

FERAS

Foresight-Enhanced Real estate Appraisal with Spatial intelligence

A Case Study on Predicting Price per Square Meter in the
Saudi Arabian Real Estate Market



Submitted in fulfillment of the requirements for the degree
Master of Science in Artificial Intelligence

AT

Collage of Informatics

Midocean University

UAE

Prepared By

Fahad AbuBaker Bahashwan

Student ID No. 243001702

Supervisor By

Dr. Ashraf Alsaeed

Collage of Informatics, Midocean University, Comoros

September 2025

Abstract

This thesis presents a comprehensive machine learning framework for predicting the price per square meter ($\text{ﷲ}/\text{m}^2$) of real estate properties in Saudi Arabia using advanced spatial features and algorithms. The study addresses the critical demand for accurate and automated property valuation in the rapidly expanding Saudi real estate market. Using a data set of 6,005 property transactions in major cities - including Riyadh, Jeddah, and the eastern province - multiple machine learning models were developed and evaluated, including linear regression, Ridge, Lasso, random forest, gradient boosting, XGBoost, K-nearest neighbors and decision trees.

The methodology incorporates extensive feature engineering, including geographic coordinates, distances to city centers calculated using the Haversine formula, proximity to schools and educational facilities, detailed road network analysis (counts and distances to motorways, trunk, primary, secondary, tertiary, and residential roads), as well as property type categorization. The study applies a modern hedonic pricing framework powered by machine learning algorithms to capture the implicit value of spatial and property characteristics in real estate price prediction.

The experimental results reveal that the XGBoost model achieved the highest predictive performance, with a R^2 score of 0.9525 on the test set and a Root Mean Squared Error (RMSE) of 367.25 ($\text{ﷲ}/\text{m}^2$). Furthermore, the model achieved a mean absolute percentage error (MAPE) of 9.93% and a mean absolute error (MAE) of 237.37 ($\text{ﷲ}/\text{m}^2$), reflecting robust predictive precision.

In general, this research contributes valuable information for real estate professionals, urban planners and policymakers in Saudi Arabia by demonstrating the effectiveness of machine learning approaches in property valuation while ensuring transparency through explainable AI techniques.

Keywords: Artificial Intelligence, Property Valuation, Machine Learning, XGBoost, Real Estate, Saudi Arabia, Algorithmic Fairness, Geographic Information Systems, Automated Valuation Models, Ensemble Methods

List of Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
GIS	Geographic Information System
OSM	OpenStreetMap
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
R^2	Coefficient of Determination
IQR	Interquartile Range
Latitude	Position north/south of the Equator
Longitude	Position east/west of the Prime Meridian
Haversine sphere	Great-circle distance calculation method between two points on a sphere
SHAP	Shapley Additive Explanations
XGBoost	Extreme Gradient Boosting
KNN	K-Nearest Neighbors
RFR	Random Forest Regressor
GBR	Gradient Boosting Regressor
DTR	Decision Tree Regressor
LR	Linear Regression
Ridge	Ridge Regression

Lasso	Least Absolute Shrinkage and Selection Operator
sklearn	Scikit-learn (Python ML library)
pandas	Python data manipulation and analysis library
NumPy	Python numerical computation library
Matplotlib	Python plotting library
Seaborn	Python statistical visualization library
geopy	Python library for geocoding and distance calculations
OneHotEncoder	Scikit-learn categorical feature encoder
Train/Test Split	Process of dividing data into training and testing subsets
Boxplot	Statistical data visualization showing quartiles and outliers
Heatmap	Visualization showing correlation or intensity values in a grid
Feature Engineering	Process of creating new variables to improve model performance
Outlier	Data point significantly different from others in the dataset
Visualization	Graphical representation of data
Residuals	Difference between actual and predicted values in regression
Error %	Relative percentage difference between actual and predicted values
Motorway	Major high-speed road, no cross traffic
Trunk	Major route connecting cities
Primary	Main road within or between towns
Secondary	Road linking smaller towns or districts
Tertiary	Local connector between smaller roads
Residential	Street serving housing areas
Hedonic Model	Method estimating value based on the contribution of individual features

Acknowledgments

I express my sincere gratitude to all those who have contributed to the completion of this thesis and supported me throughout my master's degree journey.

First, I extend my deepest appreciation to my advisor, Dr. Ashraf Alsaeed, for his invaluable guidance, expertise, and unwavering support throughout this research project. His insightful feedback, constructive criticism, and encouragement have been instrumental in shaping this work and my development as a researcher. His commitment to academic excellence and ethical research practices has been truly inspiring.

I am grateful to the faculty and staff of the Collage of Informatics at Midocean University for providing an exceptional academic environment and the resources necessary to conduct this research. Special thanks to the members of the thesis committee for their time, expertise, and valuable feedback that helped improve the quality of this work.

This research represents not only my individual effort, but also the collective contribution of all those who have supported and guided me along the way. Any errors or omissions are my own responsibility.

Fahad AbuBaker Bahashwan
September 2025

Contents

Abstract	i
List of Abbreviations	ii
Acknowledgments	iv
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	2
1.3 Research Objectives	3
1.4 Research Contributions	4
1.5 Thesis Organization	5
2 Literature Review	7
3 Methodology	11
3.1 Research Design and Approach	11
3.2 Dataset Description and Characteristics	12
3.3 Data Preprocessing	16
3.4 Feature Engineering and Geographic Integration	16
3.5 Machine Learning Algorithms and Implementation	16
3.5.1 XGBoost (Extreme Gradient Boosting)	16
3.5.2 Random Forest	17
3.5.3 Gradient Boosting	17
3.5.4 Linear Regression Methods	17
3.5.5 K-Nearest Neighbors	17
3.5.6 Decision Tree Regressor	17
3.6 Experimental Design and Validation	17
3.7 Fairness Evaluation Framework	17
4 Experimental Results and Analysis	18
4.1 Dataset Characteristics and Exploratory Analysis	18

CONTENTS

- 4.2 Model Performance Comparison 18
- 4.3 XGBoost Model Analysis 18
 - 4.3.1 Hyperparameter Optimization 18
 - 4.3.2 Feature Importance Analysis 18
- 4.4 Geographic Feature Impact Analysis 18
- 4.5 Fairness and Bias Evaluation 19
 - 4.5.1 Geographic Fairness Assessment 19
 - 4.5.2 Property Type Fairness Assessment 19
 - 4.5.3 Price Range Equity Analysis 19
- 4.6 Model Deployment Considerations 19

- 5 Discussion 20**
 - 5.1 Interpretation of Results 20
 - 5.2 Implications for Property Valuation Practice 20
 - 5.3 Geographic Feature Integration Insights 20
 - 5.4 Fairness and Bias Implications 20
 - 5.5 Limitations and Challenges 20
 - 5.6 Regulatory and Compliance Considerations 20
 - 5.7 Future Research Directions 21

- 6 Conclusion and Future Work 22**
 - 6.1 Summary of Findings 22
 - 6.2 Research Contributions 22
 - 6.2.1 Methodological Contributions 22
 - 6.2.2 Technical Contributions 22
 - 6.2.3 Practical Contributions 22
 - 6.3 Implications for Industry and Policy 22
 - 6.3.1 Industry Implications 23
 - 6.3.2 Policy Implications 23
 - 6.4 Limitations and Future Research Directions 23
 - 6.4.1 Data and Methodology Limitations 23
 - 6.4.2 Technical Research Directions 23
 - 6.4.3 Application and Deployment Research 23
 - 6.5 Final Remarks 23

List of Figures

- 3.1 AVM framework for predicting real estate price per square meter. 12
- 3.2 Monthly transactions volume (July 2024 – July 2025). 13
- 3.3 Spatial distribution of transactions across 16 cities. 14
- 3.4 Price (ﷲ/m²) distributions in Riyadh. 15
- 3.5 Price (ﷲ/m²) distributions in Makkah and Jeddah. 15

List of Tables

- 3.1 Transactions by city (July 2024 – July 2025). 13
- 3.2 Transactions by property type. 14
- 3.3 Data dictionary for features. 16

Chapter 1

Introduction

1.1 Background and Motivation

The real estate industry is on the threshold of a technological revolution, driven by the convergence of artificial intelligence, big data analytics, geospatial technologies such as GIS, and the growing demand for efficient, transparent and equitable property valuation systems. Traditional property appraisal methods, although long-established and professionally established, face growing challenges in meeting the requirements for speed, consistency, and scalability in modern markets. These approaches often rely heavily on manual inspections, subjective judgment, and limited comparable data, making them less adaptable to rapidly changing urban environments and large-scale property transactions. In contrast, the emergence of AI-driven valuation systems offers unprecedented opportunities to overcome these limitations by leveraging big data, spatial features, and machine learning algorithms to provide faster, more consistent, and scalable valuations. However, adopting artificial intelligence (AI) in property valuation also raises important issues of fairness, transparency, and regulatory compliance. This creates a dual challenge: to use technological innovation effectively while ensuring that it is applied ethically and equitably.

Property valuation represents one of the most critical functions in real estate markets, shaping mortgage lending decisions, investment strategies, tax assessments, and overall market transparency. The accuracy and reliability of valuations exert far-reaching effects on financial stability, economic development, and social equity. Traditional appraisal methods, typically conducted by licensed professionals through comparative market analysis and standardized procedures, provide essential experience and professional judgment. However, they remain constrained by human limitations in processing vast and heterogeneous datasets, ensuring consistency across different appraisers, and scaling to meet the demands of increasingly dynamic and complex real estate markets. These challenges are particularly pronounced in rapidly developing regions

such as Saudi Arabia, where rapid urban expansion and national initiatives such as Vision 2030 and large-scale housing and infrastructure projects have increased the demand for valuation systems that are scalable, data-driven and able to integrate spatial features for more accurate assessments.

In this context, the Saudi Arabian real estate market presents a particularly compelling case for investigating AI-driven property valuation systems. As Saudi Arabia advances its Vision 2030 economic diversification strategy, the sector has undergone significant growth and transformation, introducing both new challenges and opportunities for valuation practices. The combination of rapid urban expansion, growing foreign investment and ongoing regulatory reforms has increased the need for more efficient, consistent, and data-driven valuation systems capable of supporting Saudi Arabia's long-term economic development objectives.

Saudi Arabia's property valuation system has the same problems seen worldwide, but also has special issues and unique characteristics specific to the region. Historically, Saudi Arabia has relied on conventional appraisal methods that, although grounded in professional expertise, lack the capacity to handle the volume and complexity of modern transactions. The increasing sophistication of investments, the emergence of new property types, and the accelerating pace of market activity highlight the limitations of traditional methods in delivering timely and consistent valuations, raising concerns about their capacity to sustain an efficiently functioning market.

1.2 Problem Statement

The current property valuation landscape in Saudi Arabia and comparable emerging markets faces multiple interconnected challenges that undermine efficiency, consistency, and scalability. Traditional appraisal methods, while grounded in professional expertise and market knowledge, are hindered by time-intensive processes that often take days or weeks to complete, reducing their relevance in fast-moving market conditions. Moreover, the subjective nature of human judgment—though valuable in highly complex or non-standard cases—frequently introduces inconsistency and potential bias in routine valuations, underscoring the limitations of traditional approaches in meeting the demands of modern real estate markets.

In addition to these limitations, the growing volume of real estate transactions in Saudi Arabia driven by Vision 2030 initiatives and broader economic diversification efforts has put additional pressure on traditional valuation systems. The limited number of qualified appraisers relative to market demand creates bottlenecks in transaction processing, slowing economic activity, and reducing overall market efficiency. This capacity constraint is particularly acute in rapidly developing regions, where the pace of new construction and investment activity far exceeds the availability of local valuation

expertise.

Moreover, traditional valuation methods struggle to capture the complex relationships between property characteristics and market values in dynamic real estate markets. The growing availability of diverse data sources, including geographic information systems, demographic data, and market analytics, creates opportunities for more sophisticated approaches capable of integrating multiple factors simultaneously. However, the manual processing constraints of conventional methods prevent full utilization of these rich data resources, further limiting their effectiveness in modern valuation practice.

The issue of consistency and potential bias in property valuations constitutes another significant challenge. Appraisers may reach divergent conclusions about the same property, influenced by individual experience, judgment, and interpretations of market conditions. This variability generates uncertainty for market participants and can unintentionally introduce systematic biases, leading to inequitable outcomes across property types, locations, or market segments.

Together, these challenges highlight the urgent need for innovative, data-driven valuation approaches capable of providing faster, more consistent, and scalable property assessments.

1.3 Research Objectives

This research aims to address the identified challenges by investigating AI-driven property valuation systems within the context of the Saudi Arabian real estate market. The primary objective is to develop and rigorously evaluate machine learning approaches capable of delivering accurate, efficient, and fair valuations, while also aligning with the practical requirements necessary for real-world implementation.

The first research objective focuses on establishing the optimal machine learning algorithm for property price prediction in the Saudi Arabian context. Through systematic comparison of multiple algorithms including ensemble methods, linear regression approaches, and tree-based models, our aim is to identify the approach that maximizes accuracy while ensuring explainability and practical applicability. This objective requires a comprehensive evaluation of algorithm performance in multiple metrics, including predictive accuracy, computational efficiency, and generalization capabilities.

The second objective focuses on the integration of geographic and spatial features into property valuation models. Since location is a fundamental determinant of property value, this research seeks to develop methodologies to incorporate spatial relationships, accessibility measures, and geographic characteristics into machine learning frameworks. Specifically, this involves building distance-based features, proximity indicators, and regional attributes that can improve predictive accuracy while ensuring

transparency and practical interpretability.

The third objective focuses on the comprehensive evaluation of fairness and bias in AI-driven valuation systems. Recognizing the critical importance of equitable treatment in real estate, this research seeks to design and apply rigorous methodologies for detecting and measuring bias between property types, geographic regions, and market segments. To achieve this, fairness metrics will be established, bias detection procedures implemented, and validation will be performed to ensure that automated valuation systems deliver equitable results across all relevant categories.

The fourth objective involves developing practical implementation frameworks that address the real-world requirements to deploy AI-driven valuation systems. This includes evaluating computational efficiency, scalability requirements, integration challenges, and operational considerations that influence successful adoption. The framework must address technical infrastructure needs and data quality requirements.

Collectively, these objectives establish a comprehensive framework for developing AI-driven valuation systems that are accurate, transparent, fair, and practically deployable within the Saudi real estate market.

1.4 Research Contributions

This research makes several significant contributions to both academic knowledge and practical practice in the field of AI-driven property valuation. Specifically, it advances theoretical understanding, introduces methodological innovations, provides practical implementation frameworks, and delivers contextual insights within the Saudi Arabia real estate market.

Methodological Contributions: This research offers a comprehensive evaluation of machine learning algorithms for property valuation in the Saudi Arabian real estate market, addressing an underexplored area of the literature and providing insights of relevance for emerging markets in the Middle East. By systematically comparing eight algorithms under consistent evaluation criteria, the study establishes methodological benchmarks for future research and offers practical guidance for algorithm selection in comparable market contexts.

Technical Contributions: This research advances the literature on spatial analysis in real estate valuation by developing methodologies to integrate geographic characteristics into machine learning models. Incorporating distance-based measures, proximity indicators, and accessibility metrics demonstrates practical approaches to capture location effects, establishing a foundation for future research on spatial relationships in property pricing.

Fairness and Ethics Contributions: This research strengthens the emerging field of algorithmic fairness in real estate valuation by performing a comprehensive fairness

analysis that demonstrates how well-designed AI systems can achieve equitable results along with improved efficiency and accuracy. The multidimensional framework for bias detection and fairness assessment establishes transferable methodologies that can be applied to other AI applications in regulated domains, reinforcing the importance of ethical considerations in technology adoption.

Practical Contributions: This research delivers practical frameworks and methodologies for implementing AI-driven property valuation systems in real-world contexts, narrowing the gap between academic exploration and industry application. By systematically evaluating deployment factors, including computational efficiency, scalability, and integration requirements, the study provides actionable guidance for organizations and institutions considering the adoption of automated valuation systems.

1.5 Thesis Organization

This thesis is organized into six main chapters, each addressing specific aspects of AI-driven property valuation research.

Chapter 1: Introduction provides the introduction, presenting the background, motivation, problem statement, research objectives, contributions, and the overall structure of the thesis.

Chapter 2: Literature Review provides a comprehensive survey of existing research on property valuation, automated valuation models, machine learning applications in real estate, and algorithmic fairness. It establishes the theoretical foundation for this study and identifies the key gaps in current knowledge that the research seeks to address.

Chapter 3: Methodology presents the research design, detailing the dataset characteristics, the pre-processing procedures, the feature engineering approaches, the machine learning algorithms evaluated, and the experimental design. The chapter ensures methodological transparency and provides sufficient detail to enable replication and validation of the results.

Chapter 4: Experimental Results and Analysis reports the evaluation outcomes, including algorithm performance comparisons, feature importance analysis, fairness evaluations, and model validation procedures. The chapter offers a detailed examination of the experimental findings and discusses their implications for property valuation practices.

Chapter 5: Discussion interprets the research findings in relation to the existing literature, examines their implications for property valuation practice, acknowledges limitations and challenges, and offers recommendations to practitioners and policy makers.

Chapter 6: Conclusion and Future Work summarizes the key findings, outlines research contributions, discusses future research directions, and provides final remarks

Chapter 1. Introduction

on the potential of AI-driven property valuation systems.

The thesis also includes comprehensive appendices that provide detailed documentation of datasets, performance metrics, geographic analysis results, feature importance rankings, and fairness evaluation procedures.

Chapter 2

Literature Review

The story of property valuation begins with a fundamental challenge that has persisted across centuries: how to accurately and fairly determine the value of real estate in an ever-changing market. This narrative traces the evolution from traditional appraisal methods to sophisticated artificial intelligence systems, revealing both the promise and the risks of technological transformation in one of the most critical economic functions of society. For decades, property valuation was based on three established approaches that formed the foundation of professional practice. Sales comparison, cost and income capitalization approaches provided the framework through which licensed appraisers navigated the complex task of price determination. These approaches became the global standard of valuation practice and remain influential today, although increasingly challenged by the growing scale and complexity of modern real estate markets [1].

These methods, grounded in economic theory and professional judgment, operate on well-established principles such as substitution, which posits that rational buyers will not pay more for a property than the cost of acquiring a comparable substitute [2]. The comparative market analysis (CMA) methodology emerged as a practical application of these principles, carried out by real estate professionals with intimate knowledge of local conditions [3]. Although grounded in sound reasoning, such approaches depend heavily on localized expertise and limited *comparables*, which constrains scalability and consistency.

Although these traditional approaches served the industry for generations, their limitations became increasingly apparent in modern markets shaped by rapid urbanization and increasingly complex data. The subjectivity inherent in human judgment, while valuable in complex or unique cases, frequently introduced inconsistency and potential bias in routine valuations, creating uncertainty for market participants and, in some cases, inequitable results across property types, locations, and demographic segments [4, 5]. Moreover, traditional methods struggled to capture complex relationships between property characteristics and market values in dynamic environments, particularly as diverse data sources such as geographic information systems (GIS), demographic

indicators, and market analytics became more widely available [5, 6].

A more systematic foundation emerged with Sherwin Rosen’s hedonic pricing model, which conceptualized a property as a bundle of attributes, with buyers valuing individual characteristics such as size, location, and amenities [7]. Observed market prices can therefore be interpreted as the aggregate result of marginal contributions of specific attributes [7, 8]. In its simplest form:

$$P_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ij} + \epsilon_i,$$

where P_i is the price of the property i , X_{ij} is the attribute j of property i , β_j is the implicit marginal contribution of attribute j , and ϵ_i is an error term. This formulation became the basis for hedonic regression and ultimately for modern machine learning applications in property valuation.

The integration of GIS with hedonic models marked the next chapter, as researchers recognized that location factors could be systematically quantified and incorporated into the valuation [9]. This development addressed a fundamental real-estate principle: location is a primary determinant of value. The ability to capture spatial relationships, accessibility measures, and neighborhood characteristics through GIS enabled more sophisticated analytical approaches that would emerge in subsequent decades.

As computational power increased and data became more abundant, the limitations of traditional approaches catalyzed technological innovation. Automated Valuation Models (AVMs)—computer-based systems that apply statistical, econometric, or machine learning techniques to property and market data—ushered in a paradigmatic shift toward systematic, data-driven evaluation, producing valuations with greater speed, scalability, and consistency than traditional appraisal methods [10]. Research has increasingly demonstrated the ability of AVMs to improve value assessment, support decision-making, and provide market information on development, investment, land administration, and taxation [11]. Comparative investigations across markets further showed that ensemble learners (e.g. random forests and gradient boosting) often achieve higher R^2 and lower MAE than linear baselines, reflecting their ability to model nonlinear interactions among property and location attributes [12].

Among these learners, gradient-boosted trees (notably XGBoost) gained prominence due to strong regularization, scalability, and performance in heterogeneous tabular features [13]. However, accuracy alone is insufficient in high-stakes settings like valuation; interpretability is essential. SHAP (SHapley Additive exPlanations) offers a unified game-theoretic framework that attributes the positive or negative contribution of each feature to an individual prediction - its local attributions sum to the model output - making the AVMs of boosted trees more transparent and auditable [14]. Evidence

from Ukraine similarly illustrates the performance advantages of boosted ensembles in real-estate pricing tasks (e.g., gradient boosting achieved $R^2 = 0.724$) [15].

Parallel to these algorithmic advances, the centrality of spatial features has been repeatedly confirmed. Studies show that location-based factors account for a substantial share of price variation. For example, research in Seoul demonstrated that proximity to subway stations and schools increases apartment values, confirming accessibility and amenities as decisive determinants [16]. Similar patterns appear in European markets, where neighborhood quality and transport connectivity explain a sizable fraction of the variation in machine learning models [11]. In Saudi Arabia, the incorporation of geospatial characteristics, such as distance to city centers and road networks, significantly improves predictive precision [17]. These findings validate the principle that “location is value” and support systematic GIS integration in AVMs.

Beyond simple distance features, explicit spatial structure can be modeled directly. A two-stage pipeline, the clustering to define submarkets followed by kriging to capture spatial dependence, has shown a strong performance for *mass appraisal* in Medellín, highlighting the value of the geostatistical structure alongside the learned models [18]. Large-scale empirical work also reveals a strong spatial autocorrelation in housing prices and heterogeneous determinants (for example, land cost) between regions, using spatial regression and Geographical Detector (GeoDetector) to identify varying effects [19]. Together, these results motivate standardized and reproducible pipelines for spatial feature engineering (e.g. distances, network travel times, neighborhood indices) in operational AVMs [20].

Recent studies also tackle valuation when there are few recorded sales. Geography-guided self-supervised methods (e.g., DoRA) first learn from unlabeled data and then improve accuracy even with small labeled samples—useful for areas or property types with limited transactions [21].

Multimodal learning extends the feature space beyond the structured attributes. Self-supervised vision transformers can fuse interior, exterior, and streetscape imagery with tabular features to capture quality cues that are difficult to encode numerically [22], while street-view visual semantics significantly improve land or housing price estimation [23]. Complementing these findings, adding AI-derived visual scores from property photos to a hedonic baseline improves accuracy and supports price transparency objectives [24].

The GCC context provides practical momentum. Across the region, rapid economic diversification and aggressive digitization have created fertile conditions for AI-driven valuation. In the UAE, the Dubai Land Department’s “Instant Sale” operationalizes AI-based smart valuation to complete key steps in seconds, demonstrating that end-to-end digital valuation is practical at scale [25]. Regionally, industry research notes an acceleration of AI adoption in real estate and PropTech, with institutional users pri-

oritizing speed, auditability, and data integration [26]. For Saudi Arabia, the national data-and-AI strategy under Vision 2030 positions AI as an explicit enabler in government and industry, while sectoral digital platforms and open transaction data provide the practical foundation for modernizing **property valuation**—including REGA’s Digital Real Estate Indicators and the Ministry of Justice’s real estate market data (with price per square meter) - and professional governance by TAQEEM sets standards for the valuation profession [27, 28, 29, 30]. Empirical work with Saudi transactions already shows strong machine learning performance relative to traditional baselines, indicating readiness for production-grade systems [17]. Local PropTech platforms (e.g., Wasalt) report large-scale usage and AI-powered capabilities (instant valuation, personalized descriptions, ROI prediction), reflecting market appetite and feasibility [31, 32].

As adoption scales, fairness becomes central. Recent public-sector research documents that AVMs can exhibit larger errors in majority black communities - even when protected attributes are not explicit - implying that overall predictive accuracy does not guarantee equitable outcomes [33]. Regulators have responded: In 2024, six federal agencies in the USA issued a final rule mandating AVM quality control standards, including testing, documentation, and controls to guard against bias, formalizing governance expectations for automated home valuations [34].

Putting it all together, the literature indicates that effective AI-driven valuation combines (i) strong tabular learners (e.g., gradient boosting), (ii) standardized spatial feature engineering with GIS, (iii) optional multimodal inputs (e.g., imagery), (iv) interpretable outputs (e.g., SHAP), (v) spatio-temporal modeling with uncertainty estimates, and (vi) fairness evaluation aligned with emerging regulations. What is still missing, especially for Saudi Arabia, is a unified and standardized spatial pipeline with localized fairness and uncertainty diagnostics integrated into one practical framework. This thesis addresses this gap by proposing *FERAS*, derived from "firāsa" meaning foresight, a framework that leverages spatial features and explainability while incorporating fairness and uncertainty assessments tailored to the Saudi real estate market.

Chapter 3

Methodology

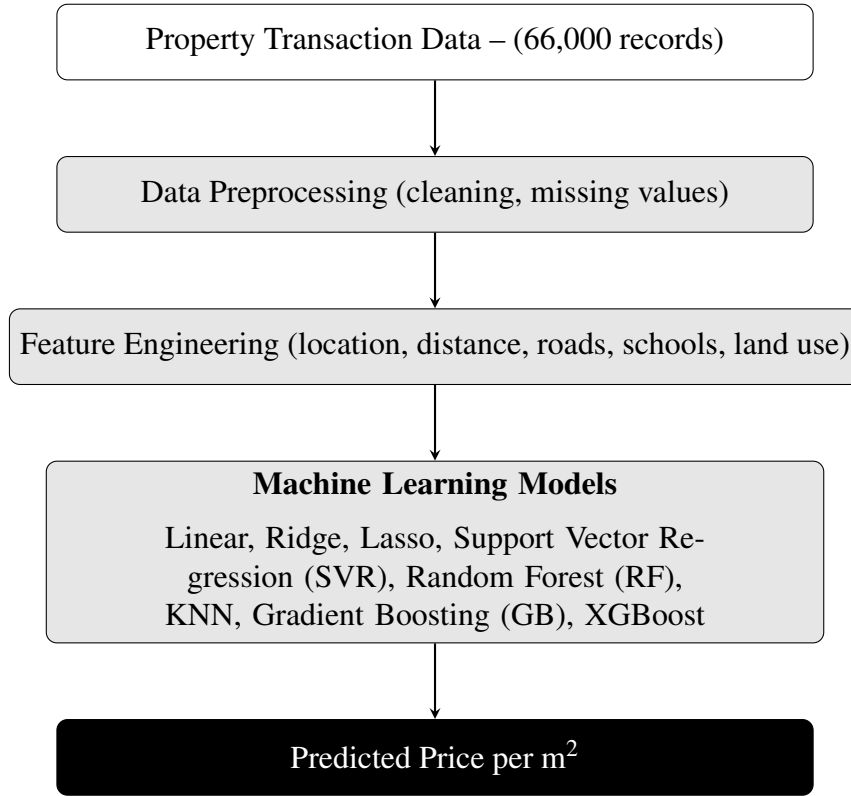
3.1 Research Design and Approach

This study uses a quantitative experimental design to evaluate AI-driven automated valuation models (AVMs) to predict the price per square meter ($\text{﷼}/\text{m}^2$) in the Saudi real estate market. The pipeline covers data acquisition, preprocessing, spatial and temporal feature engineering, model training with leakage-aware validation, and fairness and robustness assessment. A single reproducible feature pipeline is fixed across models, with identical data splits, tuning budgets, and random seeds to ensure a controlled comparison.

The overall research workflow is summarized in Figure 3.1, which illustrates the conceptual AVM pipeline applied in this study. It begins with the collection of property transaction data, followed by preprocessing and feature engineering. Spatially enriched features are then passed into multiple machine learning models for training and evaluation, leading to final predictions of price per square meter.

We adopt a comparative framework for eight algorithms representing major families. OLS, Ridge, Lasso, Support Vector Regression, Random Forest, Gradient Boosting, XGBoost, and LightGBM (Tree Boosting per 13). All models use the same inputs and evaluation procedures. Validation follows (i) a scheme *time-aware* (train in earlier periods, test later) and (ii) a scheme *spatial-block* (hold out entire districts/tiles) to assess geographic generalization under spatial autocorrelation [35]. We report R^2 , RMSE, MAE and MAPE in holdout sets and use paired error tests to assess significance.

Figure 3.1: AVM framework for predicting real estate price per square meter.



Because deployment in regulated settings requires not only accuracy but also equity and transparency, we evaluate both predictive performance and fairness. We compute SHAP values for explanations per prediction and aggregate them for global importance [14]. For uncertainty, we produce 90% prediction intervals via distribution-free conformal prediction and empirical audit coverage [36]. For fairness in regression, we assess error parity, signed bias, coverage parity, and calibration across observable groups (districts, property types, price bands), following survey guidance on bias and mitigation in machine learning. An ablation (no spatial $\rightarrow$ +proximity $\rightarrow$ +neighborhood/lag) quantifies the incremental value of standardized GIS-based features.

3.2 Dataset Description and Characteristics

This study uses a large-scale multicity dataset of Saudi Arabian property transactions suitable for machine learning analysis in a regional real estate context. The data set contains 66,000 verified transactions drawn from multiple sources in 16 major Saudi cities, providing extensive geographic diversity for modeling and evaluation¹. The prediction target is the price per square meter ($\text{ﷲ}/\text{m}^2$), calculated as $y = \frac{\text{sale_price}}{\text{area}}$ after standardizing units and filtering implausible or missing areas.

¹Where applicable, data sources include official transaction feeds [28, 29].

Temporal coverage. The dataset spans July 2024 through July 2024, capturing multiple market phases and policy environments. This broad period ensures that models are trained and tested under heterogeneous conditions rather than a single economic regime. Monthly transaction volume over the study window is shown in Figure 3.2.

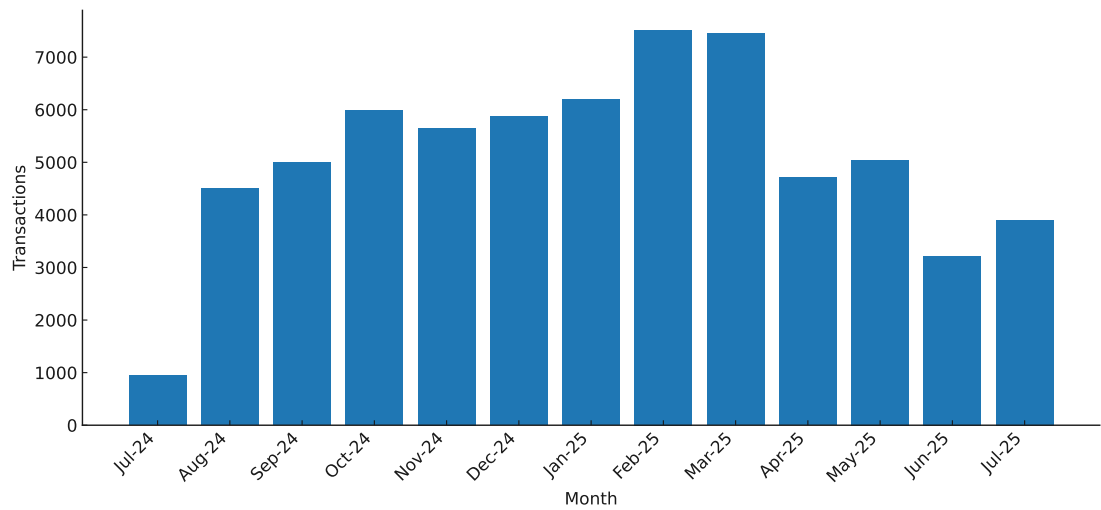


Figure 3.2: Monthly transactions volume (July 2024 – July 2025).

Geographic coverage. Transactions are distributed across key urban centers: Riyadh (61.1%), Jeddah (9.7%), Makkah (6.3%), Dammam (6%), Buridah (4.3%), Khobar (3.3%), with the remaining 9.3% in eleven additional cities. This distribution reflects major population and economic hubs while preserving the cross-city variability needed for spatial generalization. The counts and shares at the city level are reported in Table 3.1, and their spatial distribution is visualized in Figure 3.3.

Table 3.1: Transactions by city (July 2024 – July 2025).

City	Count	Share (%)
Riyadh	40,332	61.1
Jeddah	6,442	9.7
Makkah	4,151	6.3
Dammam	3,980	6.0
Buridah	2,862	4.3
Khobar	2,195	3.2
Other (11 cities)	6,043	9.3
Total	66,005	100

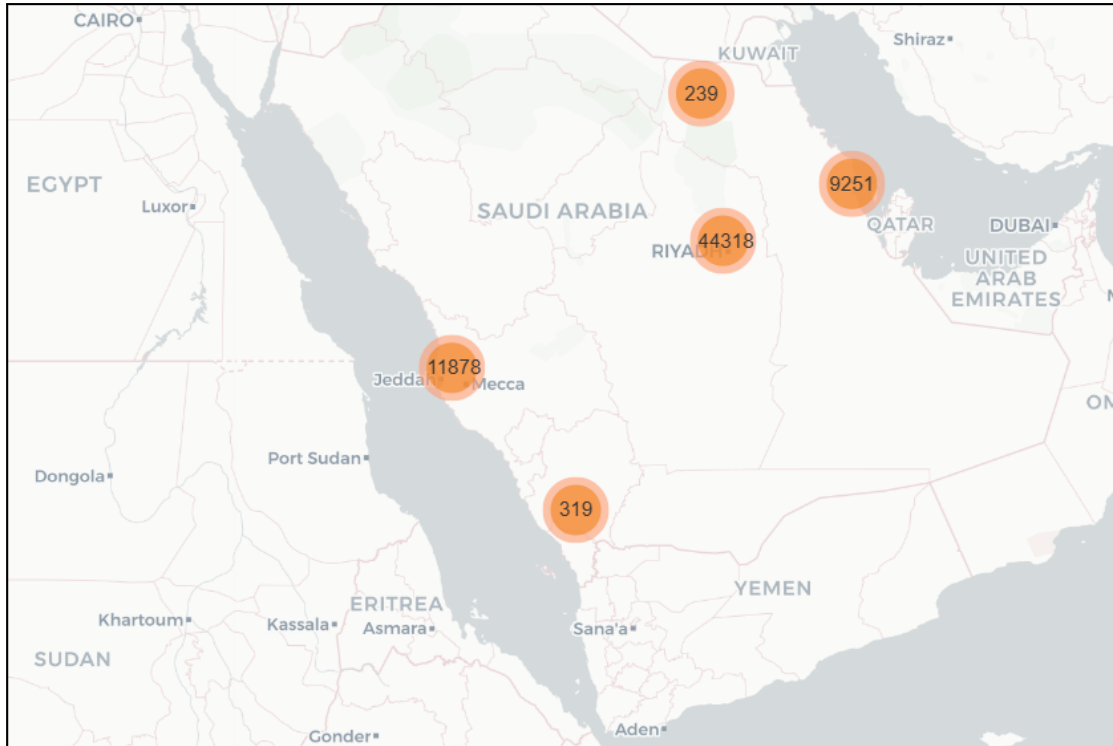
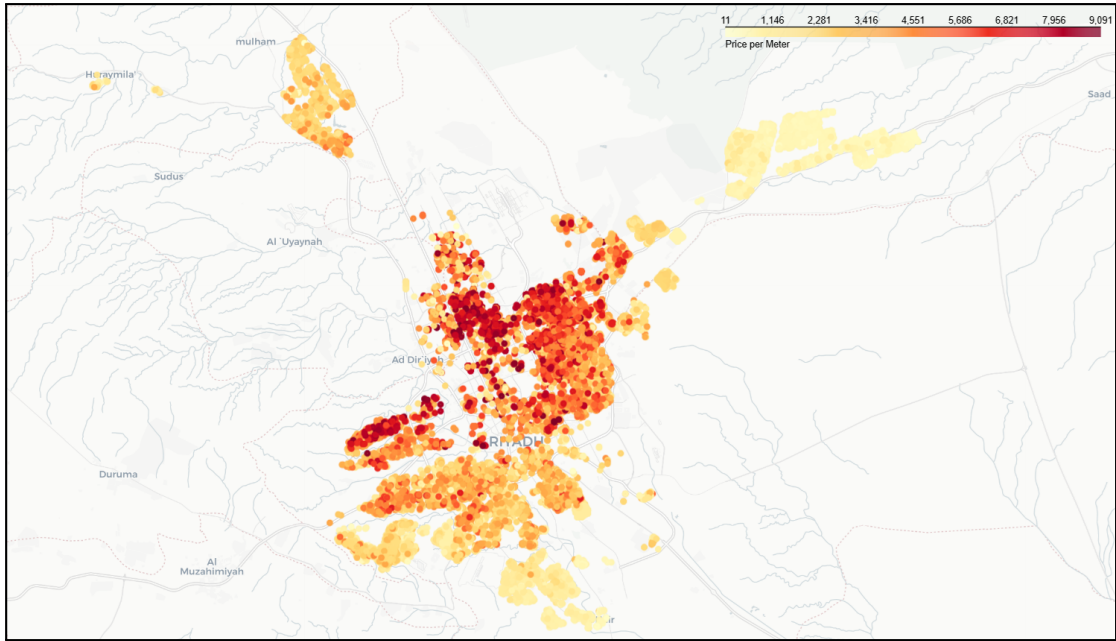
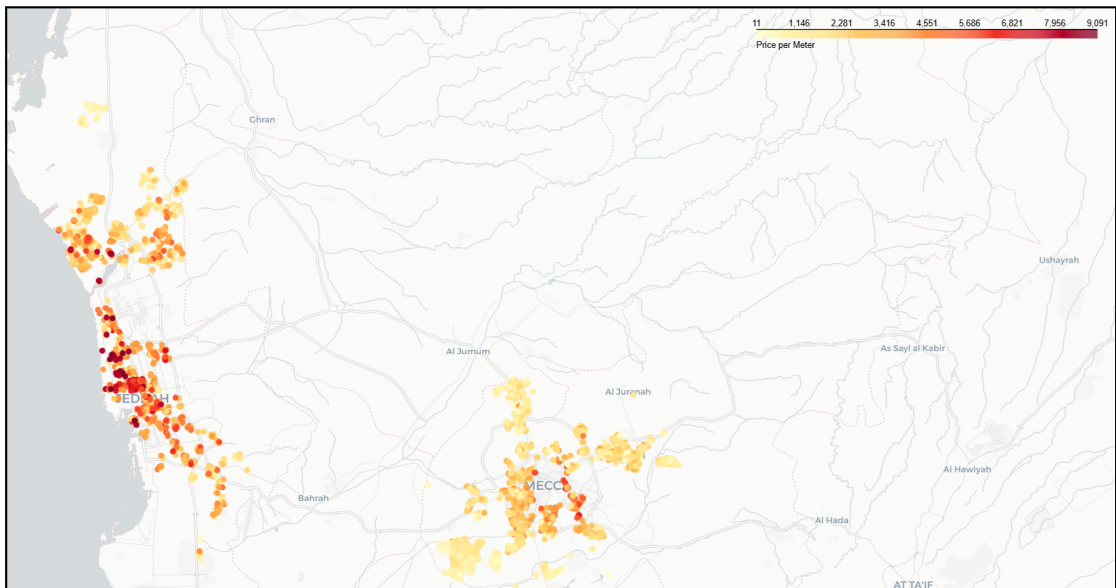


Figure 3.3: Spatial distribution of transactions across 16 cities.

Property types. Five principal categories are represented: apartments (25.52%), villas (7.12%), residential land (40.42%), commercial properties (12.59%), and other types (12.75%). This mixture supports model evaluation across distinct market segments and built forms. The distribution by property type is summarized in Table 3.2, and the price per m² variability in Riyadh, Makkah and Jeddah are shown in Figures 3.4, 3.5.

Table 3.2: Transactions by property type.

Property type	Count	Share (%)
Apartments	16,845	25.52
Villas	4,701	7.12
Residential land	27,737	42.02
Commercial	8,310	12.59
Other	8,403	12.75
Total	66,005	100

Figure 3.4: Price ($\text{ﷲ}/\text{m}^2$) distributions in Riyadh.Figure 3.5: Price ($\text{ﷲ}/\text{m}^2$) distributions in Makkah and Jeddah.

Attributes and units. Each transaction includes core structural attributes (e.g., built area, parcel area as applicable, precise geographic coordinates (WGS84), and the realized transaction price in Saudi Riyals (ﷲ). All prices reflect closed transactions, not listings, improving the reliability of the label for supervised learning. The core variables are documented in Table 3.3.

Table 3.3: Data dictionary for features.

Variable	Type/Unit	Description
sale_price	฿	Final transaction price (closed deal)
area	m^2	Built or parcel area (as appropriate)
price_per_m2	฿/ m^2	Target: sale_price/area after validation
lat, lon	degrees (WGS84)	Geographic coordinates of the property
sale_date	ISO date	Date of transaction
property_type	categorical	Apartment, villa, residential land, commercial, other
TBC later		

Data sources. Aggregated from official transaction feeds [29, 28].

3.3 Data Preprocessing

The data preprocessing pipeline implements comprehensive procedures to ensure data quality, consistency, and suitability for machine learning analysis. The preprocessing workflow addresses missing values, outlier detection, data type conversions, and feature standardization while preserving the integrity and interpretability of the original data.

This section is currently in active development.

3.4 Feature Engineering and Geographic Integration

This section is currently in active development.

3.5 Machine Learning Algorithms and Implementation

This section is currently in active development.

3.5.1 XGBoost (Extreme Gradient Boosting)

This section is currently in active development.

3.5.2 Random Forest

This section is currently in active development.

3.5.3 Gradient Boosting

This section is currently in active development.

3.5.4 Linear Regression Methods

This section is currently in active development.

3.5.5 K-Nearest Neighbors

This section is currently in active development.

3.5.6 Decision Tree Regressor

This section is currently in active development.

3.6 Experimental Design and Validation

This section is currently in active development.

3.7 Fairness Evaluation Framework

This section is currently in active development.

Chapter 4

Experimental Results and Analysis

4.1 Dataset Characteristics and Exploratory Analysis

This section is currently in active development.

4.2 Model Performance Comparison

This section is currently in active development.

4.3 XGBoost Model Analysis

This section is currently in active development.

4.3.1 Hyperparameter Optimization

This section is currently in active development.

4.3.2 Feature Importance Analysis

This section is currently in active development.

4.4 Geographic Feature Impact Analysis

This section is currently in active development.

4.5 Fairness and Bias Evaluation

This section is currently in active development.

4.5.1 Geographic Fairness Assessment

This section is currently in active development.

4.5.2 Property Type Fairness Assessment

This section is currently in active development.

4.5.3 Price Range Equity Analysis

This section is currently in active development.

4.6 Model Deployment Considerations

This section is currently in active development.

Chapter 5

Discussion

5.1 Interpretation of Results

This section is currently in active development.

5.2 Implications for Property Valuation Practice

This section is currently in active development.

5.3 Geographic Feature Integration Insights

This section is currently in active development.

5.4 Fairness and Bias Implications

This section is currently in active development.

5.5 Limitations and Challenges

This section is currently in active development.

5.6 Regulatory and Compliance Considerations

This section is currently in active development.

5.7 Future Research Directions

This section is currently in active development.

Chapter 6

Conclusion and Future Work

6.1 Summary of Findings

This section is currently in active development.

6.2 Research Contributions

This section is currently in active development.

6.2.1 Methodological Contributions

This section is currently in active development.

6.2.2 Technical Contributions

This section is currently in active development.

6.2.3 Practical Contributions

This section is currently in active development.

6.3 Implications for Industry and Policy

This section is currently in active development.

6.3.1 Industry Implications

This section is currently in active development.

6.3.2 Policy Implications

This section is currently in active development.

6.4 Limitations and Future Research Directions

This section is currently in active development.

6.4.1 Data and Methodology Limitations

This section is currently in active development.

6.4.2 Technical Research Directions

This section is currently in active development.

6.4.3 Application and Deployment Research

This section is currently in active development.

6.5 Final Remarks

This section is currently in active development.

Bibliography

- [1] Frederick M. Babcock. *The Valuation of Real Estate*. McGraw-Hill Book Company, New York, 1932. Available at: <https://archive.org/details/valuationofreale0000fred>.
- [2] Gallagher Mohan Real Estate. Real estate valuation techniques: Comparing appraisal methods, 2023. Available at: <https://www.gallaghermohan.com/real-estate-valuation-techniques>.
- [3] Superior School of Real Estate. Comparative market analysis vs. property appraisal, 2023. Available at: <https://www.superiorschoolnc.com/blog/comparative-market-analysis-vs-appraisal>.
- [4] Nick French and Laura Gabrielli. Valuation uncertainty: Common professional standards and methods. *Journal of Property Investment & Finance*, 22(2):150–163, 2004. Available at: <https://doi.org/10.1108/14635780410525105>.
- [5] Eleni Pagourtzi, Vassilis Assimakopoulos, Thomas Hatzichristos, and Nick French. Real estate appraisal: A review of valuation methods. *Journal of Property Investment & Finance*, 21(4):383–401, 2003. Available at: <https://doi.org/10.1108/14635780310483656>.
- [6] Hasan Selim. Determinants of house prices in turkey: Hedonic regression model. *Doğuş Üniversitesi Dergisi*, 10(1):65–76, 2009. Available at: <https://dergipark.org.tr/en/pub/doujournal/issue/6619/89063>.
- [7] Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974. Available at: <https://doi.org/10.1086/260169>.
- [8] Stephen Malpezzi. Hedonic pricing models: a selective and applied review. *Housing Economics and Public Policy*, pages 67–89, 2003. Available at: <https://doi.org/10.1002/9780470752145.ch5>.
- [9] Yongheng Deng, Pengyu Liu, and Qi Wang. Geospatial analysis in real estate: A review of gis applications in property valuation. *Habitat International*, 71:149–159, 2018. Available at: <https://doi.org/10.1016/j.habitatint.2017.11.010>.

BIBLIOGRAPHY

- [10] Mary L. Downie, Kenneth Gibb, and Moira Munro. Automated valuation models: An international perspective. *Journal of European Real Estate Research*, 12(1):92–111, 2019. Provides a comprehensive international review of Automated Valuation Models (AVMs), their methodologies, and implications for property markets. Available at: <https://doi.org/10.1108/JERER-05-2018-0022>.
- [11] Nils Kok, Elina-Leena Koponen, and Carmen A. Martínez-Barbosa. Big data in real estate? from manual appraisals to automated valuation. *The Journal of Portfolio Management*, 43(6):202–211, 2017. Explores how AVMs and big data enhance property valuation, investment analysis, and decision-making across real estate sectors. Available at: <https://doi.org/10.3905/jpm.2017.43.6.202>.
- [12] Jeonghyeon Kim, Youngho Lee, Myeong-Hun Lee, and Seong-Yun Hong. A comparative study of machine learning and spatial interpolation methods for predicting house prices. *Sustainability*, 14(15):9056, 2022.
- [13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. PDF: <https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf>.
- [14] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. Introduces SHAP values as a unified framework for model interpretability. Available at: <https://arxiv.org/abs/1705.07874>.
- [15] O. Shevchuk and O. Budiak. Property valuation with the use of machine learning methods: The case of ukraine. *Baltic Journal of Real Estate Economics and Construction Management*, 10(1):158–172, 2022. Available at: <https://doi.org/10.2478/bjreecm-2022-0010>.
- [16] Sangwoo Yoo and Phaedon Kyriakidis. Gis-based machine learning for real estate valuation. *Sustainability*, 12(9):3779, 2020. Available at: <https://doi.org/10.3390/su12093779>.
- [17] Elham Alzain, Ali Saleh Alshebami, Theyazn H. H. Aldhyani, and Saleh N. Alsubari. Application of artificial intelligence for predicting real estate prices: The case of saudi arabia. *Electronics*, 11(21):3448, 2022. Available at: <https://doi.org/10.3390/electronics11213448>.
- [18] Hernán D. Villada-Medina, Juan F. Rendón-García, César A. Ramírez-Dolores, and Gerardo Alcalá. A house price modeling based on clustering and kriging: The medellín case. *Review of Regional Studies*, 52(3):321–343, 2022. Available at: https://www.researchgate.net/publication/366696927_A_House_Price_Modeling_Based_on_Clustering_and_Kriging_The_Medellin_Case.

BIBLIOGRAPHY

- [19] Yang Wang, Shaojian Wang, Guangdong Li, Hongou Zhang, Lixia Jin, Yufeng Su, and Kangmin Wu. Identifying the determinants of housing prices in china using spatial regression and the geographical detector technique. *Applied Geography*, 79:26–36, 2017. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0143622816307597>.
- [20] Gabriela Droj, Nicolae Ionescu, and Ioan Pop. A comprehensive overview regarding the impact of gis on real estate valuation activities. *ISPRS International Journal of Geo-Information*, 13(6):175, 2024.
- [21] Wei-Wei Du, Wei-Yao Wang, and Wen-Chih Peng. Dora: Domain-based self-supervised learning framework for low-resource real estate appraisal. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2023. Available at: <https://doi.org/10.1145/3583780.3615470>.
- [22] Mahdiah Yazdani and Maziar Raissi. Real estate property valuation using self-supervised vision transformers. *arXiv preprint arXiv:2302.00117*, 2023.
- [23] Chunxu Zhao, Yuhei Ogawa, Shintaro Chen, Taikan Oki, and Yoshihide Sekimoto. Quantitative land price analysis via computer vision from street view images. *Engineering Applications of Artificial Intelligence*, 123:106294, 2023.
- [24] William C. Wheaton and Cunjia Xu. Using ai to improve price transparency in real estate valuation. Technical Report Research Paper No. 24/16, MIT Center for Real Estate, 2024. Available at: <https://doi.org/10.2139/ssrn.5056315>.
- [25] Dubai Land Department. Dld launches ‘instant sale’ feature using smart evaluation to simplify procedures and improve customer experience. <https://dubailand.gov.ae/en/news-media/dld-launches-instant-sale-feature-using-smart-evaluation-to-simplify-p> 2023. Accessed 19 Aug 2025.
- [26] JLL Middle East & North Africa. How ai is shaping the future of real estate (mena insight). <https://www.jll-mena.com/en/trends-and-insights/cities/how-ai-is-shaping-the-future-of-real-estate>, 2022. Accessed 19 Aug 2025.
- [27] Saudi Data & AI Authority (SDAIA). National strategy for data & ai. <https://sdaia.gov.sa/en/SDAIA/SdaiaStrategies/Pages/NationalStrategyForDataAndAI.aspx>, 2025. Strategy linked to Vision 2030; accessed 20 Aug 2025.
- [28] Real Estate General Authority (REGA). Real estate indicators — rega platforms. <https://rega.gov.sa/en/reg-services/platforms/real-estate-indicators/>, 2025. Digital indicators at Kingdom/city/neighborhood levels with comparative views; accessed 20 Aug 2025.

BIBLIOGRAPHY

- [29] Ministry of Justice (KSA). Real estate market — live transaction feed (service details). <https://www.moj.gov.sa/English/eServices/Pages/ServiceDetailsNew.aspx?itemId=433>, 2025. Lists transaction fields including price per square meter, location, area, and time; last modified 28 Jun 2025; accessed 20 Aug 2025.
- [30] Saudi Authority for Accredited Valuers (TAQEEM). About us. <https://taqeem.gov.sa/en/about-us>, 2025. Regulator of the valuation profession; role, mission, and standards in line with Vision 2030; accessed 20 Aug 2025.
- [31] Quara Holding. Wasalt achieves major growth milestone in saudi arabia’s proptech market. <https://www.quaraholding.com/wasalt-achieves-major-growth-milestone-in-saudi-arabias-proptech-market>, 2025. Reports 750,000+ registered users and 100,000 monthly leads; Accessed 19 Aug 2025.
- [32] TradeArabia. Digital real estate platform wasalt posts significant growth. <https://www.tradearabia.com/News/429985/Digital-real-estate-platform-Wasalt-posts-significant-growth>, 2025. Summarizes AI features incl. instant valuation; Accessed 19 Aug 2025.
- [33] Linna Zhu, Michael Neal, and Caitlin Young. Racial disparities in automated valuation models: New evidence using property condition and machine learning. *Cityscape: A Journal of Policy Development and Research*, 26(1):297–320, 2024.
- [34] CFPB, OCC, FRB, FDIC, NCUA, FHFA. Quality control standards for automated valuation models (final rule). <https://www.consumerfinance.gov/rules-policy/final-rules/quality-control-standards-for-automated-valuation-models/>, 2024. Final rule published Aug 7, 2024; Accessed 19 Aug 2025.
- [35] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, Jose J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Methods in Ecology and Evolution*, 8(4):539–552, 2017. Guidance on cross-validation under temporal and spatial dependence; supports time-aware and spatial-block validation designs. Available at: <https://doi.org/10.1111/2041-210X.12674>.
- [36] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2023. Introductory tutorial on conformal prediction for valid prediction intervals without distributional assumptions. Available at: <https://arxiv.org/abs/2107.07511>.