, Ersoz et al. (2018) studied the Turkish real estate using data mining and explored a number of features among them; real estate type (land, residential, shop), building age, distance from city center, building protection system, popularity demand, parking and garden, number of floors, building heat protection, presence of a lift, area of location, landscape etc. They considered the Chi-square Automatic Interaction Detector (CHAID) and Classification & Regression Trees (C&RT) algorithms for data mining analysis. The two algorithms encompassed the following prediction models; Neural Net, Decision Tree, Linear & Logistic Regression and Bayes Classifiers. According to C&RT algorithm's output, the distance to the city centre of 0.7 Kms was the most valuable feature in determining the real estate unit value. On the other hand, according to the CHAID algorithm's output, whether the house's area of location is popular was the most important feature.

Other important features included; size of the building, its protection system, and availability of park and garden. Generally, the study established that the C&RT algorithm had the highest prediction success with a Mean Absolute Error (MAE) of 72.22 compared to that of the CHAID algorithm which was determined to be 97.62. They further argued that the use of advanced technologies in ML, statistical modelling and automatic value estimation are important in determining the real value of the real estate.

In Canada, (Manasa, Gupta, & Narahari, 2020) undertook a real estate prediction-based study in Montreal to forecast house prices based on a number of variables namely: number of rooms, police station proximity, fire station proximity, location among others. Ultimately, (Manasa, Gupta, & Narahari, 2020) study leveraged various algorithms, among them Linear Regression, Support Vector Regression, K-Nearest Neighbour Regression, Regression Decision Trees, Extreme Gradient Boosting (XG Boost) Regression Model, and Random Forest Regression.

To assess and compare the predictive performance of the models based on the test set, three evaluation metrics were used, namely; Coefficient of determination R-Squared, Root Means Squared Error (RMSE), and Root Mean Squared Logarithmic Error (RMLSE). The study reported that; Linear Regression had an R-Squared of -2.12, RMSE of 0.2077, and RMLSE of 0.03493, Ridge Regression Model had an R-Squared of 0.4358, RMSE of 0.5224, and RMLSE of 0.040701, Lasso Regression had an R-Squared of 0.4430, RMSE of 0.5224, and RMLSE of 0.04069, Support Vector Regression had an R-Squared of 0.6630 and RMLSE of 0.03170 while XGBoost Regression had an R-Squared of .7584, RMSE of 0.3462, and RMLSE of 0.03170. The study concluded that the tested features and algorithms used concurrently yield better results in predicting house pricing.

2.4.2 Machine Learning

Machine Learning (ML) is a special discipline of AI that automates analytical models resulting in algorithms and systems that learn from data, improve with experience, and perform specific tasks without being explicitly programmed (Dey, 2016). Broadly, ML algorithms enable analytical models and designed systems to make decisions autonomously independent of external programming support.

There are three primary building blocks of ML. These are: Supervised Learning; Unsupervised Learning and Reinforcement Learning. This categorization is based on the type of input and output data, the learning technique and the problem being solved.

2.4.2.1 Supervised Learning

Supervised Learning encompasses classification and regression algorithms, which ingest data from features and output categorical or continuous values of the target variable. This is referred to as the *training set* from which the model adopts the mapping from the input space to the output space. This learning is referred to as the *training phase* from which the algorithm trains

on how to classify or regress new unseen data (Sah, 2020). Figure 2.3 below shows classification of data points into different classes.

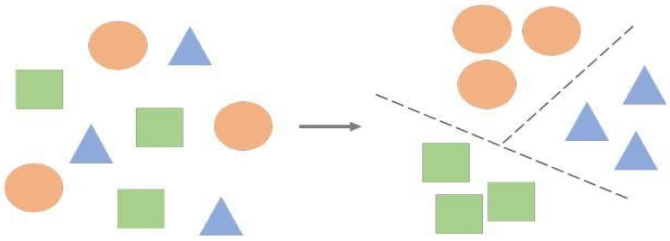


Figure 2.3 Overview of Supervised Learning in a Classification Task

Source: (Sah, 2020)

2.4.2.2 Unsupervised Learning

Unsupervised Learning encompasses clustering algorithms that consider data from different dimensions and group it into segments based on similarities across different aspects. Such models capture the underlying trends and patterns in the data so as to infer more on the data properties. Thus, this kind of Machine Learning does not utilize labelled data unlike the supervised learning techniques (Sah, 2020). Figure 2.4 below illustrates clustering of data points into different homogeneous segments.

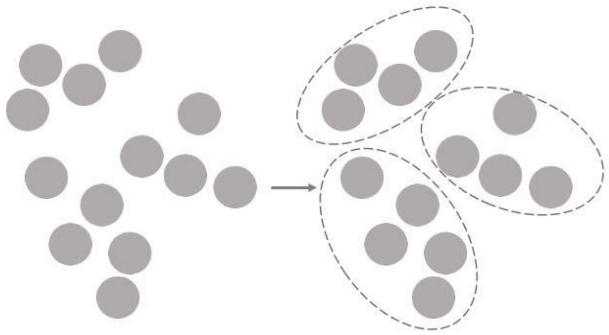


Figure 2.4 Overview of Unsupervised Learning

Source: (Sah, 2020)

2.4.2.3 Reinforcement Learning

In Reinforcement Learning, a model trains on data without the correct output key to sequentially make data-based decisions towards an ultimate reward. Depending on the actions performed, an agent gets either a penalty or a reward with the ultimate goal being to minimize the penalties and maximize the total reward (Francois-lavet, Islam, Bellamare, & Pineau, 2018). Figure 2.5 illustrates an artificial agent operating in an environment of rewards/penalties based on the output of an action. In a bid to maximize the overall reward, an agent in Reinforcement Learning studies the environment's state and performs actions to optimize the reward process.

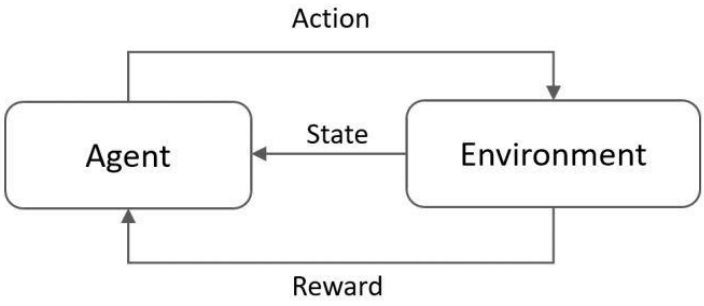


Figure 2.5 Overview of Reinforcement Learning

Source: (Sah, 2020)

2.5 Research Gaps

Review of current literature on determinants of housing prices and the prediction established that several studies employed AI and particularly ML tools for data processing and analysis leading to more stable and optimal prediction models (Khalafallah, 2008); (Park & Bae, 2015); (Lim, Wang, Wang, & Chang, 2016); (Ravikumar, 2018) (Manasa, Gupta, & Narahari, 2020); (Abirami & Chitra, NA). Nevertheless, to the best of our knowledge, insufficient research had been directed towards leveraging modern technologies such as ML and DL in modelling and predicting housing prices in Kenya. These modelling architectures have the capacity to capture deep dependencies and patterns in huge and complex data sets resulting in intelligent systems.

Further, this literature review has established that most studies have paid more attention to macro-economic factors vis-à-vis micro-economic factors in modelling and prediction of housing prices in Kenya. The review also shows that earlier studies tended to focus on

traditional data analytics tools and time series and regression modelling techniques, among them Vector Autoregressive Model (VAR), Vector Error Correlation Model (VECM), Multiple Regression, and Hedonic Regression dominate literature in the Kenyan case (Makena, 2011); (Kihumba, 2017); (Kosgei, 2018); (Ungayi, 2019).

These approaches are, however, manual in operation, unstable and unreliable, especially in the wake of huge, complex, and highly unstructured datasets. The proposed study, therefore, seeks to fill these gaps by leveraging ML and DL architectures such as the SVR, which is effective in addressing the problem of data over-fitting and empirically has a high prediction efficiency; Random Forest Regression which combines the predictive efficiency of different Decision Tree leaners through Ensemble Learning resulting in a superior and more efficient learner; and the MLP Regression which is based on Non-linear Neural Networks that can find intelligent shortcuts to achieve computationally expensive solutions rich in actionable data insights. These ML and DL architectures will automate the housing prices modelling and prediction procedures in Nairobi County by using as much data (structured, semi-structure, and unstructured) as is available. These details are summarised in Table 2.1 below.

Table 2.1 Summary of Empirical Literature Gaps

Author	Title	Findings	Research Gap
(Chan & Xiong, 2007)	Offered a conceptual definition of housing features in the Beijing Courtyard Housing scheme.	House structures are made more or less attractive and costly depending on availability of features such as availability of wheelchair access, security, parking spaces, size of corridors, play areas for children, recreation facilities, etc.	Focused on housing elements that could be used to determine how attractive and pricey courtyard houses that are shaped by cultural conventions should be. These houses are increasingly challenged by; Cultural changes, family planning and social changes, New technology and changes in building materials, Economic changes, Population growth, housing planning strategies etc.
(Makena, 2011); (Waari & Kosgei, 2018); (Kihumba,	Embraced traditional data analysis tools including Vector Autoregressive Model	The methods for housing price prediction in these studies though	There was need for novel AI & ML approaches, such as Chen & Guestrin's

2017); (Ungayi, 2019)	(VAM), Vector Error Correlation Model (VECM), etc. in housing price modelling and prediction	popular, they are unstable, unreliable sub-optimal in the wake of technological advancements and availability of highly unstructured data on housing features.	(2016), XGBoost Regressor Algorithm, Brieman's (2021) Random Forest Regressor, etc. that have the technological capacity to process and analyse complex datasets.
(Park & Bae, 2015)	Produced first scientific housing data set in Pakistan using ML algorithms. Consequently, explored eleven ML algorithms and used them to predict housing prices for the real estate market in Pakistan utilizing Islamabad's city housing data.	Elements of housing determine attractiveness and pricing of specific houses. Such include; swimming pool, location of house, size of living area, number of bathrooms, bedrooms, dining area, presence of a lounge etc.	Cited as their own limitation, they needed to complement the textual dataset with house's visual data to build more robust novel ML and perhaps Deep Learning models for house price prediction purposes.
(Belke & Keil, 2017)	Appraised key determinants of real estate prices in 100 German cities.	Established the correlation between supply and demand in housing.	There was need to focus more on micro economic elements that affect demand as opposed to just macro elements that affect supply.
(Ersoz, Ersoz, & Soydoan, 2018)	Focused on Turkey's housing terrain, using data mining.	Established that; city centre proximity, house size and its age are the most important elements of housing in deciding demand and setting their prices.	There was need to pay particular attention to these elements in designing housing estates, and determining their prices.
(Manasa, Gupta, & Narahari, 2020)	Attempted real estate prediction-based study in Montreal to predict house pricing, asking and setting prices based on key elements using various algorithms.	Tested features and algorithms used concurrently yielded better results in predicting house pricing.	There was need to leverage more advanced ML models such as those under Bagging Ensemble Learning e.g. Random Forest Regression and Deep Learning models to improve accuracy of predictions.

2.6 Conceptual Framework

Figure 2.6 below presents the study's conceptual framework. It gives a visual representation of the relationship between features and the target variable in the study.

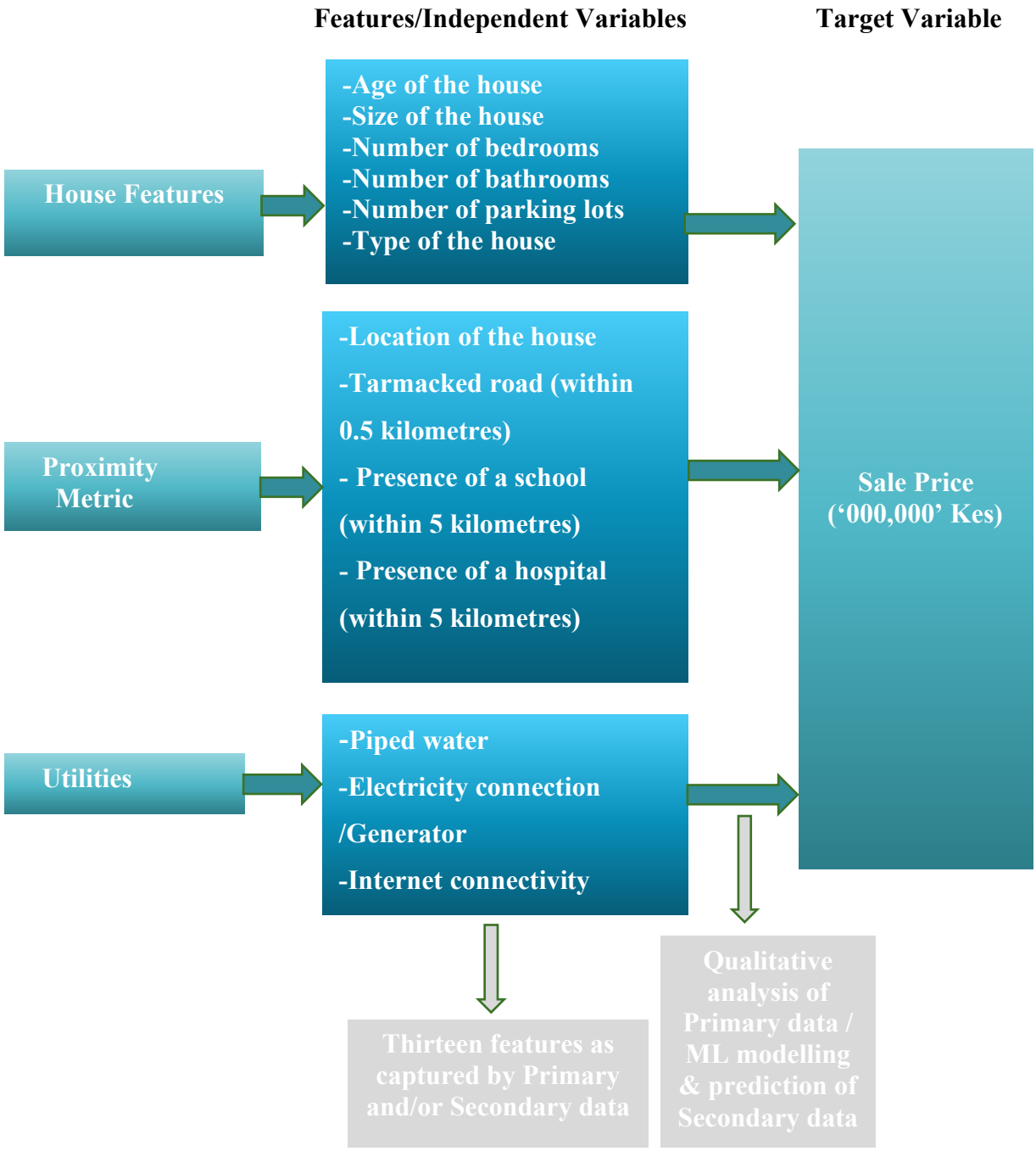


Figure 2.1 Conceptual Framework

Source: Researcher, (2023)

The framework shows the house features, proximity metric and utilities as the predictors and the house sale price as the target variable. It further shows that the presence of the afore-mentioned independent variables determine the price of the house. Thirdly, the Conceptual Framework further demonstrates that various factors influence the price per square meter. The operationalization of the variables is captured in Table 2.2 below.

Table 2.2 Operationalization of Study Variables

Variable	Variable Measures	Research Instrument	Data Analysis
House features	- Age of the house - Size of the house - Number of bedrooms - Number of bathrooms - Number of parking lots - Type of the house	- Questionnaires - Ordinal Scale (Likert Scale) - Key Informant Interviews - Web Scrapping	- Descriptive analysis - Correlation - Regression
Proximity Metric	- Location of the house - Tarmacked road (within 0.5 kilometers) - Presence of a school (within 5 kilometers) - Presence of a hospital (within 5 kilometers)	- Questionnaires - Ordinal Scale (Likert Scale) - Web Scrapping	- Descriptive analysis - Correlation - Regression
Utilities	- Piped water - Electricity connection/Generator - Internet connectivity	- Questionnaires - Ordinal Scale (Likert Scale)	- Descriptive analysis

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter presented the procedures and methodology of the study. The chapter aimed to ascertain how the use of ML could be used to determine housing pricing in Nairobi County. Specifically, the chapter described critical elements of an academic research, including research philosophy, research design, population and sampling, data collection instruments, and data collection methods. Other elements dealt with in the chapter were the research quality, data analysis tools, ethical considerations of the study, its limitations and assumptions.

3.2 Research Philosophy

This research employed a mixed method approach as the underpinning philosophy. Thus, the study relied on both qualitative and quantitative data in order to gain a more complete understanding of the issues and to get the perception of the participants. The integration of both qualitative and quantitative data drew on the strengths of both methods, using a convergent design whereby the qualitative and quantitative data was collected and analysed at similar times (Cresswell, 2014). This approach had a triangulation element in-built in both qualitative and quantitative methods of surmounting singular weaknesses of each and complementing each data set's strengths.

The mixed method approach thus ensured credibility of research findings and enhanced integration making the research problem clearer to the researcher in order to address the research questions with greater authority and in more detail (Hesse-Biber & Johnson, 2015). This philosophical foundation informed the study since the researcher collected secondary data independently and analysed with ML tools.

3.3 Research Design

According to (Saunders, Lewis, & Bristow, 2019) research design is a blueprint used to provide answers to research questions. It encompasses methods used in the gathering, analyzation and interpretation of data. This study employed an experimental research design in which Machine Learning modelling and prediction techniques were explored in establishing the underlying relationship between the target variable and the features. The choice of this research design

was informed by its renowned reputation as the "gold standard" by virtue of its being vigorous (Saunders, Lewis, & Bristow, 2019).

The researcher then applied a qualitative approach by processing and analysing the qualitative data from in-person interviews through filling questionnaires by existing home owners and potential home owners. The researcher also used Key Informant Interviews (KIIs) to collect information from the Principal Secretary and the Housing Secretary at the State Department of Housing and Urban Development (State Department) as well as the directors and assistant directors. This State Department is responsible for the implementation of policy and delivering houses to Kenyans on behalf of the Government of Kenya (GOK). The experiments were conducted based on the following steps: data scraping, data pre-processing and feature engineering, model training and optimization, model testing (in-sample), and model testing on real world data points in production.

3.4 Target Population and Sample Size

3.4.1 Target Population

Target population refers to the entire population of persons, events or objects showing common observable features (Mugenda & Mugenda, 2003). According to the State Department, there is a housing demand of 250,000 houses in Nairobi County (Ministry of Transport, Infrastructure, Housing and Urban Development, 2018). Therefore, for the primary data collection the target population was 250,000 homeowners and potential homeowners.

For the secondary data retrieval, the target population was the finitely many property listings that are published in a given property website at a particular point in time or within a period of time.

3.4.2 Sample Design and Sample Size

According to (Kothari, 2004), random sampling is the scientific practice of examining an entire population by investigating a subset of the population. Sampling is important due limited time and resources. (Mugenda & Mugenda, 2003) avert that for descriptive studies, 10%-30% of the population is an adequate sample size.

The researcher applied two forms of sampling in the collection of primary data. Firstly, simple random sampling was applied where members of the public in Kibra and Shauri Moyo were

selected to be part of the sample. Secondly, the purposive sampling technique was used to collect data from the State Department officers for expert opinions.

From the target population, $N = 250{,}000$ a sample size of $n = 600$ participants from Kibra and Shauri Moyo was arrived at using the sample size formula by (Yamane, 1967).

$$n = \frac{N}{(1 + Ne^2)} = 400{+}200$$

where $e = 5\%$ level of significance (for 95% confidence interval) is the Margin of Error. The Researcher arrived at a sample size of 400 participants but opted to interview 200 more participants.

Additionally, 8 participants were selected for KIIs from the State Department of Housing. Respondents in this sample were: the Principal Secretary and the Housing Secretary at the State Department as well as the directors and assistant directors. The respondents were chosen on the basis that they were best placed to give credible and accurate required information; since they were knowledgeable and are at the helm of the State Department. In this respect, (Campbell D. , 1955) posits that key informants must be conversant on the issues being studied.

For the secondary data, the researcher retrieved as much information on property listings of research interest as had been published on www.property24.co.ke as at January 31st, 2023. Thus, the sample size in this case was a function of time and could not be determined with certainty beforehand.

3.5 Data Collection Methods

The research relied on both qualitative and quantitative research data. Resultantly, primary was collected and secondary data scrapped from a website. The instruments used to gather data are crucial to the success of the research (Kerlinger & Lee, 2000). In this regard, the researcher took into account the topic, response rate, time and the targeted population.

Primary data was collected using closed ended and open-ended questionnaires. The researcher used a five-point Likert scale technique. The researcher approached the respondents for permission to conduct the research. Once permission was granted, the respondents were given a chance to go through the questions and date was collected on the spot. The questionnaires

were structured based on reviewed literature as well as the operationalization of the variables of the study.

Key Informant Interviews (KIIs) were also undertaken in the form of google form (virtually) questionnaires. The researcher approached the Principal Secretary, Housing Secretary, directors and assistant directors in the State Department and sought permission to conduct the research. Once permission was shared the researcher sent the questionnaire in the form of Google forms aimed at reinforcing electronic data collection process. This form of data collection allowed for confidentiality and reduced bias on the part of the researcher; since the questions are administered by a research assistant. The respondents were allowed seven days to complete and submit the form.

Secondary data was retrieved from the website www.property24.co.ke regarding housing and property listings in Nairobi County through web scrapping methods in Python.

3.5.1 Primary Data

Primary data was collected from home owners and potential home owners in Nairobi County, as well as from officers at the State Department. Data in Nairobi County was collected from Kibra and Shauri Moyo which comprise of areas containing social housing. Additionally, this was informed by the fact that data pertaining to housing in these locations is not readily available thereby creating a gap on housing information owned by majority of Nairobi County dwellers. Data collection exercise was carried out between January 9th, 2023 and February 13th, 2023 whereby 600 respondents from Shauri Moyo and Kibra regions in Nairobi and 8 officers from the State Department of Housing were interviewed utilizing Google forms as the data collection tools. Collection of this data was important as it allowed interaction with home owners and potential home owners through first-hand provision of original, specific and unique insights.

3.5.2. Secondary Data

Secondary data entailed online publication of property listings on various aspects of housing in Nairobi County. The researcher retrieved data on residential houses from www.property24.co.ke. The data contained property listings published on the website as at January 31st, 2023. The dataset consisted of 51,055 observations and 8 variables namely; Address, Region, Bedrooms, Bathrooms, Parking lots, Floor area, Type of the house and the

Sale price of residential houses. The residential houses included townhouses, villas and apartments. This played a key role in providing expert perspectives and insights from already existing data which was consequently used to train ML models for use in house price prediction goals

3.6 Validity Tests

Validity is crucial for determining whether the research truly measures that which it was intended to measure before data is seen. The researcher sought to ascertain the validity of the selected instruments using a sample of eleven questions in a pilot study in which the ambiguity relating to the questions was amended. The research supervisor was involved in coming up with the research questionnaire for the content validity index in order to determine the worthiness of the instrument. The supervisor assisted in the review and correction of errors in the instrument that may have impacted the content validity of the questionnaire.

3.7 Data Analysis and Presentation

Different approaches were utilized in analysing quantitative and qualitative data types. In both cases, Python Programming language was used for ML model development while Microsoft Power BI was leveraged for Exploratory Data Analysis (EDA).

3.7.1 Analysis of Qualitative Data

Questionnaires were used to collect qualitative data whereby Microsoft Power BI was used for Exploratory Data Analysis (EDA). Data were process and analysed for descriptive statistics as well as an interactive dashboard developed. Additionally, thematic analysis was done on the Key Informants data.

3.7.2 Analysis on Quantitative Data

For quantitative data, feature significance was explored through a correlation heatmap to compute the strength of the relationship between housing features and the target variable, house price. Subsequently, data pre-processing was carried out and ML models trained through supervised learning, optimized and used for prediction of housing prices in Nairobi County. The analyses outputs were then presented in the form of charts and tables.

Finally, the trained and optimized ML/DL models were evaluated on the test set based on the metrics; Root Mean Squared Error (RMSE) and the R-Squared score for predictive ability.

3.7.3 Machine Learning Design Workflow (Model)

The data processing and analytics was sequential and adopted an experimental design. The ML experimental design was as shown in Figure 3.1 below.

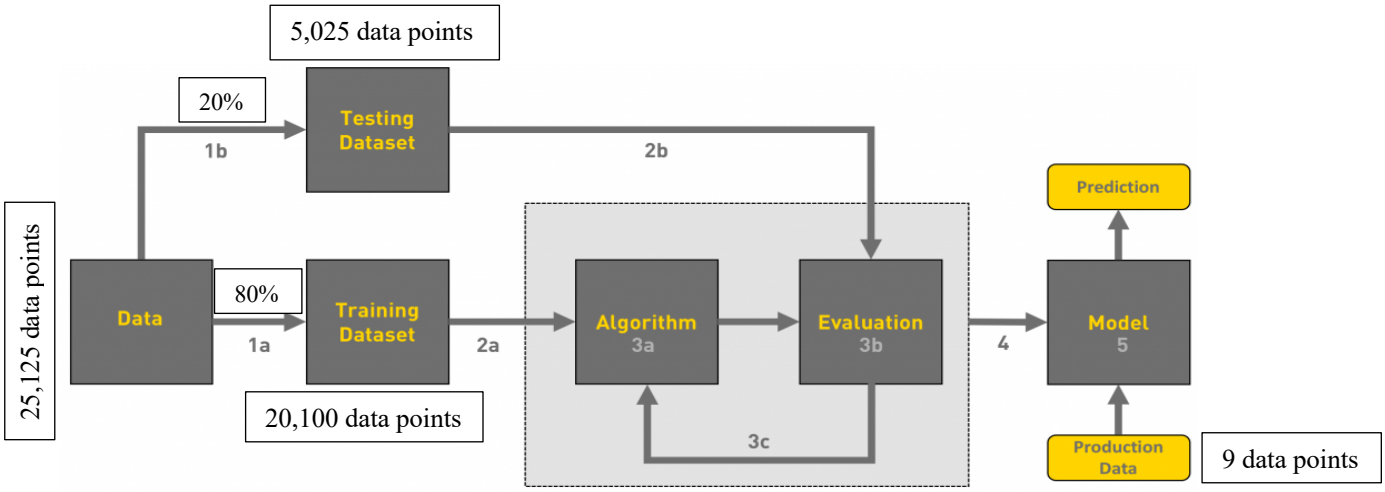


Figure 3.1 Overview of Machine Learning Workflow

Source: Towards Data Science (2019)

In **Stage 1**, raw data was gathered and EDA carried out to obtain a bird's view of the key characteristics and patterns of target variable and the features. Furthermore, the data was pre-processed to obtain clean data that is feasible for ML Models. In this stage, processes such as handling of missing values, encoding of variables, and feature scaling through normalization or standardization was carried out. Lastly, the dataset was split into a training set (1a) and a test set (1b) in the ratio 80% to 20% respectively. The training/testing ratio of 80/20 was the most ideal one for training and validating ML models (Nguyen, et al., 2021).

Stage 2 involved parsing the pre-processed data into well-researched ML models for housing price modelling and performance evaluation. In **Stage 3**, the ML algorithms training took place using 80% of the data while 20% of the data which formed the test set and was used to model performance evaluation. Moreover, in this stage model optimization took place through hyper-parameter tuning. In **Stage 4**, ML model deployment to production and continual testing took place. In **Stage 5**, the optimized model was used for prediction using new unseen data from the model deployment to production setting.

3.8 Ethical Considerations

Hoyle, Stephenson, Palmgreen, Lorch, and Donohew (2002) posited that ethical standards are honourable practices embraced in research undertakings. The researcher was guided by ethical standards to ensure that no rights of any respondents were violated.

Approval to proceed to the field for data collection was sought from the Strathmore Business School in the form provided as ***Appendix I***. Subsequently, a research permit was sought and obtained from National Commission for Science, Technology and Innovation (NACOSTI). Written permission to collect data was also sought from the State Department.

The researcher recruited and trained two research assistants who supported the researcher in administering the questionnaires and Key Informant Guide to identified respondents in Nairobi County and the State Department. The research assistants received a training package consisting of an introduction and understanding of the study's background and design, interviewing skills, ethical issues related to the study, consenting process and a thorough understanding of questionnaire items in the survey questionnaire. The researcher assured the respondents of their anonymity (where desired) and confidentiality. Participation during data collection was voluntary having obtained prior informed consent from the respondents who were guaranteed that all responses obtained were to be utilized for academic purposes only.

CHAPTER FOUR

DATA ANALYSIS, PRESENTATION AND DISCUSSION

4.1 Introduction

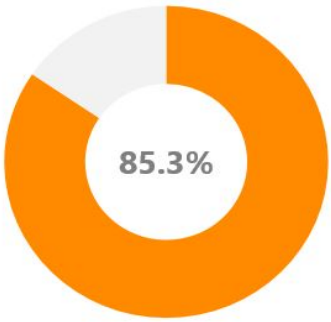
This chapter focuses on various findings obtained following the analytics of the study's datasets. The chapter organizes main findings in subsections presented under each study objective. The subsections comprise of demographic results, the summary of the descriptive results, ML modelling and prediction results as well as the inferential tests applied in determining the relationship of the study variables.

4.2 Primary Data Findings

Primary data was collected from home owners and potential homeowners in Kibra and Shauri Moyo in Nairobi County as well from officers at the State Department. The data obtained from the survey and the consequent analysis of findings are presented below.

4.2.1 Response Rate

The target sample for this study was 600 respondents comprising home owners and potential homeowners in Kibra and Shauri Moyo in Nairobi County, as well as the officers at the State Department. The study received 512 respondents consisting of home owners and potential home owners who filled in-person questionnaires and six (6) officers from the State Department who filled Google forms in the data collection process. Resultantly, the response rate in this study was 85.3% (n=512) with only 14.7% of the sample participants not responding to the research. Figure 4.1 below illustrated this.



Key: Non-Response Response

Figure 4.1 Response Rate

4.2.2 Cronbach's Alpha

This is a measure of scale reliability that seeks to measure the internal consistency of a questionnaire or survey. It ranges between 0 and 1 with higher values indicating more reliability and the threshold s 0.7. The formula is as follows[1]:

$$\alpha = \frac{\mathrm{N}\bar{c}}{\bar{v} + (\mathrm{N} - 1)\bar{c}}$$

Where: N is the number of items, $\bar{c}$ is the mean covariance between items and $\bar{v}$ is the mean item variance.

The Nairobi County results comprising of Kibra and Shauri Moyo survey had a Cronbach's alpha of 77.72% (0.7772) which was acceptable and the 95% confidence interval of [0.748, 0.804].

4.2.3 Profile of Respondents

This section provides details of respondents regarding their ability to buy homes. The study analysed the respondents in respect of their age, marital status, status of employment, number of children, income per month, level of education and home ownership status as is more particularly shown below.

4.2.3.1 Age of Respondents

Figure 4.2 below shows that a majority of the respondents were aged 31-40 representing 38.5% followed by persons aged between of 21-30 years who represented 32.36%. Persons aged between 41-50 years represented 20.93% while persons aged 51-60 represented 8.14%.

[1] https://stats.oarc.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/

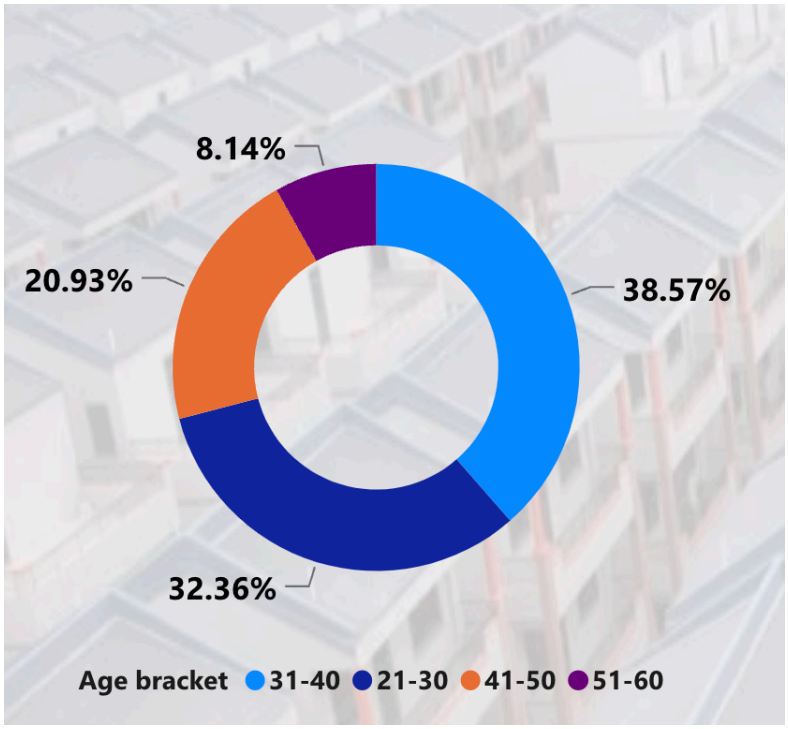


Figure 4.2 Age of the Respondents

4.2.3.2 Marital Status of Respondents

The research further shows that more than half, 68.34% of the respondents were married while 31.8% were not; as shown in Figure 4.3 below.

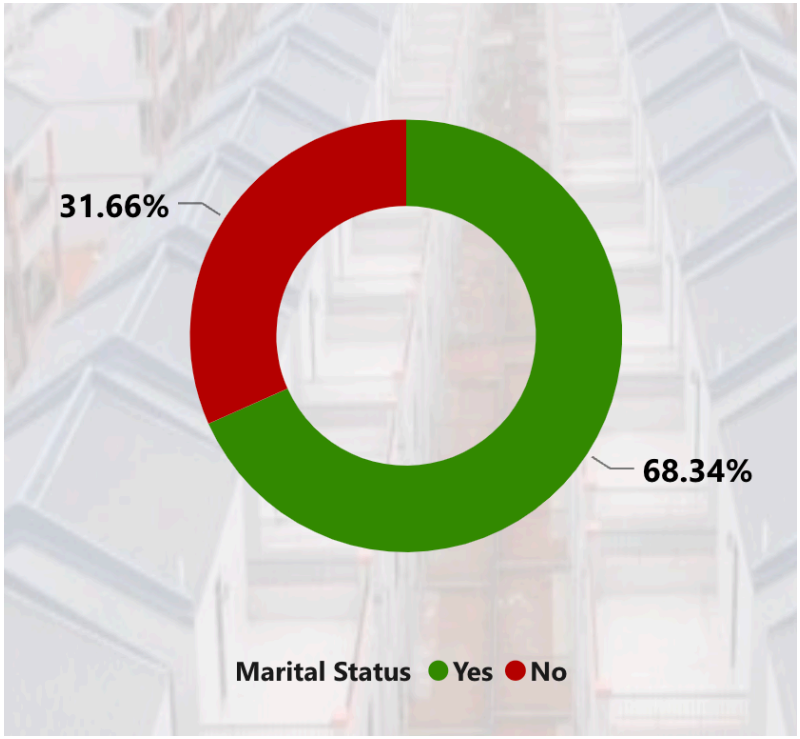


Figure 4.3 Marital Status of the Respondents

4.2.3.3 Status of Employment

Figure 4.4 below shows that close to half of the respondents, 46.56% were self-employed, 30.57% were employed at the time of the survey, while those not employed represented 30.57% of the total survey participants were not employed.

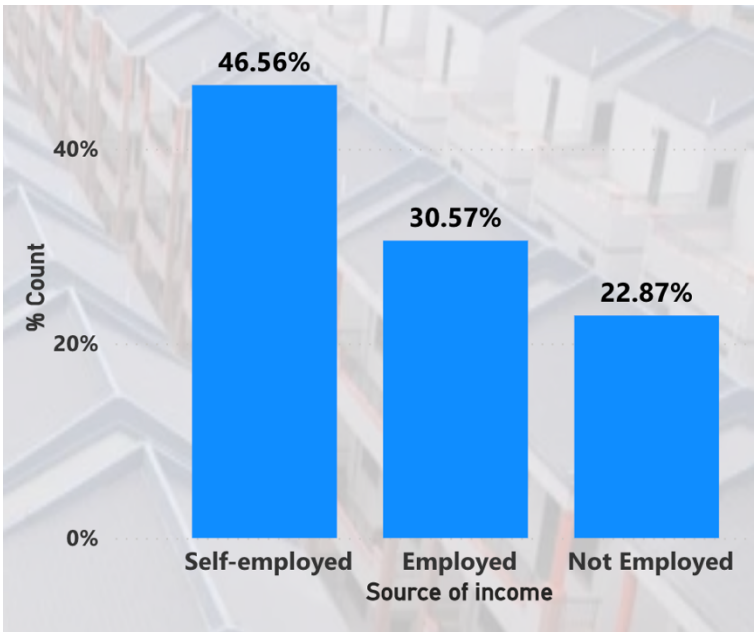


Figure 4.4 Employment Status

4.2.3.4 Number of Children

This study shows that a quarter specifically 25.30% of the respondents had no children. Further, 20.04% of the respondents had two (2) children, 18.62% had three (3) children, 11.34% had one (1) child, 10.32% had four (4) children, 8.50% had five (5) children, 3.64% had six (6) children, 1.62% had seven (7) children, 0.40% had eight (8) children while 0.20% had eleven (11) children at the time of the survey. Figure 4.5 below shows these findings.

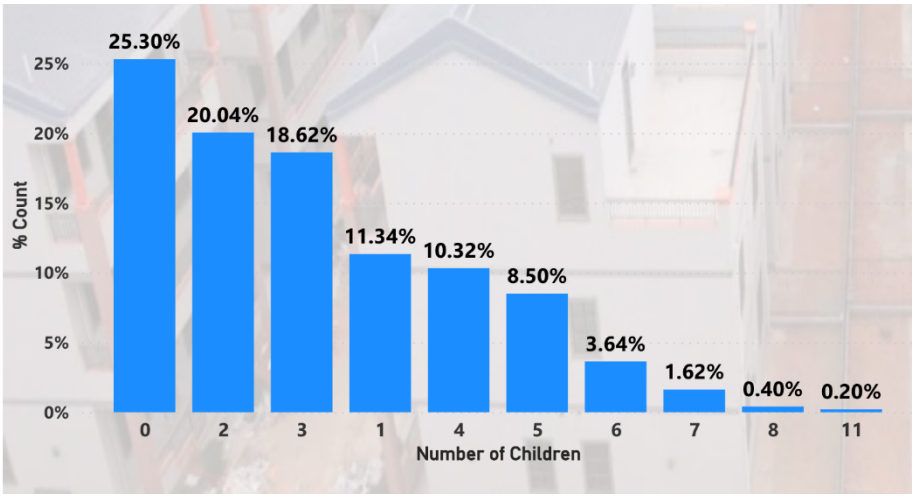


Figure 4.5 Number of Children

4.2.3.5 Monthly Income

Further, the study results as presented in Figure 4.6 below indicated that majority of the respondents, 39.83%, earned between Kes. 1,000-14,999 at the time of the survey, while slightly over a quarter (28.93%) earned Kes. 15,000-49,999. These were followed by 26.83% of the respondents who did not disclose their level of income. Finally, less than a tenth 4.40% earned Kes. 50,000-99,000.

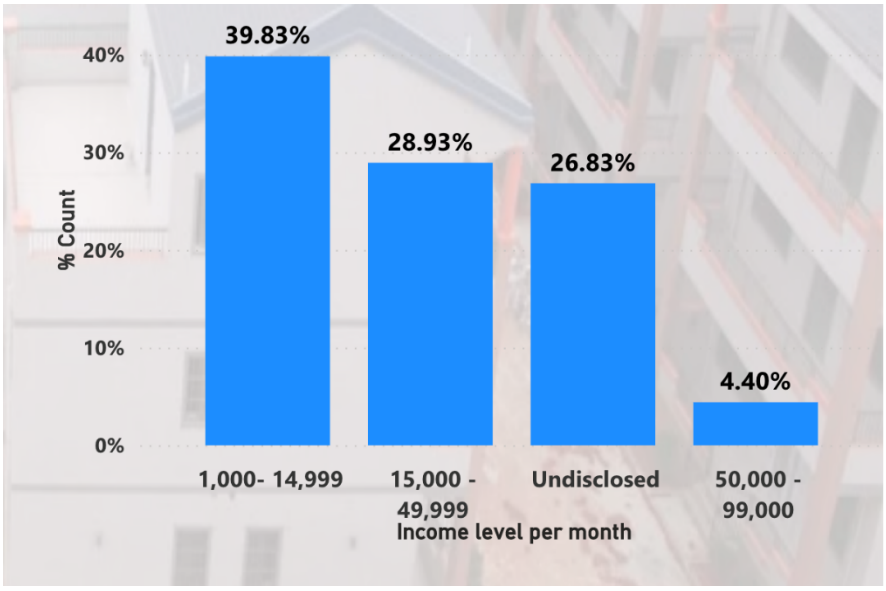


Figure 4.6 Monthly Income Levels

4.2.3.6 Level of Education

The study results as illustrated in Figure 4.7 below show that a majority of the survey participants, 43.44%, had attained secondary school education at the time of the study, while slightly more than a third of the participants had attained a University Bachelor's degree. Another 11.58% of the respondents had primary school education, while 7.72% of the survey respondents had attained other levels of education.

Out of the 7.72% of those who had attained other levels of education, more than three-quarters had attained tertiary college certificates; 5.13% had artisanship qualifications; 2.56% had a Master's degree, while 7.69% had no education at the time of this research study.

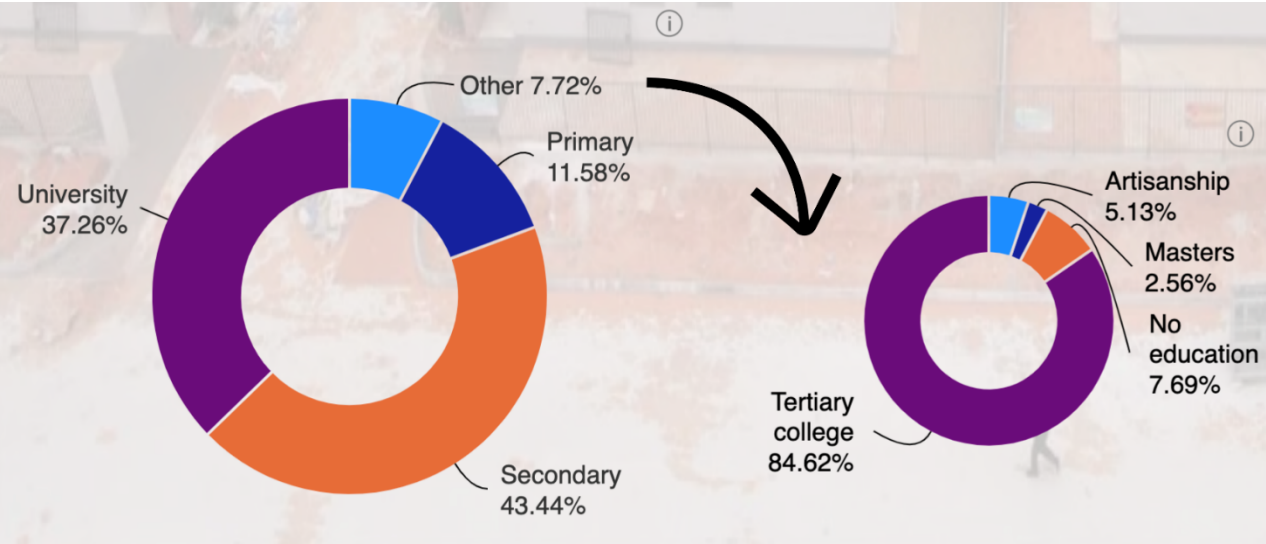


Figure 4.7 Level of Education

4.2.3.7 Home Ownership Status

The findings further reveal that half of the respondents comprising, specifically 50.62%, were potential home owners, while 42.14% were home owners. Nonetheless, 7.34% were neither potential home owners nor actual home owners at the time of the survey, as indicated in Figure 4.8 below.

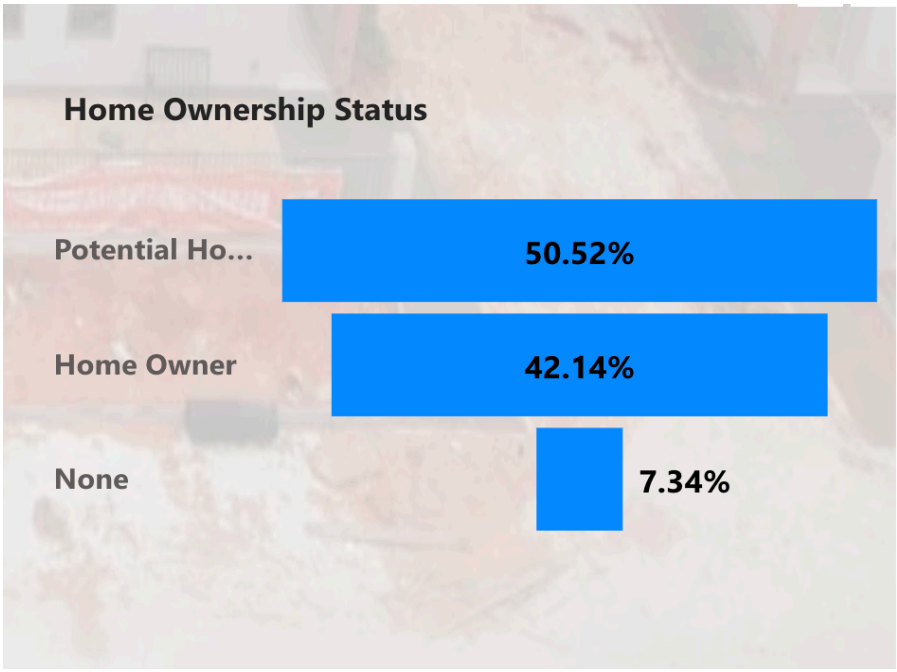


Figure 4.8 Home Ownership Status

4.3 Information on Housing Features

The first objective in this study sought to establish the features that significantly determine housing prices in Nairobi County. Therefore, this section strived to ascertain to what extent respondents agreed that housing features determined housing pricing in Nairobi County. This data was collected in the form of a Likert scale where the scores were represented as follows: 1 Strongly Disagree (SD), 2 Disagree (D), 3 Neutral, 4 Agree (A), and 5 Strongly Agree (SA). This was completed by the secondary data findings.

4.3.1 Size of the House, Income of the Owner and Number of Rooms

Following descriptive analysis, Figure 4.9 below showed that 34.11% of the respondents strongly agreed that the size of the house was an important factor in determining pricing, 17.89% disagreed with the statement, 17.05% agreed with the statement, 15.79% strongly disagreed with the statement while 15.16% were neutral on whether or not the size of the house was an important factor in determining pricing.

Additionally, 34.47% considered the income of the home owner as a significant consideration in house pricing, 23.19% agreed that the income of the home owner was an important factor in determining housing pricing, 16.38% disagreed with the statement, 14.89% were indifferent

while 11.06% strongly disagreed that the income of the home owner was an most important factor in determining housing pricing.

Further, 30.80% of the participants strongly agreed that the number of rooms was an important factor in determining house pricing in Nairobi County, 20.46% of the participants disagreed with this statement, 17.93% were neutral on whether or not the number of rooms was an important factor in determining house pricing, 15.82% of them strongly disagreed with the statement and 14.98% simply agreed that the number of rooms was a significant factor in determining house pricing.

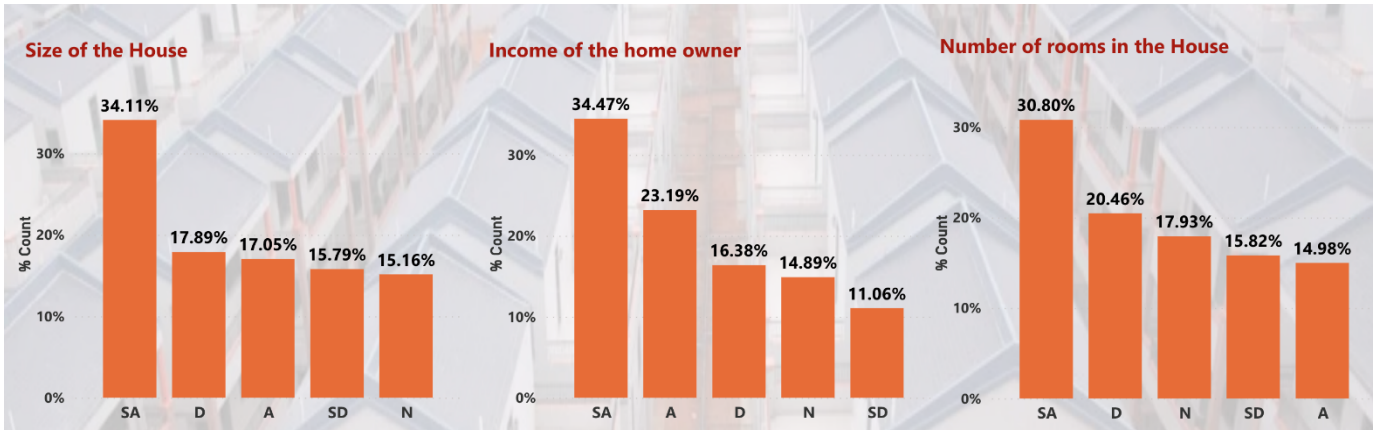


Figure 4.9 Size of the House, Income of the Owner and Number of Rooms

4.3.2 Location of the House, Age of the House and Presence of Tarmacked Road

On further analysis, Figure 4.10 below demonstrated that almost half, 49.47% of the respondents strongly agreed that location of the house was a key determinant of house pricing in Nairobi County, 34.11% simply agreed with the statement, 10.11% were indifferent on whether or not location of the house was a key determinant of house pricing in Nairobi County, 3.37% of them strongly disagreed that location of the house was a key determinant of house pricing in Nairobi County while 2.95% of the participants disagreed with the claim that location of the house was a key determinant of house pricing in Nairobi County.

Slightly more than half, 54.60% of the participants strongly agreed that the age of the house was a key feature that determines house pricing in Nairobi County, 23.98% of the participants agreed that the age of the house was a key feature that determines house pricing, 11.13% were neutral on whether or not the age of the house was a key feature that determines house pricing, 6.42% disagreed with the statement that the age of the house was a key feature that determines house pricing while 3.85% strong disagreed with this statement.

Additionally, slightly more than three-quarters, 75.16% of the survey participants strongly agreed that the presence of a tarmacked road within 0-5 kilometres was an important feature in determining housing pricing in Nairobi County, 16.42% of them simply agreed with the statement that the presence of a tarmacked road within 0-5 kilometres was an important feature in determining housing pricing, 5.47% of them were neutral on whether or not they thought that the presence of a tarmacked road within 0-5 kilometres was an important feature in determining housing pricing, 1.47% of the participants disagreed with the statement and a similar number strongly disagreed with the statement.

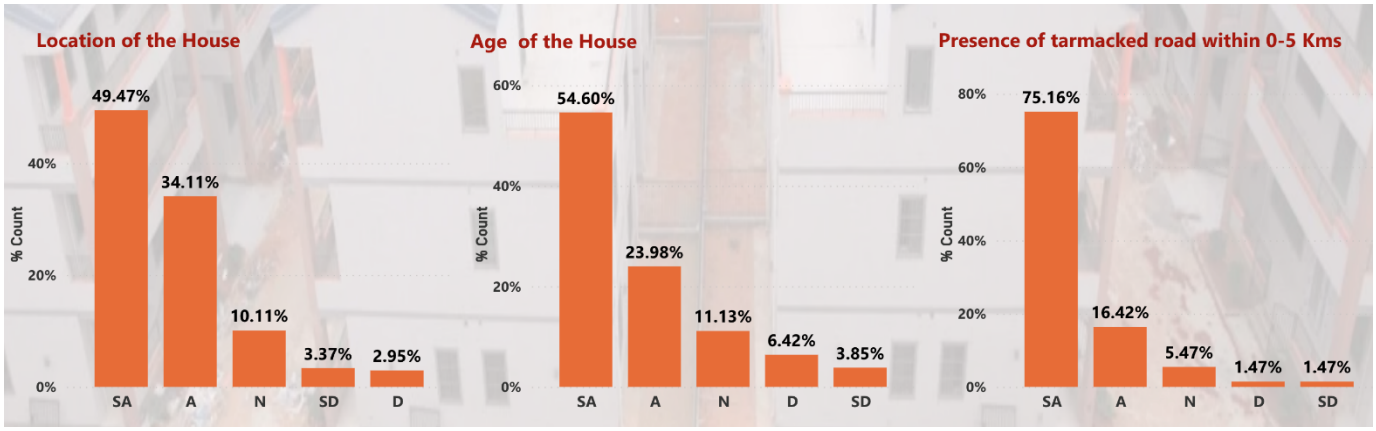


Figure 4.10 Location of the House, Age of the House and Proximity to a Tarmacked Road within 0.5 Kilometres

4.3.3 Presence of a School, Hospital and Piped water within 0-5 Kilometres

Further analysis in Figure 4.11 below showed that, more than half, 64.00% of the survey participants strongly agreed that the presence of a school within 0-5 Kilometres was an important determinant of house pricing in Nairobi County, 14.95% of them agreed with the statement, 10.11% of them were neutral about whether or not the presence of a school within 0-5 Kilometres was an important determinant of house pricing, 7.79% of the participants disagreed with the statement that the presence of a school within 0-5 Kilometres was an important determinant of house pricing in Nairobi County while 3.16% strongly disagreed with the statement.

More than half, 64.84% of the participants strongly agreed that the presence of a hospital within 0-5 Kilometres was an important feature that determined housing pricing in Nairobi County, 13.68% simply agreed with this statement, about a tenth, 10.32% were neutral on this opinion, 5.89% disagreed with the statement that the presence of a hospital within 0-5 Kilometres was an important feature that determined housing pricing in Nairobi County while almost a similar

number. 5.26% strongly disagreed with the statement that the presence of a hospital within 0-5 Kilometres was an important feature that determined housing pricing.

Less than half, 38.53% of the participants strongly agreed with the statement that the presence of piped water was an important feature that determined house pricing in Nairobi County, 18.95% of the participants remained neutral on whether or not the presence of piped water was an important feature that determined house pricing, 17.47% of them strongly disagreed with this statement, 15.16% disagreed with the statement while 9.89% of the participants agreed that indeed the presence of piped water was an important feature that determined house pricing in Nairobi County.

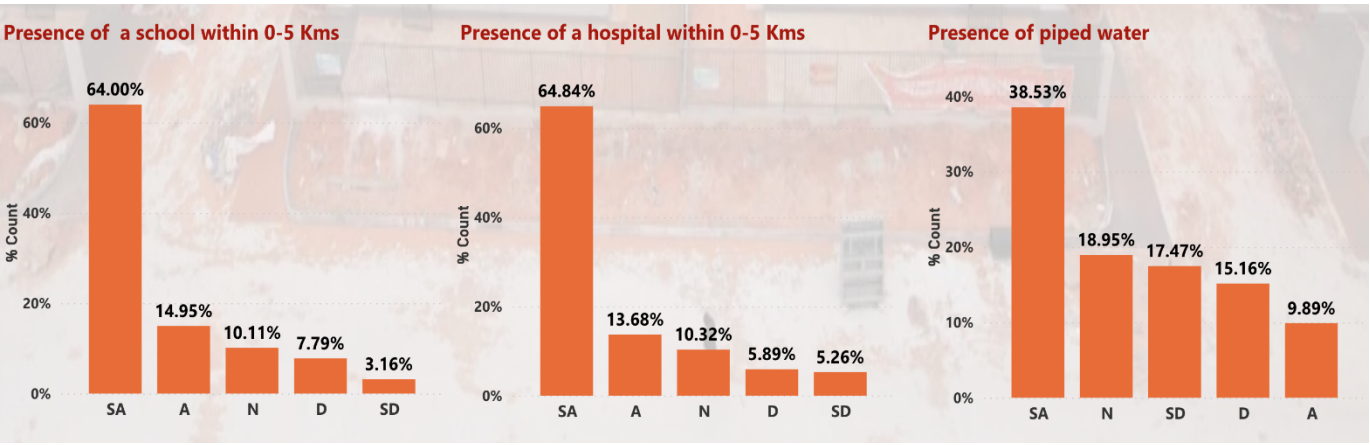


Figure 4.11 Presence of a School, Hospital or Piped Water within 0-5 Kilometres

4.4.4 Electricity Connection, Generator and Internet Connectivity

Figure 4.12 presents additional analysis on housing utilities. Three-quarters, 74.95% of the participants strongly agreed that electricity connection was an important determinant of housing pricing in Nairobi County, 18.74% of the participants agreed that electricity connection was an important determinant of housing pricing in Nairobi County, 3.37% were neutral on whether or not they thought electricity connection was an important determinant of housing pricing, 2.11% disagreed with this statement while a meagre 0.84% strongly disagreed that electricity connection was an important determinant of housing pricing in Nairobi County.

Slightly more than half, 55.49% of the participants indicated that they strongly agreed that the presence of internet connectivity was a key determinant of house pricing in Nairobi County, 12.66% disagreed with this statement, 11.60% agreed that the presence of internet connectivity

was a key determinant of house pricing in Nairobi County, 10.55% of the participants strongly agreed with the statement while 9.70% of the survey participants were indifferent.

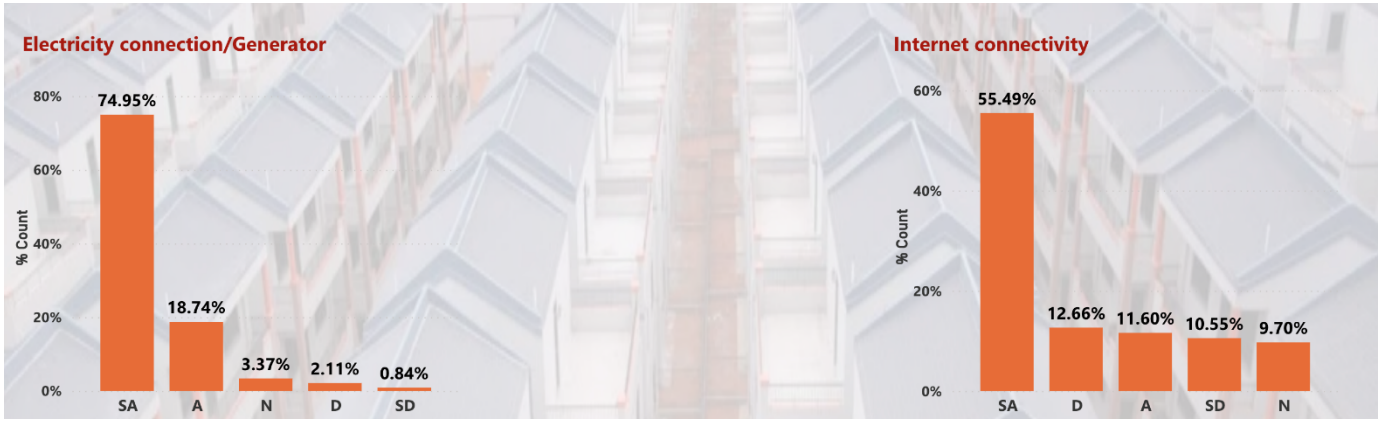


Figure 4.112 Electricity Connection/ Generator and Internet Connectivity

4.4 Extent to which housing features are key determinants of housing price

This study also sought to establish to what extent the respondents agreed that housing features mattered to them and that other stakeholders such as developers and policy makers are aware of this. Figures 4.13 and 4.14 below summarize the findings.

Figure 4.13 Features Considered in Determining House Pricing

As presented in Figure 4.13 above, less than half, 38.69% the respondents strongly agreed that features are taken into account while determining housing prices, 23.04% of the participants agreed that features are taken into account while determining housing prices, 21.14% of them remained neutral, 11.84% disagreed with the statement that features are taken into account while determining housing prices while 5.29% strongly disagreed with the statement.

Furthermore, close to half, 44.84% of the participants strongly agreed that features are key determinants of housing prices, about a quarter of them, 25.89% agreed that features are key determinants of housing prices, 17.68% were neutral as to whether or not features are key determinants of housing prices, 9.89% of them disagreed with the statement while 1.68% strongly disagreed that features are key determinants of housing prices.

Figure 4.14 Developers' Awareness of Importance of Housing Features to Respondents

As summarized in Figure 4.14, close to half, 45.24% of the respondents strongly agreed with the statement that developers are aware of these features, 20.93% of the participants remained neutral as to whether or not they thought that developers are aware of these features, 20.30% agreed with the statement that developers are aware of these features, 8.46% disagreed with the statement while 5.07% of the survey participants strongly disagreed that developers are aware of these features that determined house pricing in Nairobi County.

Additionally, more than half, 66.52% of the survey participants strongly agreed that housing features that determine housing prices matter to them individually, 19.04% of the participants agreed that housing features that determine housing prices matter to them individually, 6.65% of the participants remained indifferent on whether or not housing features that determine housing prices matter to them individually, 5.03% of the participants disagreed with this statement while 2.84% of the survey participants strongly disagreed with the claim that housing features that determine housing prices matter to them individually.

4.5 Machine Learning Modelling and Prediction

In this section we presented the main results of the ML models constructed to model and predict housing prices in Nairobi County. The modelling stage involved training each and every ML/DL model on housing data from Nairobi County and consequently fine-tuning some of the models' hyper-parameters to optimize the models. On the other hand, the prediction stage

involved utilizing part of the data that was set aside (test dataset) before training the models to forecast house sale prices based on the features in the test dataset. This was done in order to assess each model's predictive ability by means evaluation metrics. Additionally, prediction constituted forecasting house sale prices based on real world data.

Firstly, qualitative data analysis was conducted to give a bird's view of the secondary dataset obtained from https://www.property24.co.ke/. The second part of this section presented the quantitative results of the Machine Learning models utilized in modelling and predicting the house sale price and consequently meeting the study's objectives.

4.5.1 Exploratory Data Analysis

In this section, quick summary statistics of the target variable and the quantitative features were provided. Graphical methods were also adopted to identify underlying data distributions, patterns and trends.

4.5.1.1 Dataset Description

The dataset was retrieved from Property24.co.ke, a platform where different types of houses and other properties are listed for rent and sale. Web Scrapping of the secondary data was carried out using Python's Beautiful Soup library on January 31st, 2023. The initial dataset consisted of 51,055 observations and 8 variables of property listings on Maisonettes/Villas, Townhouses and apartments that were published on the website as at January 31st, 2023.
Table 4.1 below presented a short description of the dataset.

Table 4.1 Nairobi County Dataset

Variable	*Description*
Title	Description of the type of the house and its location.
Address	A detailed description of the house's specific location.
Region	Region in Nairobi County or suburbs where house is located
Bedrooms	Number of bedrooms in the house
Bathrooms	Number of bathrooms in the house
Parking	Number of parking slots of the house
Floor Area	Size of the floor area in squared meters
Price	The price of the house
Short Description	A short description of the house features and location.
Type	The type of the house.

Data cleaning and feature engineering was done in the data pre-processing steps that included; dropping of missing values, one-hot-encoding of categorical features, excluding extreme values using the Inter-Quartile Range (IQR) method, scaling the target variable to million Kenya Shillings, and feature scaling through normalization for ML models and standardization of the DL model.

Consequently, the final dataset utilized for modelling and making predictions reduced to 25,125 observations and 6 variables namely; sale price (target variable), location, bedrooms, bathrooms, house type and parking (features).

Table 4.2 below gave a random sample of 7 observations, the scaled target variable and the 5 features. The first column represents the identities (keys) of the seven randomly selected data points in the training set. The target variable, Sale_Price is scaled in Kes. 1,000,000. For the features, Location is a categorical variable containing 103 different locations in Nairobi County as values, Bedrooms is a feature that measures the number of bedrooms in the house, Bathrooms is a feature that measures the number of bathrooms in the house, Type is a categorical feature with values; Maisonette/Villa, Townhouse, and Apartment while Parking is a feature that measures the number of parking lots for a given house.

Table 4.2 Random sample of the dataset

```
df7.sample(7)
```

	Location	Bedrooms	Bathrooms	Type	Parking	Sale_Price
7696	Kileleshwa	2.0	3.0	Apartment	2.0	7.0
5509	Kiambu Road	4.0	4.0	Maisonette/Villa	3.0	30.0
12603	Kilimani	1.0	1.0	Apartment	1.0	3.5
20573	Langata	3.0	3.0	Apartment	2.0	21.0
41231	Spring Valley	5.0	6.0	Maisonette/Villa	3.0	110.0
12519	Kilimani	1.0	1.0	Apartment	1.0	3.5
27039	Lavington	5.0	5.0	Townhouse	3.0	70.0

Additionally, Table 4.3 provided a summary of descriptive statistics of the numerical measures of location and dispersion in the dataset that was used for ML modelling and prediction. The first column presents the various descriptive measures.

Table 4.3 Descriptive Statistics

```
df7.describe()
```

	Bedrooms	Bathrooms	Parking	Sale_Price
count	25125.000000	25125.000000	25125.000000	25125.000000
mean	3.386406	3.700911	2.226209	32.532572
std	1.255183	1.436839	0.979970	33.631013
min	1.000000	1.000000	0.000000	3.000000
25%	3.000000	3.000000	2.000000	9.800000
50%	3.000000	4.000000	2.000000	17.000000
75%	4.000000	5.000000	2.000000	46.000000
max	7.000000	7.000000	6.000000	145.000000

As presented in Table 4.3, the average number of bedrooms was 3, average number of bathrooms was 4, and average number of parking slots was 2 while the mean house sale price was Kes. 32.533 million. The minimum number of bedrooms, bathrooms, parking slots and sale price was 1, 1, 0 and Kes. 3.0 million while the maximum was 7, 7, 6 and Kes. 145.0 million respectively.

4.5.1.2 Distribution of the Target Variable

Data pre-processing process involved removal of extreme values and missing values among other steps. The Inter-Quartile Range (IQR) method was used to identify and exclude outliers in the original dataset. Following this process, it was determined that the values of the target variable (sale price) that were greater than the upper whisker (Kes. 147.8 million) of the IQR measure accounted for 6.08% of the total observations. This result was further compounded by the data distribution of house sale price as shown in Figure 4.15 below.

As presented in Figure 4.15, the distribution of the target variable was skewed to the right implying that the average house selling price was greater than 50% of the sorted house selling prices list. Notably, most of the listed houses in the dataset were valued under Kes. 115 million.

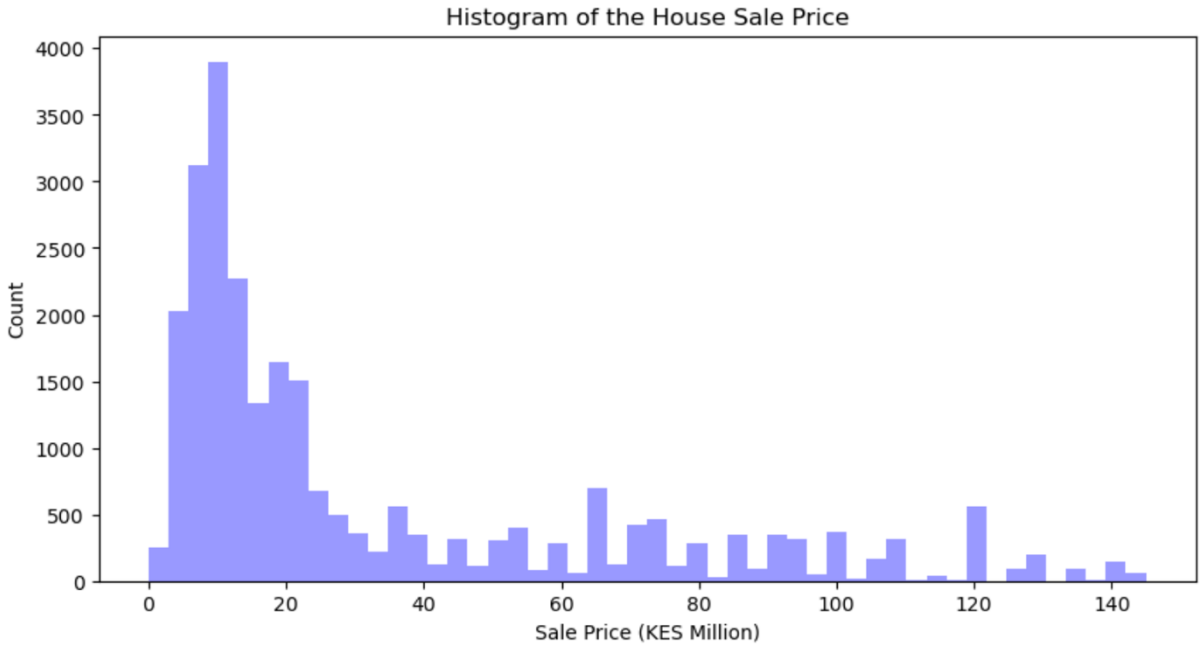


Figure 4.15: Histogram of House Sale Price

4.5.1.3 Correlation Analysis

A correlation heat-map of the features and the target variable was generated to examine which of the features influenced the target variable more as well as the degree of association among the features themselves. The results are shown in Figure 4.16.

Figure 4.16: Correlation Analysis

The features considered were the quantitative features namely; number of bedrooms, bathrooms and parking lots. Notably, the features; type of house and location were excluded in the correlation analysis as they were converted into dummies during the one-hot-encoding data pre-processing stage. As a result, the features; type of house resulted into 3 other features while location of the house results into 103 other features hence correlation analysis in this case was not feasible.

As shown in Figure 4.16, the feature with the strongest linear association with the target variable sale price was number of bedrooms recording a correlation coefficient of 0.71. Number of bathrooms was the next most significant feature in determining the level of sale price with a correlation coefficient of 0.61 while number of parking slots was the least significant in determining the level of sale price with a correlation coefficient of 0.47. Notably, all the features were positively correlated with the target variable indicating that a positive change in the features translated to a positive change in the target variable. The correlation coefficients between the features themselves were; bedrooms vs bathrooms = 0.88, bedrooms vs parking = 0.55, parking vs bathrooms = 0.49. Likewise, all feature pairs had positive correlations.

4.5.2 Machine Learning Model Training, Optimization, and Testing

This section addressed objective two whereby the main results from developing, training & fine-tuning (modelling), and testing the ML models were presented. Consequently, the best ML candidates were selected based on their predictive performance on the test set. The models were trained on 80% of the dataset and tested on 20% of the data. Two performance metrics were used to evaluate the models' performance on the test set. The Root Mean Squared Error (RMSE) provided the average model prediction error in the values of the target variable sale price and the absolute fit of the model to the data while R-Squared provided the proportion of the target variable's variation that was explained by the set of features utilized during the prediction phase.

The RMSE is the square root of the sum of squares of the residuals as shown in the formula below;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{i(actual)} - \hat{y}_{i(predicted)})^2}$$

The R-Square Score determines the prediction accuracy of the model in terms of the residual distances as shown in the formula below;

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i(actual)} - \hat{y}_{i(predicted)})^2}{\sum_{i=1}^{n}(y_{i(actual)} - \bar{y}_{i(actual)})^2}$$

A total of 14 Machine Learning architectures were developed, optimized and tested. They include;

a) Stand-alone Learning Models

Linear Regression (LR)

Support Vector Regression (SVR)

K – Nearest Neighbors (K-NN) Regression

Decision Trees (DT) Regression

b) Bagging Ensemble Learning Models

Extra Decision Trees (Ext-DT) Regression

Random Forest (RF) Regression

c) Boosting Ensemble Learning Models

Gradient Boosted Machine (GBM) Regression

Extreme Gradient Boosting (XG-Boost) Regression

Light Gradient Boosted Machine (Light GBM) Regression

Category Boosting (CatBoost) Regression

d) Penalized Regression

Ridge Regression (Ridge)

Lasso Regression (Lasso)

Elastic Net Regression (ENet)

e) Deep Learning

Multi-Layer Perceptron (MLP) Regression

Firstly, stand-alone models were trained, optimized and evaluated. Next, ensemble machine learners were considered and comparisons made to the base models. Ensemble learning is a meta-approach to ML that combines multiple models, usually weaker learners, for better predictive performance of the final model. Penalized regression was also carried out to assess the effect of an expanded feature space. Penalized regression involves variable selection and/or shrinkage for better predictive capacity of the model. Lastly, a deep learning model was constructed and comparison in performance made to the ML architectures. Deep Learning is a special case of ML where models are constructed to mimic the functioning of the human brain's neurons.

Model predication evaluation results on the test dataset of each model are provided in terms of the RMSE and R-Squared scores.

4.5.2.1 Linear Regression (LR)

A Linear Regression model was trained on the training set and its prediction performance on the test set recorded. The model had a RMSE of 15.086041 and an R-Square score of 79.47%.

4.5.2.2 Support Vector Regression (SVR)

A Support Vector Regression model was trained and a polynomial kernel utilized to transform the linearly inseparable data to linearly separable data and also aid in reducing the computational cost in high dimensional space of the features. Performance results of the model on the test set showed that the SVR model had a RMSE of 12.606531 and an R-Square score of 85.66%.

4.5.2.3 K-Nearest Neighbors Regression (KNN)

A K-Nearest Neighbors Regression model was trained on the training set and the Grid Search Cross-Validation method used to optimize the model by fitting a set of different models for specific values of the neighbors hyper-parameter. Performance results of the model on the test set showed that the optimal value of the neighbors hyper-parameter was 10 and the model had a RMSE of 11.634257 and an R-Square score of 87.79%.

4.5.2.4 Decision Trees Regression (DT)

A Decision Trees Regression model was trained on the training set and the prediction results of the model on the test set noted. The model had a RMSE of 12.022199 and an R-Square score of 86.96%.

To improve the predictive performance of the ML models, Ensemble Learning was adopted where weaker Machine Learners were combined through bagging and boosting to construct stronger learners. Extra Decision Trees Regression and Random Forest Regression were constructed through bagging of a set of individual Decision Trees. Bagging/Bootstrap aggregation is an ensemble learning technique that combines multiple weak learners in a parallel framework where each model is independent of each other during the training session. The ultimate predictions are obtained by averaging all the predictions from the weak learners. This learning technique helps to reduce variance in the resultant stronger learner hence by extension it addresses the challenge of over-fitting.

4.5.2.5 Extra Decision Trees Regression (Ext-DT)

An Extra Decision Trees Regression model was trained on the training set and the hyper-parameter values of estimators, minimum number of samples splits and minimum number of samples leaves set at 100, 2 and 1 respectively. Performance results of the model on the test set were then obtained. Notably, the Extra Decision Trees Regression model performed slightly better than the base Decision Trees Regression and recorded a RMSE of 11.49913 and an R-Squared score of 88.27%.

4.5.2.6 Random Forest Regression (RF)

A Random Forest Regression model was trained on the training set and the Grid Search Cross-Validation technique utilized to optimize the model by examining different fits for optimal values of the hyper-parameters. Performance results of the model on the test set showed that the optimal values of the estimators, minimum number of samples splits and minimum number of samples leaves were 200, 15 and 1 respectively. Additionally, the model outperformed the base Decision Trees Regression model and the Extra Decision Trees Regression model. It recorded a RMSE of 11.247315 and an R-Square score of 88.59%.

Moreover, ensemble learners based on the gradient boosting procedure were considered. Boosting ensemble learning involves combining weak learners in a sequential fashion. In this case each subsequent learner attempts to correct and minimize the errors of the preceding learner. Models are basically learned sequentially with the final stronger learner being a weighted mean of the weak learners. Essentially each learner boosts the predictive performance of the ensemble.

The first model considered was the base Gradient Boosted Machine Regression model followed by XGBoost, Light GBM and finally CatBoost.

4.5.2.7 Gradient Boosted Machine Regression (GBM)

A Gradient Boosted Machine Regression model was trained on the training set and the hyper-parameter values; estimators, maximum depth of the tress and minimum number of samples splits set at 500, 100 and 10 respectively. Evaluation results of the model on the test set data were then obtained. From the results, the GBM regression model had a RMSE of 11.609453 and an R-Squared Score of 87.84%.

4.5.2.8 Extreme Gradient Boosting Regression (XG Boost)

An Extreme Gradient Boosted Machine Regression model was trained on the training set and the values of the hyper-parameters; maximum tress depth, minimum child weight and eta set at 100, 12 and 0.3 respectively. Evaluation results of the model on the test set were obtained The XG Boost regression model slightly out-performed the base GBM regression model and recorded a RMSE of 11.412821 and an explained variance proportion in sale price of 88.25%.

4.5.2.9 Light Gradient Boosted Machine Regression (Light GBM)

A Light Gradient Boosted Machine Regression model was fitted on the training set and the values of the hyper-parameters; feature fraction, maximum depth, number of leaves and maximum bin set at 0.9, 8, 256 and 512 respectively with 100,000 iterations being achieved. Evaluation results of the model on the test set showed that the Light Gradient Boosted Machine regression model out-performed the XG-boost regression model on the test set and recorded a RMSE of 11.216350 and an R-Squared score of 88.65%.

4.5.2.10 Category Gradient Boosting Regression (CatBoost)

A Category Gradient Boosting Regression model was trained on the training set and the Grid Search Cross-Validation technique utilized to optimize the model by examining different fits for optimal values of the hyper-parameters. Performance results of the model on the test set showed that CatBoost regression model recorded the optimal values of the hyper-parameters depth, iterations, learning rate and the lasso leaf regression as 10, 200, 0.1 and 0.2 respectively. Moreover, the model recorded a RMSE of 11.317758 and a variance in the sale price score of 88.44% on the test set.

During the data pre-processing stage, dummies for the features; house location and house type were obtained using the one-hot-encoding technique. As a result, there was an expansion in the feature space which could potentially lead to the *curse of high dimensionality*. This refers to the event where data becomes sparser in high-dimensional feature spaces making models developed for low-dimensional spaces to be infeasible (Li & Liu, 2017). To counter this, often penalized regression is carried out where penalties are applied on the feature coefficients based on the features' relevance in explaining the target variable. Three penalized regression models were trained namely; the Ridge, Lasso and Elastic Net penalty which is a weighted combination of former two.

4.5.2.11 Ridge Regression (Ridge)

A Ridge Regression model was trained on the training set using the l2-penalty. The l2 penalty value was set at 0.01 after trying a set of different values for the highest model prediction accuracy. Performance results of the model on the test set were consequently obtained. The Ridge regression model had a RMSE of 15.086667 and an R-Squared score of 79.46% on the test set.

4.5.2.12 Lasso Regression (Lasso)

A Lasso Regression model was fitted on the training set utilizing the l1-penalty. The penalty value was set at 0.0015 after trying a set of different values for the highest model prediction accuracy. Evaluation results of the model on the test set were obtained. The results showed that the Lasso regression model performed more or less the same as the Ridge regression model. The model recorded a RMSE of 15.085996 and an R-Squared score of 79.47% on the test set.

4.5.2.13 Elastic Net Regression (E-Net)

An elastic Regression model was trained on the training set and the values of the l1 and l2 penalties set at 0.01 and 0.9 respectively. This followed a Cross-Validation process to obtain the optimal pair of l1 and l2 penalties. Performance results of the model on the test set were the recorded. The results showed that the Elastic Net regression model performed slightly poorly than both Ridge and Lasso models. The model recorded a RMSE of 15.405066 and an R-Squared score of 78.59% on the test dataset.

Upon training the ML models, a Deep Learning model was trained to examine if there was a significant difference in model performance compared to the ML models. In particular, a Multi-Layer Perceptron was constructed using Python's Keras functional API.

4.5.2.14 Multi-Layer Perceptron Regression (MLP)

A MLP was constructed utilizing three hidden layers with neurons; 30, 30 and 16 and the Rectified Linear Unit (ReLU) activation function in the hidden layers as well as the output layer. Consequently, the DL architecture was compiled using the Adams optimizer and the mean squared error loss function. The MLP was trained at 500 epochs and a batch size of 128. Evaluation results of the model on the test set were obtained. The results indicated that the MLP had a RMSE of 11.799989 and an R-Squared score of 87.44%.

4.5.3 Machine Learning Model Performance Comparison

In this section, a comparison analysis was done to select the best performing ML models based on the prediction results on the test dataset. Table 4.4 gives a summary of the ML models' prediction performance results.

Table 4.4 Machine Learning Model Performance Comparisons

ML/DL Model	RMSE Score	R-Squared Score
Linear Regression	15.086041	79.47%
Support Vector Regression	12.606531	85.66%
K-Nearest Neighbors	**11.634257**	**87.79%**
Decision Trees	12.022199	86.96%
Extra Decision Trees	11.40013	88.27%
Random Forest	**11.247315**	**88.59%**
Gradient Boosted Machine	11.609453	87.84%
XGBoost	11.412821	88.25%
Light GBM	**11.21635**	**88.65%**
CatBoost	11.317758	88.44%
Ridge	15.086667	79.46%
Lasso	**15.086099**	**79.47%**
E-Net	15.405066	78.59%
Multi-Layer Perceptron	11.799989	87.44%

Key

Bolded **Best model in the category**

- Stand - alone Learners
- Bagging Ensemble Learning
- Boosting Ensembe Learning
- Penalized Learning
- Deep Learning

As presented in Table 4.4, the K-Nearest Neighbors Regression Model was the best performing predictor in the class of stand-alone learners with the least RMSE of 11.634235 and the highest R-Squared Score of 0.8779.

For the case of Ensemble Learning using the bagging method, Random Forest Regression Model was the best performing predictor in the test set with the least RMSE of 11.247315 and an R-Squared Score of 0.8859. On the other hand, the best model in Ensemble Learning using the boosting architectures was the Light Gradient Boosted Machine Regression Model with the least RMSE of 11.216350 and highest R-Squared Score of 0.8865 in that category of models.

Additionally, the best performing Penalized Regression model was the Lasso Regression with a RMSE of 15.086099 and an R-Squared Score of 0.7959. Lastly, the Deep Learning

architecture MLP did not have a benchmark model in Deep Learning to compare its predictive performance to other than the fit Machine Learning Models. However, the MLP recorded a RMSE of 11.799989 with an R-Squared Score of 0.8744.

Following this analysis, it was concluded that the overall best performing ML prediction models based on this dataset were the Light GBM, Random Forest Regression, K-Nearest Neighbors Regression and the Lasso Regression in that order in the various categories.

Penalized Regression Models performed relatively poorly compared to the rest. This is attributable to the fact that Penalized Regression is ideal for problems in higher dimensional feature spaces while in our case the feature space was relatively small compared to the sample data points, that is, 59 features against 25,125 sample data points. Additionally, DL requires relatively huge datasets for the models to perform automated feature extraction and identify the underlying relationships and patterns in the dataset. This explains why our MLP performed relatively poorer compared to a majority of ML models above.

4.5.4 Machine Learning Prediction Results in Production

In this section we addressed objective 3 by utilizing the trained models to predict real housing pricing in practice. The top performing models in each of the categories were utilized in predicting real world data points.

The following locations were randomly selected to represent real world data points for housing price prediction purposes once the ML models are deployed to production.

The first cadre locations randomly chosen were; Karen, Runda and Kitisuru. The second cadre locations randomly chosen were Kilimani, Lavington and Kiambu Road while the third cadre locations randomly chosen were South C, Embakasi and Langata.

Firstly, we made house sale price predictions for the first cadre locations. Table 4.5 below presents various prediction instances for a Maisonette/Villa in Karen with 5 bedrooms, 5.5 bathrooms and 2 parking slots and a Maisonette/Villa in Runda or Kitisuru with 6 bedrooms, 7 bathrooms and 2 parking slots.

Table 4.5 Housing price predictions in Karen, Runda, and Kitusuru

Location	Model	Predicted sale price
Karen Runda Kitisuru	K-NN	Kes. 81.4 million Kes. 115.2 million Kes. 120.8 million
Karen Runda Kitisuru	Random Forest	Kes. 97.3 million Kes. 113.89 million Kes. 115.87 million
Karen Runda Kitisuru	Light GBM	Kes. 92.16 million Kes. 112.21 million Kes. 115.04 million
Karen Runda Kitisuru	Elastic Net	Kes. 90.32 million Kes. 104.38 million Kes. 96.28 million

As shown in Table 4.5, using a K-NN regression model, a Maisonette/Villa in Karen with 5 bedrooms, 5.5 bathrooms and 2 parking slots would cost Kes 81.4 million. Additionally, a Maisonette/Villa in Runda with 6 bedrooms, 7 bathrooms and 2 parking slots would cost Kes. 115.2 million while a Maisonette/Villa in Kitusuru with 6 bedrooms, 7 bathrooms and 2 parking slots would sale for Kes. 120.8 million.

Using a Random Forest regression model, a Maisonette/Villa in Karen with 5 bedrooms, 5.5 bathrooms and 2 parking slots would cost Kes. 97.3 million. Additionally, a Maisonette/Villa in Runda with 6 bedrooms, 7 bathrooms and 2 parking slots would cost Kes. 113.89 million while a Maisonette/Villa in Kitusuru with 6 bedrooms, 7 bathrooms and 2 parking slots would cost Kes. 115.87 million.

Using a Light GBM regression model, a Maisonette/Villa in Karen with 5 bedrooms, 5.5 bathrooms and 2 parking slots would cost Kes. 92.16 million. Additionally, a Maisonette/Villa in Runda with 6 bedrooms, 7 bathrooms and 2 parking slots would cost Kes. 112.21 million while a Maisonette/Villa in Kitusuru with 6 bedrooms, 7 bathrooms and 2 parking slots would cost Kes. 115.04 million.

Lastly, using a Lasso regression model, a Maisonette/Villa in Karen with 5 bedrooms, 5.5 bathrooms and 2 parking slots would cost Kes. 90.32 million. Additionally, a Maisonette/Villa

in Runda with 6 bedrooms, 7 bathrooms and 2 parking slots would cost Kes. 104.38 million while a Maisonette/Villa in Kitusuru with 6 bedrooms, 7 bathrooms and 2 parking slots would cost Kes. 96.28 million.

Secondly, we made house sale price predictions for the second cadre locations. Table 4.6 below presents various prediction instances for a Townhouse in Kilimani, Lower Kabete or Kiambu Road with 5 bedrooms, 4 bathrooms and 4 parking slots.

Table 4.6 Housing price predictions in Kilimani, Lower Kabete, and Kiambu Road

Location	Model	Predicted sale price
Kilimani	K-NN	Kes. 45.3 million
Lower Kabete		Kes. 94.3 million
Kiambu Road		Kes. 40.19 million
Kilimani	Random Forest	Kes. 48.59 million
Lower Kabete		Kes. 62.56 million
Kiambu Road		Kes. 37.26 million
Kilimani	Light GBM	Kes. 55.79 million
Lower Kabete		Kes. 78.09 million
Kiambu Road		Kes. 38.31 million
Kilimani	Elastic Net	Kes. 60.81 million
Lower Kabete		Kes. 83.3 million
Kiambu Road		Kes. 42.62 million

As shown in Table 4.6, using a K-NN regression model, a Townhouse in Kilimani with 5 bedrooms, 4 bathrooms and 4 parking slots would cost Kes 45.3 million. Additionally, a Townhouse in Lower Kabete with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 94.3 million while a Townhouse in Kiambu Road with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 40.19 million.

Using a Random Forest regression model, a Townhouse in Kilimani with 5 bedrooms, 4 bathrooms and 4 parking slots would cost Kes. 45.3 million. Additionally, a Townhouse in Lower Kabete with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 94.3 million while a Townhouse in Kiambu Road with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 47.8 million.

Using a Light GBM regression model, a Townhouse in Kilimani with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 55.79 million. Additionally, a Townhouse in Lower

Kabete with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 78.09 million while a Townhouse in Kiambu Road with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 38.31 million.

Lastly, using a Lasso regression model, a Townhouse in Kilimani with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 60.81 million. Additionally, a Townhouse in Lower Kabete with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 83.3 million while a Townhouse in Kiambu Road with 5 bedrooms, 4 bathrooms and 4 parking slots would sale for Kes. 42.62 million.

Thirdly, we made house sale price predictions for the third cadre locations. Table 4.7 below presents various prediction instances for an Apartment in South C, Lang'ata or Embakasi with 3 bedrooms, 3 bathrooms and 2 parking slots.

Table 4.7 Housing price predictions in South C, Langata, and Embakasi

Location	Model	Predicted sale price
South C	K-NN	Kes. 15.42 million
Lang'ata		Kes. 20.2 million
Embakasi		Kes. 9.58 million
South C	Random Forest	Kes. 15.37 million
Lang'ata		Kes. 20.37 million
Embakasi		Kes. 10.85 million
South C	Light GBM	Kes. 14.96 million
Lang'ata		Kes. 19.62 million
Embakasi		Kes. 7.22 million
South C	Elastic Net	Kes. 14.83 million
Lang'ata		Kes. 14.83 million
Embakasi		Kes. 11.16 million

As shown in Table 4.7, using a K-NN regression model, an Apartment in South C with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 15.42 million. Additionally, an Apartment in Langata with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 20.2 million while an Apartment in Embakasi with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 9.58 million.

Using a Random Forest regression model, an Apartment in South C with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 15.37 million. Additionally, an Apartment in Langata with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 20.37 million while an Apartment in Embakasi with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 10.85 million.

Using a Light GBM regression model, an Apartment in South C with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 14.96 million. Moreover, an Apartment in Langata with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 19.62 million while an Apartment in Embakasi with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 7.22 million.

Lastly, using a Lasso regression model, an Apartment in South C with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 14.83 million. Similarly, an Apartment in Langata with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 14.83 million while an Apartment in Embakasi with 3 bedrooms, 3 bathrooms and 2 parking slots would sale for Kes. 11.16 million.

4.6 Key Informants Data

Objective four in this research paper sought to determine what policy recommendations can be put in place to influence housing prices determination in Nairobi County and beyond. This section partly addresses objective four by thematically analysing data that was obtained from Key Informants at the State Department.

Six informants were interviewed on housing prices in Nairobi County. According to each informant, the price ranges for homes varied across Nairobi County, as shown in the Table 4.8.

Table 4.8 Key Informants Price Range Estimations

Informant Identity	Price Range
1	3-100 million – General
2	1-4 million – 1 bedroom
3	10-250million – General
4	5-100 million - General
5	1.55-7.5 million – 1-2 bedroom

Furthermore, the following elements were listed as having an impact and being taken into account when determining property prices:

i) Housing supply and demand are influenced by dwelling types and sizes in the market.
ii) Government's directives incorporated into the development framework.
iii) Cost of land on which the property sits.
iv) Property's location.
v) Cost of professional fees.
vi) Cost of construction.
vii) Cost of onsite infrastructure.
viii) Cost of marketing.
ix) Cost of financing by partnering financial institutions.
x) Developer's profit margin.

Informants further stated that the aspects and features taken into account in construction of the affordable houses were:

i) House affordability, factoring in the estimated income level of intended beneficiaries.
ii) The land size and house size
iii) The location of projects which affects the cost of land.
iv) Availability of related infrastructure.
v) The availability of government funding.
vi) The quality of the house and finishing.
vii) The speed of construction.
viii) The housing demand.

Following these interviews, the general consensus was that Kenyans are unaware of how housing prices are determined. Half of the informants thought housing prices were poorly determined, while the other half thought otherwise. Another observation was that amenities were poorly taken into account while determining housing prices. Further, given that the government is a non-profit entity, there were numerous subsidies in the pricing model compared to housing pricing in the private sector. This resulted in disparities in housing prices in the two cases. Only two of the six informants supported the subsidy-based pricing system; while the other four objected this on the grounds that building materials prices and housing market prices fluctuated. One of the informants proposed the use of hedonic pricing models for the determination of housing prices.

To the best of the informant's knowledge, majority of stakeholders were not privy to ML models being used in housing price determination. They acknowledged that in the local environment, Microsoft Excel based financial models have flooded the market as the go-to tools in housing price forecasting. This is as a result of varying priorities of market players, lack of technological knowledge as well as awareness of the workings of the models and the set government directives. Although, the majority (4) believe that implementation of ML technologies would lead to more accurate estimation of housing prices, the remaining two were unsure of the outcome since they believed that accuracy depended on technical knowhow of adopting and implementing AI/ML technologies as well as access to current and correct datasets in the housing sector.

Following these findings, consensus existed among the six participants that the Government should commit resources and efforts for the adoption and implementation of new age technologies such as Artificial Intelligence and Machine Learning in guiding housing price determination in the housing sector.

4.7 Discussions of Findings

In this section an overview of the research study's findings was presented. Firstly, we discussed findings from the primary data followed by findings from the secondary data.

4.7.1 Features that significantly determine housing pricing

This study was aimed at examining how ML technology can be leveraged in modelling and predicting housing prices in Nairobi County using secondary datasets. Additionally, primary data was collected from Shauri Moyo and Kibra areas in Nairobi County as well the State Department so as to obtain actionable insights on features that significantly influence housing pricing from both the supply and demand sides of the housing sector.

Following the analysis of the primary datasets, the following housing features were found to be significant in determining housing pricing in Nairobi County: See Table 4.9.

Table 4.9 Ranking of housing features

Feature	Proportion of participant's that **Strongly Agreed** that the feature significantly determined housing pricing
Presence of tarmacked road within 0-5 KMs	75.16%
Access to electricity connection/generator	74.95%
Presence of a hospital within 0-5 KMs	64.84%
Presence of a school within 0-5 KMs	64.00%
Internet connectivity in the house	55.49%
Age of the house	54.60%
Location of the house	49.47%
Presence of piped water	38.53%
Size of the house	34.11%
Number of rooms in the house	30.80%

Following the analysis of secondary data, a correlation heat-map was generated to assess which feature significantly influenced housing pricing based on the degree and direction of association as measured by the correlation coefficient. Table 4.8 presented a summary of features and their corresponding correlation coefficients vis-à-vis the target variable house sale price.

Table 4.10 Correlation measure of feature significance

Feature	**Correlation Coefficient** (Feature vs House Sale Price)
Number of Bedrooms	0.71
Number of Bathrooms	0.61
Number of Parking slots	0.47

Notably, all the features were positively correlated with the target variable indicating that a positive change in the features translated to a positive change in the target variable. The most significant features in changing the level of house sale price were the number of bedrooms, followed by the number of bathrooms and finally number of parking lots.

Following ML modelling, all the 14 ML/DL candidates considered recorded an R-Squared score above 75% on the test set. R-Squared Score measures the amount of variation in the target variable that is explainable by the variations in the set of features used in the regression analysis. As such, in all the ML/DL architectures that were developed and trained, in addition to number of bedrooms, number of bathrooms and number of parking lots, the features; location of the house and type of the house (Apartment, Maisonette/Villa, Townhouse) all accounted for at least three-quarters of the changes in the predicted house sale price. Notably, half of the trained ML models had an R-Squared score of at least 88% indicating that the 5 features; bedrooms, bathrooms, parking slots, location, and type of house significantly explained 88% of the changes in the predicted house sale price while only 12% of the changes in the predicted house sale price was accounted for by other features not included in the ML development in each case.

4.7.2 Trained and Evaluated Machine Learning models

Fourteen Supervised Learning models were considered namely: Linear Regression (LR), Support Vector Regression (SVR), K–Nearest Neighbor (KNN) Regression, Decision Trees (DT) Regression, Extra Decision Trees (Ext-DT) Regression, Random Forest (RF) Regression, Gradient Boosted Machine (GBM) Regression, Extreme Gradient Boosting (XG-Boost) Regression, Light Gradient Boosted Machine (Light GBM) Regression, Category Boosting (CatBoost) Regression, Ridge Regression (Ridge), Lasso Regression (Lasso), Elastic Net Regression (E-Net), and the Multi-Layer Perceptron (MLP) Regression. These models were trained, optimized and evaluated on a dataset from Property24.co.ke a site that contains listings of among other properties, houses and apartments for rent and sale in Nairobi County and other major towns in Kenya. To measure the models' performance in making predictions on the test dataset, two statistical scores were used namely: Root Mean Squared Error (RMSE) and R-Squared Score.

As per the results, the overall best performing ML models based on the predictive accuracy were:

i) Light GBM Regression model - Ensemble Learning (Boosting).
RMSE = 11.216350 and R^2 = 88.65%

ii) Random Forest Regression model - Ensemble Learning (Bagging)
RMSE = 11.247315 and R^2 = 88.59%

iii) K-Nearest Neighbors Regression model - Stand-alone model
RMSE =11.634257 and R^2 = 87.79%

iv) Lasso Regression model - Penalized Regression
RMSE =15.086099 and R^2 =79.47%

4.7.3 Prediction of housing prices using trained ML models based on unseen data

To examine how robust the trained and optimized ML models were, the four ML candidates discussed above were used to predict sale house prices based on unseen features during the training phase. The inputs were user-defined.

The four ML models provided predictions on real world data points that did not differ as much. Therefore, our proposal is; going by the housing sector domain knowledge, one could make such predictions with minimal error if they were to average the predictions from across the four ML models or even select the model that had the overall lowest RMSE in this case Light GBM.

CHAPTER FIVE

DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

Discussion of the research findings, conclusions made and recommendations drawn from the research were presented in this section. Summarized outcomes for each study objective have been presented in line with this study's objectives. Lastly, recommendations arrived at from findings of the study at hand are presented and identified knowledge gaps are highlighted for research in the future.

5.2 Summary of findings

This section summarizes research results based on the first three objectives that the research aimed to achieve. The final objective will be under policy recommendations in the next section.

5.2.1 Features that significantly determine housing pricing in Nairobi County.

The study revealed that the following housing features, in no specific order, significantly determine housing pricing in Nairobi County:

i) Presence of tarmacked road within 0-5 KMs
ii) Access to electricity connection/generator
iii) Presence of a hospital within 0-5 KMs
iv) Presence of a school within 0-5 KMs
v) Internet connectivity in the house
vi) Age of the house
vii) Location of the house
viii) Presence of piped water
ix) Size of the house
x) Number of bedrooms
xi) Number of bathrooms
xii) Number of parking slots
xiii) Type of the house (for example, Apartment, Maisonette/Villa, Townhouse)

These results were backed by evidence in both the qualitative and quantitative data analysed.

5.2.2 Review of the Trained and Evaluated Machine Learning models

Following the training, fine-tuning and evaluation of fourteen Supervised Learning models namely: Linear Regression (LR), Support Vector Regression (SVR), K–Nearest Neighbors (KNN) Regression, Decision Trees (DT) Regression, Extra Decision Trees (Ext-DT) Regression, Random Forest (RF) Regression, Gradient Boosted Machine (GBM) Regression, Extreme Gradient Boosting (XG-Boost) Regression, Light Gradient Boosted Machine (Light GBM) Regression, Category Boosting (CatBoost) Regression, Ridge Regression (Ridge), Lasso Regression (Lasso), Elastic Net Regression (E-Net), and the Multi-Layer Perceptron (MLP) Regression, a ranking was done based on model predictive accuracy on the test dataset as measured by RMSE and R-Squared Scores.

Table 5.1: ML/DL model ranking based on predictive accuracy

ML/DL Model	RMSE	R-Squared
Light GBM	11.216350	88.65%
Random Forest	11.247315	88.59%
CatBoost	11.317758	88.44%
Extra Decision Trees	11.400130	88.27%
XGBoost	11.412821	88.25%
Gradient Boosted Machine	11.609453	87.84%
K-Nearest Neighbors	11.634257	87.79%
Multi-Layer Perceptron	11.799989	87.44%
Decision Trees	12.022199	86.96%
Support Vector Regression	12.606531	85.66%
Linear Regression	15.086041	79.47%
Lasso	15.086099	79.47%
Ridge	15.086667	79.46%
E-Net	15.405066	78.59%

As summarized above, the best ML candidate based on the two prediction evaluation metrics was the Light Gradient Boosted Machine (Light GBM) regression model with a RMSE of 11.216350 and an R-Squared value of 88.65%. On the other hand, the least performing model was the Elastic Net (E-Net) regression model which recorded a RMSE of 15.405066 and an R-Squared score of 78.59%.

Generally, ML models based on the ensemble learning techniques were the best performers (top six) in predicting housing prices in Nairobi County based on the test dataset. On the other hand, base Linear Regression model and the Penalized Regression models were the least performers.

5.2.3 Review of ML prediction of housing prices based on unseen data

To examine the applicability of the trained and optimized ML models in the real world, four ML candidates that emerged as top performers in the categories; Stand-alone models, Ensemble Learning (Bagging), Ensemble Learning (Boosting) and Penalized regression were utilized.

It was observed that the predictions did not differ by much in most cases. For instance, random prediction results from the real-world data showed that;

i) A 3-bedroom Apartment with 2 bathrooms and two parking lots in South C was projected to sale for;
   Kes. 15.42 million (K-NN regression)
   Kes. 15.37 million (Random Forest regression)
   Kes. 14.96 million (Light GBM regression)
   Kes. 14.83 million (Lasso regression)
ii) A 5-bedroom Maisonette/Villa with 5.5 bathrooms and 2 parking lots in Karen was projected to sale for;
   Kes. 81.4 million (K-NN regression)
   Kes. 97.3 million (Random Forest regression)
   Kes. 92.16 million (Light GBM regression)
   Kes. 90.32 million (Lasso regression)
iii) A 5-bedroom Townhouse with 4 bathrooms and 4 parking lots in Kiambu Road was projected to sale for;
   Kes. 40.19 million (K-NN regression)
   Kes. 37.26 million (Random Forest regression)
   Kes. 38.31 million (Light GBM regression)
   Kes. 42.62 million (Lasso regression)

As presented in the summary above, it was noted that variations in the predictions from the four models were relatively high in first and second cadre locations vis-à-vis the third cadre location.

5.3 Conclusions

The study established that there exists considerable association between certain features of a house and the sale price of the house. The study further revealed that there are several ML/DL models used in prediction of housing prices in the developed world as presented in the literature review that were implementable in the local housing sector particularly in predicting housing prices in Nairobi County. These ML-based housing price prediction frameworks are a paradigm shift from classical housing price modelling and forecasting tools. AI and ML tools are at the core of technological advancements and availability of big data which characterizes today's housing sector among other sectors in the Kenya economy.

The efficiency of these ML-based prediction models is up-scaled as the models are put in more use to predict housing prices. Constant model exposure to new data in the real world translates to more self-training and retraining consequently resulting to a more robust and efficient prediction model. This integrated process is facilitated by incorporating data pipelines and automating data pre-processing, training, optimization and testing frameworks of the models.

The findings in this study are important as they will enable real estate investors, appraisers, tax assessors, mortgage lenders, insurers and policy makers to derive actionable insights and reach informed decisions on housing prices. The study does so through leveraging ML technologies to yield data-driven scalable solutions that inform policies in the housing sector. Lastly, the study will also contribute towards the housing agenda by the GOK.

5.4 Recommendations

5.4.1 Recommendations to Policymakers

Based on the findings, the researcher makes recommendations to policy makers in the housing sector to allocate resources and direct efforts towards collecting and recording as much quality data as possible on features that determine housing pricing as well as the application of new-age data modelling techniques such ML in respect of the same. As noted in this study, implementation of ML technologies requires sufficient data in order to adequately train and

optimize them to yield quality results that would guide policy decisions. This would be achieved through engaging key stakeholders namely: developers; investors; appraisers; tax assessors; real estate market participants including mortgage lenders and insurers on cutting age technologies then sensitizing them on the importance of data collection, sharing, and recording keeping as this information would aid them all. This information should then cascade downwards to all to enable data-driven decisions on features that determine housing pricing.

The study further recommends that the Kenyan Government institute data safety policies to enhance the processing of data for actionable insights through ML/AI usage by stakeholders in the real estate sector.

Additionally, Benchmarking exercises should also be undertaken in countries such as Singapore, United Kingdom and the United States where the use of Machine Learning in prediction of real estate pricing has proven very useful. Further, Government through the State Department for Housing should train its officers on the use of data and data-driven technologies for ease and effective prediction of housing pricing in Nairobi County.

From the study, it is evident that price is a great determinant of home ownership in Kenya. There is an urgent need for the government to implement and/or boost policies focused towards affordability of homes by Kenyans from all cadres of society. This can be attained through: identification of more public land for housing; reduction of registration costs at the Lands Registry; reduction of fees for various approvals required by developers which amounts are passed down to end buyers.

The government of Kenya should fast track digitization of processes of the Lands Registry as this is where ownership of homes is registered. This would encourage home ownership by: ensuring sanctity of titles as fraud would be negated; fast tracking registration processes and boosting the confidence required by banks in providing mortgages to home owners. Once the use of machine learning is achieved in Nairobi County as proposed by this study, it should be replicated across the other 46 Counties in Kenya.

5.5 Suggestion for Further Studies

In the future, the kind of dataset used in this work can be enriched by houses' visual features, such as house interior and exterior images to construct more efficient predictors in the housing sector in Kenya based on advance ML/DL frameworks such as the Convolutional Neural Networks. Furthermore, this research work can be used as a foundation for other kinds of studies in areas such as; the stock exchange, petroleum and oil markets, banking and insurance industries and the manufacturing processes for demand forecasting etc.

REFERENCES

Abirami, S., & Chitra P., (2020). Energy-efficient edge based real-time healthcare support system. In Advances in Computers (1st ed., Vol. 117, Issue 1). Elsevier Inc.

Are, O., Kjartan, H., (2017). What is a housing bubble? *Economics Bulletin 37*(2), 806-836.

Belke, Ansgar; Keil, Jonas (2017): Fundamental determinants of real estate prices: A panel study of German regions, Ruhr Economic Papers, No. 731, ISBN 978-3-86788-851-6, RWI - Leibniz-Institut für Wirtschaftsforschung, Essen, https://doi.org/10.4419/86788851

Biau´, G. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research* 13, 1063-1095.

Biau, G., & Fr, G. B. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research*, *13*, 1063-1095.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). Training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, (COLT'92), Pittsburgh, 27-29 July 1992, 144-152.

Botello B., Bigelow S. J., Big Data. The Ultimate Guide to Big Data for Businesses, 2022 Retrieved on July 11th 2022 from https://www.techtarget.com/searchdatamanagement/definition/big-data

Botello, M. L. et al., (2021). Chronic patient remote monitoring through the application of big data and internet of things, *Health Informatics Journal, 27, 3*, July-September 2021.

Breiman, L. (2001). Random Forests. *Machine Learning 2001 45*(1), 5-32.

Brown, Roger. (2019, July 15). *What are Features in Machine Learning and Why it is Important?* https://cogitotech.medium.com/what-are-features-in-machine-learning-and-why-it-is-important-e72f9905b54d

Brown, N. & Sandholm, T. (2019). Superhuman AI for multiplayer poker. *Science 365*, 885–890

Bundhun, M. (2020). A comparison of Fundamental Analysis and Technical analysis in analyzing stock returns on the Stock Exchange of Mauritius (SEM). B.A. Thesis, University of Mauritius https://www.researchgate.net/publication/344198625_A_comparison_of_Fundamental_analysis_and_Technical_analysis_in_analyzing_stock_returns_on_the_Stock_Exchange_of_Mauritius_SEM

Cabrera, J., Wang, T. & Yang, J. (2011) Linear and Nonlinear Predictability of International Securitized Real Estate Returns: A Reality Check, *Journal of Real Estate Research*, 33:4, 565-594.

Cagan, P. (1956). *The Monetary Dynamics of Hyperinflation.* In: Friedman, M. (Ed.), *Studies In the Quantity Theory of Money* (pp. 25-117), University of Chicago Press.

Case, K., Shiller, R. (1988). The behaviour of home buyers in boom and post-boom markets. *New England Economic Review*, 29-46.

Campbell, J. & Shiller, R. (1988). The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors. *The Review of Financial Studies*. 1, 3, 195-228.

Campbell, D.T. (1955). "The Key Informant in Quantitative Research", *The American Journal of Sociology*. Vol. 60, No. 4. Pp. 339 – 342. Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research. *The Qualitative Report*, 8(4), 597-606.

Chan, C. & Xiong, Y. (2007). The Features and Forces that Define, Maintain, and Endanger Beijing Courtyard Housing", *Journal of Architectural and Planning Research*, Vol. 24, No. 1, 42-64 Stable URL: https://www.jstor.org/stable/43030789

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *arXiv:1603.02754*

Chen, C., Chen, X. & Huang, S. (2013) Chinese Guanxi: An Integrative Review and New Directions for Future Research. *Management and Organization Review* 9, 1, 167-207.

Cooper, D., & Schindler, P.S. (2014). *Business Research Methods* (12th ed.). McGraw-Hill Education. India.

Collins, H. (2010). Creative Research: The Theory and Practice of Research for the Creative Industries. Singapore: AVA Publications.

Creswell, J. W. (2014). Research Design: Qualitative, Quantitative and Mixed Methods Approaches (4th ed.). Thousand Oaks, CA: Sage.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). *Ensemble Machine Learning*. *Ensemble Machine Learning*, *February 2014*.

Davis, E. P. & Zhu, H. (2004): "Bank lending and commercial property prices: some cross country evidence", BIS Working Paper No 150.

Development Framework Guidelines, (2018). The Affordable Housing Program. https://www.housingandurban.go.ke/wp-content/uploads/2018/11/Development-Framework-Guidelines-Release-Version.pdf

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179.

Dipasquale, D., & Wheaton, W. C. (1996). *Urban economics and real estate markets.* Prentice Hall.

Ersoz, F., Ersoz, T., & Soydan, M. (2018). Research on Factors Affecting Real Estate Values by Data Mining. *Baltic Journal of Real Estate Economics and Construction Management*, *6*(1), 220-239. https://doi.org/10.2478/BJREECM-2018-0017

François-lavet, V., Henderson, P., Islam, R., Bellemare, M. G., François-lavet, V., Pineau, J., & Bellemare, M. G. (2018). *An Introduction to Deep Reinforcement Learning*. (arXiv:1811.12560v1 [cs. LG]) http://arxiv.org/abs/1811.12560 *Foundations and Trends in Machine Learning*, *II* (3-4), 1-140.

Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2016, 518–522. https://doi.org/10.1109/FSKD.2016.7603227

Gaspareniene, L., Venclauskiene, D., & Remeikiene, R. (2014). *Critical Review of Selected Housing Market Models Concerning the Factors that Make Influence on Housing Price Level Formation in the Countries with Transition Economy*. *Procedia-Social and Behavioral Sciences*, *110*, 419-427.

Habitat for Humanity. (2021). *The Affordable Housing Crisis in Developing Countries.* (n.d.). https://www.habitat.org/emea/about/what-we-do/affordable-housing

Hastie, T., Tinshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Books.

Hayes, A. (2022). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guilford Press.

Herd Publication. (2016). https://www.herd.org.np/uploads/frontend/Publications/PublicationsAttachments1/148540Group%20Discussion0.pdf

Hesse-Biber, S. N., & Johnson, R. B. (2015). *The Oxford Handbook of multimethod and Mixed methods research inquiry*. Oxford University Press.

Hoyle, R. H., Stephenson, M. T., Palmgreen, P., Lorch, E. P., & Donohew, R. L. (2002). Reliability and validity of a brief measure of sensation seeking. *Personality and Individual Differences*, *32*(3), 401-414. https://doi.org/10.1016/S0191-8869(01)00032-0.

Hwang, M., Quigley, J.M. (2010). Housing Price Dynamics in Time and Space: Predictability, Liquidity and Investor Returns. *J Real Estate Finan Econ* 41, 3–23.

Iacobucci, D. & Churchill, G. (2010). *Marketing Research: Methodological Foundations*. Ohio: Northwestern University.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Springer Texts in Statistics.

Kaiser, M. & Cressie, N. (1997). Modeling Poisson variables with positive spatial dependence. *Statistics & Probability Letters*, 35, 4, 423-432.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of Behavioural Research* (4th ed.). Holt, NY: Harcourt College Publishers.

Khalafallah, A. (2008). Neural Network Based Model for Predicting Housing Market Performance. *Tsinghua Science and Technology*, *13*(SUPPL. 1), 325-328.

Kihumba, M. K. (2017). *A Prototype for predicting real estate investment performance in Kenya* [MSc Dissertation, Strathmore University]. Strathmore University SU+ @ Strathmore University Library.

Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. Sage Publications.

Kosgei, M. (2018). "Determinants of Residential House Prices in Nairobi City County, Kenya." PhD Thesis, Moi University.

Kothari, C. R. (2004). *Research methodology*: *Methods and Techniques*. New Age International (P) Limited Publishers.

Li, J., & Liu, H. (2017). Challenges of Feature Selection for Big Data Analytics. *IEEE Intelligent Systems*, *32(2)*, 9–15.

Lim, W. T., Wang, L., Wang, Y., & Chang, Q. (2016). Housing price prediction using neural networks. *Proceedings of the 12th International Conference on Natural Computation.*

Makena, J. (2011). Determinants of Residential Real Estate Prices in Nairobi. Master of Science Dissertation, University of Nairobi.

Malpezzi, S., & Wachter, S. M. (2005). The Role of Speculation in Real Estate Cycles. *Journal of Real Estate Literature*, *13*(2), 143–164. http://www.jstor.org/stable/44103516

Manasa, J., Gupta, R., & Narahari, N. S. (2020). Machine Learning based Predicting House Prices using Regression Techniques. *Proceedings of the 2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020, February*, 624-630.

Mayr, A., Kibkalt, D., Meiners, M., & Lutz B. (2019). *Machine Learning in Production – Potentials, Challenges and Exemplary Applications.* (Towards shifted production value stream patterns through inference of data, models, and technology), 49-54.

Mayr et al., (2020). The LSC Benchmark Dataset: Technical Appendix and Partial Reanalysis. Tech. rep. Institute for Machine Learning, Johannes Kepler University Linz.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics 1943*, *5*(4), 115-133.

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, 7, 983-999.

Mohit, M. A., Al-Khanbashi Raja, A. M. M. (2014) "Residential satisfaction: concept, theories and empirical studies", Planning Malaysia: Urban Planning and Local Governance, 3, 47–66.

Moko, S. K., & Olima, W. H. A. (2014). Determinants of house construction cost in Kenya: a case of Nairobi County. *International Journal of Business and Commerce, 4*(1)2014, 19-36.

Montgomery, D.C, Peck, E., Vining, G. (2021). *Introduction to Linear Regression Analysis*, 6th Edition. Wiley-Blackwell.

Mugenda, O.M. and Mugenda, A.G. (2003) *Research Methods, Quantitative and Qualitative Approaches*. ACT, Nairobi.

Nettleton, D. (2014). *Commercial Data Mining*, Morgan Kaufmann. https://www.sciencedirect.com/science/article/pii/B978012416602809988X

Ngiam, K., & Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *Lancet Oncology*, 20, 5. NA

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Van Le, H., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, https://doi.org/10.1155/2021/4832864

OECD Affordable Housing Database-http://oe.cd/ahd OECD Directorate of Employment, Labour and Social Affairs-Social Policy Division HC1.2. HOUSING COSTS OVER INCOME. (n.d.). http://oe.cd/ahd 19th November, 2021.

Ong, M. C., *Predictive modelling of Singapore Housing Prices — Analyse Singapore's Property Price (Part III)* Retrieved on 19th November, 2021 from https://medium.com/analyticsvidhya/singaporehousingpricesmlpredictionanalysesingaporespperty-price-part-iii-bd9438077423

Onwuegbuzie, A. & Johnson, R. B. (2006). The validity issues in mixed research. *Research in the Schools 2006, 13*(1), 48-63. ResearchGate. https://www.researchgate.net/publication/228340166_The_Validity_Issues_in_Mixed_Research

Otrok, C. & Terrones, M. (2005), "Monetary Policy and House Prices: A Cross-Country Study", A Conference hosted by the Federal Reserve Bank of Atlanta, May. Available from: https://www.researchgate.net/publication/5036324_Monetary_Policy_and_House_Prices_A_Cross-Country_Study#fullTextFileContent

Oust, A. (2018) "The end of Oslo's rent control: Impact on rent level", Economics Bulletin, Volume 38, Issue 1, pages 443-458

Panigrahi, S.S., Mantri, J.K. (2017). A Hybrid of DEA-MLP Model for Stock Market Forecasting. In: Satapathy, S., Bhateja, V., Joshi, A. (eds) Proceedings of the International Conference on Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing, vol 468. Springer, Singapore.

Panigrahi, S. S., & Mantri, J. K. (2016). Epsilon-SVR and decision tree for stock market forecasting. *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, 761-766.

Pant, A. (n.d). Workflow of a Machine Learning project. *Towards Data Science.* https://towardsdatascience.com/workflowofamachinelearningprojectec1dba419b94

Park, B., & Bae, J. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928-2934.

Plakandaras, V., Gogas, P., Gupta, R., & Papadimitriou, T. (2015). US Inflation Dynamics on Long Range Data. *Applied Economics*, http://dx.doi.org/10.1080/00036846.2015.1019039

Ravikumar, A. S. (2018). *Real Estate Price Prediction Using Machine Learning.* [MA Dissertation, University College of Ireland, Dublin] https://norma.ncirl.ie/3096/

Riddel, M. Are Housing Bubbles Contagious? A Case Study of Las Vegas and Los Angeles Home Prices. *Land Economics*. Vol. 87, No. 1 (FEBRUARY 2011), pp. 126-144.

Riggs, D. W., Ansara, G. Y., & Treharne, G. J. (2020). An Evidence-Based Model for Understanding the Mental Health Experiences of Transgender Australians. *Australian Psychologist.* https://www.researchgate.net/publication/270764129_An_Evidence-Based_Model_for_Understanding_the_Mental_Health_Experiences_of_Transgender_Australians/link/5af973dd4585157136f3fcc7/download

Sah, S. (2020). Machine Learning: A Review of Learning Types. *ResearchGate*. https://www.researchgate.net/publication/342890321

Saunders, M., Lewis, P., Thornhill, A., & Bristow, A. (2019). (Eds.). *Research Methods for Business Students*. Pearson.

Taffesse, W. (2006). A Survey on Application of Artificial Intelligence in Real Estate Industry. 3rd International Conference on Artificial Intelligence in Engineering & Technology [iCAiET]. November 22-24, 2006, Kota Kinabalu, Sabah, Malaysia

Tibshirani, R. (1996[2019]). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 1, 267-288. https://www.jstor.org/stable/2346178.

Tong He, K. L. (2016). *XGBoost: An optimized distributed gradient boosting library.*

UDA-Kenya Kwanza (2022). The Kenya Kwanza Plan: The Bottom Up Economic Transformation Agenda 2022-2027. https://uda.ke/downloads/manifesto.pdf

Ungayi, H. M. (2019). Assessment of factors affecting residential real estate prices in Nairobi County (Thesis, Strathmore University). Retrieved from http://suplus.strathmore.edu/handle/11071/662

Waari, D.N. & Kosgei, D. (2018). The Conditional Effect of Experiential Encounter on the Relationship Between Loyalty Programs Benefits and Customer Satisfaction Among Hotel Patrons in Coastal Region, Kenya. *International Journal of Economics, Business and Management Research* Vol. 2, No. 05. 88-102.

Wang, S. Y. (2011). State Misallocation and Housing Prices: Theory and Evidence from China. *American Economic Review*, *101*(5), 2081-2107.

Yamane, T. (1967): Statistics: An Introductory Analysis, 2nd Ed., New York: Harper and Row.

APPENDICES

Appendix 1: Research Participation Information and Consent

PARTICIPATION INFORMATION AND CONSENT FORM

LEVERAGING MACHINE LEARNING TO DETERMINE HOUSING PRICING IN NAIROBI COUNTY

SECTION 1: INFORMATION SHEET

Investigator: JENNIFER WAMBUI NDUATI

Institutional affiliation: Strathmore Business School (SBS)

SECTION 2: INFORMATION SHEET–THE STUDY

2.1: Why is this study being carried out?

This study is being carried out for purposes of undertaking research to determine leveraging machine learning to determine housing pricing in Nairobi County. It is in part fulfilment of a Master in Public Policy Management (MPPM).

2.2: Do I have to take part?

No. Taking part in this study is entirely optional and the decision rests only with you. If you decide to take part, you will be asked to complete a questionnaire to get information on housing prices. If you are not able to answer all the questions successfully the first time, you may be asked to sit through another informational session after which you may be asked to answer the questions a second time. You are free to decline to take part in the study from this study at any time without giving any reasons.

2.3: Who is eligible to take part in this study?

- Home owners in Nairobi County.
- Potential home owners in Nairobi County.
- The State Department for Housing and Urban Development.
- Property Developers who undertake development in Nairobi County.
- Real Estate Agents who sell houses in Nairobi County.

2.4: Who is not eligible to take part in this study?

- Anyone below the age of 18 years.
- Anyone who expresses discomfort in taking part.
- Anyone whose organization excludes him/her from taking part.

2.5: What will taking part in this study involve for me?

You will be approached by Jennifer Wambui Nduati and requested to take part in the study. If you are satisfied that you fully understand the goals behind this study, you will be asked to sign the informed consent form (this form) and then taken through a questionnaire to complete.

2.6: Are there any risks or dangers in taking part in this study?

There are no risks in taking part in this study. All the information you provide will be treated as confidential and will not be used in any way without your express permission.

2.7: Are there any benefits of taking part in this study?

The information will be used to improve information available regarding housing prices in Nairobi County and how machine learning can bolster this.

2.8: What will happen to me if I refuse to take part in this study?

Participation in this study is entirely voluntary. Even if you decide to take part at first but later change your mind, you are free to withdraw at any time without explanation.

2.9: Who will have access to my information during this research?

All research records will be stored in securely locked cabinets. That information may be transcribed into our database but this will be sufficiently encrypted and password protected. Only the people who are closely concerned with this study will have access to your information. All your information will be kept strictly confidential.

2.10: Who can I contact in case I have further questions?

You can contact me, **JENNIFER WAMBUI NDUATI** at SBS, or by e-mail Jennifer.nduati@strathmore.edu or by phone +254 714239537 You can also contact my supervisor, Dr. John Olukuru, at the Strathmore Business School, Nairobi, or by e-mail jolukuru@strathmore.edu or by phone +254 788 356663

If you want to ask someone independent anything about this research please contact:

The Secretary–Strathmore University Institutional Ethics Review Board, P. O. BOX 59857, 00200, Nairobi, email ethicsreview@strathmore.edu Tel number: +254 703 034 375

I, _____, have had the study explained to me. I have understood all that I have read and have had explained to me and had my questions answered satisfactorily. I understand that I can change my mind at any stage.

Please tick the boxes that apply to you;

Participation in the research study

☐ I AGREE to take part in this research

☐ I DON'T AGREE to take part in this research

Storage of information on the completed questionnaire

☐ I AGREE to have my completed questionnaire stored for future data analysis

☐ I DON'T AGREE to have my completed questionnaire stored for future data analysis

Participant's Signature:
_____

Date: _____/_____/_____

DD / MM / YEAR

Participant's Name:
_____
(Please print name)

Time: _____/_____

HR / MN

I, _____(Name of person taking consent) certify that I have followed the SOP for this study and have explained the study information to the study participant named above, and that s/he has understood the nature and the purpose of the study and consents to the participation in the study. S/he has been given opportunity to ask questions which have been answered satisfactorily.

Signature: ……..
_____

Date: _____/_____/_____

DD / MM / YEAR

Name: JENNIFER WAMBUI NDUATI

MPPM- Strathmore University

Time: _____/_____
HR MIN

Appendix II: Research Questionnaire

1. Purpose of the Interview

The purpose of this interview is to collect information from home owners and potential home owners in Nairobi County. More particularly, the questionnaire is aimed at **identifying features that determine housing prices in Nairobi County as well as how leveraging Machine Learning modelling can be used in determining housing prices in Nairobi County.**

The data collected will be for purposes of partial fulfilment of the requirements for the award of a degree of Masters in Public Policy Management, Strathmore Business School, Strathmore University.

2. Confidentiality

Information provided in the Questionnaire by willing participants will be treated as confidential and used strictly for purposes indicated above. The student(s) involved are under oath not to disclose any information to a third party. If you wish to remain anonymous, please indicate in the text box below:

3. Due Date

This is due as soon as the research assistants involved assist you in completing this Questionnaire and furnish you with a duplicate copy of the questionnaire for your own records (should you require it).

4. Respondent

The respondents are expected to be the owners or potential owners of homes in Nairobi County.

5. Queries

Queries or assistance regarding the completion of this questionnaire should be addressed by the interviewers.

For more information contact the following:

Jennifer Nduati – 0714 239537

Jennifer.nduati@strathmore.edu

MPPM-2020

Graduate student

Strathmore University

PS: **This Questionnaire has been designed to complement the use of secondary data, particularly to identify features that determine housing prices in Nairobi County as well as how leveraging Machine Learning modelling can be used in determining housing prices in Nairobi County. This will aid in making the outputs of the study more realistic and practical.**

Please answer all questions.

SECTION 1: BACKGROUND INFORMATION

1. Please tick your age bracket

a. 21-30 ☐

b. 31-40 ☐

c. 41-50 ☐

d. 51-60 ☐

2. Please indicate whether you are married

Yes ☐

No ☐

☐

3. **Number of children**

4. **Source of income**

Employed

Self- Employed

5. **Income level**

1,000-14,999

15,000-49,999

50,000-99,000

100,000 and above

6. **Please tick your level of education**

a. Primary

b. Secondary

c. University

d. Others..... please specify.....

SECTION 2: INFORMATION ON HOUSING PRICES

PART A. FEATURES DETERMINING HOUSING PRICES

=1 Strongly disagree (SD) 2=Disagree (D) 3=Neutral (N) 4=Agree (A) 5= Strongly Agree (SA)

		SD	D	N	A	SA
NO.	**STATEMENT**	**1**	**2**	**3**	**4**	**5**
1.	To what extent do you agree that the following features are taken into account while determining house prices? i. Age of the House ii. Size of the House (Unit Plinth) iii. Number of Bedrooms iv. Number of Bathrooms v. Presence of a fire place vi. Tarmacked Road (within 0.5 kilometres) vii. Presence of a school (within 5 kilometres) viii. Piped Water ix. Electricity Connection/Generator x. Internet Connectivity					
2.	To what extent do you agree that the features are the key determinants of housing prices?					
3.	To what extent do you agree that the developers are aware of these features?					
4.	To what extent do housing features matter to you?					

Your time and participation in the research has been highly appreciated.

-END-

Appendix III: Key Informant Interview Guide

1. Purpose of the Survey

The purpose of this Key Informant Guide is to collect information from the Directors and Assistant Directors at the State Department. These are the key persons involved in delivering houses to Kenyans as well as the implementation of the Kenya Kwanza Plan aimed at providing 200,000 homes annually. This Key Informant Guide is aimed at **identifying features that determine housing prices in Nairobi County as well as how leveraging Machine Learning modelling can be used in determining house prices in Nairobi County.** This will aid in making the outputs of the study more realistic and practical.

The data collected will be for purposes of partial fulfilment of the requirements for the award of a degree of Masters in Public Policy Management, Strathmore Business School, Strathmore University. The survey is aimed at identifying the **features that determine housing prices in Nairobi County as well as how leveraging Machine Learning modelling can be used in determining housing prices in Nairobi County**

2. Confidentiality

Information provided in the questionnaire by the participants will be treated as confidential and used strictly for purposes indicated above. The student involved is under oath not to disclose any information to a third party. If you wish to remain anonymous, please indicate in the text box below:

3. Duration and Modalities of the Interview

This Key Informant Guide will take approximately 30-45 minutes. Upon completion of the survey, I shall allow you ten minutes to raise any questions which we shall respond to.

4. Queries

Queries or assistance regarding the completion of this questionnaire should be addressed by the interviewer whose contact details are below.

Jennifer Nduati – 0714 239537

Jennifer.nduati@strathmore.edu

MPPM-2020

Graduate student

Strathmore University

PS: This questionnaire has been designed to complement the use of secondary data, to identify features that determine housing prices in Nairobi County as well as how to leverage Machine Learning modelling in determining housing prices in Nairobi County.

Please answer all questions.

Key Informant Interview Guide

1. **What are the current prices of the houses being built in Nairobi County?**

……

……

2. **How are these prices arrived at?**

……

……

3. **Were any features taken into account while determining the house prices?**

……

……

4. **If yes, what features were taken into account?**

……

……

5. **Is the Kenyan public aware of how the prices for the houses are arrived at?**

……

……

6. **Do you think that the prices of houses are well determined?**

……

……

7. **In your opinion, do you agree that the method/s applied in determining house prices are sufficient and accurate?**

……

……

8. **Are any Machine Learning tools or models applied towards determining house prices?**

……

……

a. If yes, which ones?

……

……

b. If no, why not?

……

……

9. In your opinion, would Machine Learning provide more accurate house prices?

……

……

10. Would the government allocate the adequate resources for use of Machine Learning to determine house prices?

……

……

Your time and participation in the research has been highly appreciated.

END

Appendix IV: Institutional Ethics Permit

Strathmore UNIVERSITY

16th February 2023

Mrs Nduati Jennifer Wambui,
jennifer.nduati@strathmore.edu

Dear Mrs Nduati,

RE: Leveraging Machine Learning for Housing Price Prediction in NairobiCounty

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU- master's** researchproposal. Your application reference number is **SU-ISERC1550/23.** The approval period is from **16th February 2023 to 15th February 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, and MTA) will beused
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 48 hours of notification
iv. Any changes, anticipated or otherwise, that may increase the risks or affect the safety or welfareof study participants and others or affect the integrity of the research must be reported to SU-
ISERC within 48 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 NationalCommission for Science, Technology, and Innovation (NACOSTI) https://research- portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for : Dr Ben Ngoye, Secretary; SU-ISER

16-Feb-2023

Cc: Mr Ambrose Rachier,
Chairperson; SU-ISERC

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

Appendix V: NACOSTI Research Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE TECHNOLOGY & INNOVATION

Ref No: 338552

Date of Issue: 02/March/2023

RESEARCH LICENSE

This is to Certify that Ms. Jennifer Wambui Nduati of Strathmore University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: LEVERAGING MACHINE LEARNING TO DETERMINE HOUSE PRICING IN NAIROBI COUNTY for the period ending : 02/March/2024.

License No: **NACOSTI/P/23/23949**

338552

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

THE SCIENCE, TECHNOLOGY AND INNOVATION ACT, 2013 (Rev. 2014)

Legal Notice No. 108: The Science, Technology and Innovation (Research Licensing) Regulations, 2014

The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was the established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of international treaties of which Kenya is a signatory to
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
    i. Endanger national security
    ii. Adversely affect the lives of Kenyans
    iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
    iv. Result in exploitation of intellectual property rights of communities in Kenya
    v. Adversely affect the environment
    vi. Adversely affect the rights of communities
    vii. Endanger public safety and national cohesion
    viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. The license any rights thereunder are non-transferable
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or

disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.

13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and Innovation (NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 0713788787, 0735404245
E-Mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke