BABEȘ-BOLYAI UNIVERSITY OF CLUJ-NAPOCA
FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION
Doctoral School of Economics and Business Administration

SUMMARY OF THE DOCTORAL THESIS
Titled

INTEGRATING MACHINE LEARNING MODELS INTO REAL ESTATE PROPERTY VALUATION

Scientific Coordinator:
Professor Adela DEACONU, Ph.D., Habil.

PhD Candidate:
Silviu-Ionuț BĂBȚAN

Cluj-Napoca, 2025

TABLE OF CONTENTS

Introduction

This thesis explores the use of machine learning methods in real estate valuation, aiming to improve the traditional valuation process, which is often costly, time-consuming, and influenced by subjective factors. In the context of a dynamic and increasingly digitalized real estate market, the adoption of Automated Valuation Models (AVMs) becomes both a practical and strategic necessity.

It highlights that AVMs, through the integration of artificial intelligence, can deliver fast, objective, and consistent results, contributing to cost reduction and the elimination of human errors. The study proposes the application and comparison of five predictive methods—linear regression, decision trees, random forests, neural networks, and XGBoost—on a sample of apartments located in Cluj-Napoca.

The research is driven by the need for sustainable and innovative solutions in property valuation and is logically structured into four theoretical and applied chapters, followed by conclusions and future research directions. The aim of the thesis is to identify the most effective predictive method applicable to the Romanian market, while also contributing to the advancement of research in the field of automated property valuation.

Chapter 1 – Asset Valuation for Different Purposes

Real estate valuation has rapidly developed in Romania due to the growth of the property market and the increasing demand for specialized services. This professional activity, part of the broader consultancy sector, requires advanced competence, adherence to a code of ethics, and the application of international standards.

1.1 General Delimitations

1.1.1 History of the Valuation Profession

The origins of the profession trace back to the United Kingdom (1834), with its global development marked by the establishment of key organizations such as RICS (UK), the Appraisal Institute (USA), and IVSC (international level). In Europe, standardization has been consolidated through TEGoVA. In Romania, the profession was formalized after 1990 with the founding of ANEVAR, a public utility institution that regulates and supports the activity of certified valuers.

1.1.2 Key Concepts: Price, Cost, and Value

Valuation involves distinguishing between:

- **Price** – the actual amount exchanged in a transaction;
- **Cost** – the expenditure required to create an asset;
- **Value** – the market perception of an asset's usefulness.

These are influenced by factors such as utility, scarcity, preference, and purchasing power, reflecting the equilibrium between supply and demand.

1.1.3 Object of Valuation

Valuation is a professional activity based on expert judgment, expressed in an official report. It applies to distinct asset categories:

- Real estate properties;
- Movable goods;
- Businesses and equity holdings;
- Financial instruments.

This thesis focuses on residential real estate, including associated partial rights (e.g., use, lease, mortgages).

1.2 Valuation Approaches, Methods, and Techniques

Real estate valuation involves applying one or more of the three internationally recognized approaches: cost, market, and income. The choice of method depends on the property type and market context.

1.2.1 Market Approach

This method relies on the comparative analysis of recent transactions involving similar properties. It is the benchmark method for residential valuations, as it reflects actual market behavior. The process involves identifying comparable properties, adjusting differences between these and the subject property, and using appropriate units of measurement (e.g., €/sqm or €/hectare).

Key elements in the analysis include location, building condition, property rights, legal status, and market conditions. Comparability is validated by selecting recent transactions, preferably from the same area, and applying explicit price adjustments, expressed in absolute or percentage terms.

1.2.1.1 Price-Influencing Factors

A property's price is affected by a wide range of variables—from the physical characteristics of the land and building to urban regulations, utilities, location, and accessibility. Emphasis is placed on identifying and systematically correcting differences between the analyzed transactions and the subject property.

1.2.1.2 Adjustment Techniques

Adjustments can be made through quantitative methods (paired data analysis, regression, statistical trends) or descriptive techniques (relative comparison, classifications, expert interviews). Each technique helps to fine-tune values based on specific property traits.

1.2.1.3 Compliance with Romanian Standards

The direct comparison method aligns with the Romanian Property Valuation Standards (SEV) and is widely used for valuing property rights. It requires the use of relevant and market-accepted comparison units, such as €/sqm or €/room, and the coherent application of such metrics.

The valuer must pay close attention to property differences, prioritize recent transactions, and clearly describe all adjustments in the report. In the absence of direct comparisons, indirect methods (ratio, deduction) may be used, albeit with limited accuracy.

1.3 Valuation Approaches, Methods, and Techniques

In practice, three major methodological approaches are used in property valuation: market, income, and cost. Each is grounded in solid economic principles and is selected according to the property type and data availability.

1.2.1 Market Approach

This method compares the subject asset with recent transactions involving similar properties. The valuer selects relevant units of measurement (€/sqm, €/room, etc.), identifies differences, and applies specific adjustments to reflect the characteristics of the analyzed property.

Comparability is essential, and the techniques used include both quantitative analyses (regressions, paired data, graphics) and descriptive analyses (comparative classifications, interviews). It is the dominant approach in residential property valuation, complying with both Romanian and international standards. Its accuracy depends on the quality and recency of market data.

1.2.2 Income Approach

Applied to income-generating properties (e.g., office buildings, hotels), this approach converts future income flows into present value. There are two main methods:

- **Direct Capitalization** – simple, based on stable annual income and capitalization rate;
- **Discounted Cash Flow (DCF)** – updates multiple income flows over a time horizon, being more complex and suited to dynamic forecasts.

This approach requires a deep understanding of the rental market, risks, and investor expectations, and is supported by the principles of demand and anticipation.

1.2.3 Cost Approach

This method estimates a property's value starting from its replacement or reproduction cost, minus physical, functional, and external depreciation. It is preferred for new, specialized constructions or when transaction data is insufficient.

Valuation involves the following steps:

- Calculating the replacement cost (via unit cost, cost estimate, or comparison);
- Determining depreciation (recoverable or non-recoverable);
- Estimating net value.

Costs are categorized into direct (materials, labor, contractor profit) and indirect (fees, studies, taxes, marketing), which must be completely and accurately integrated into the final analysis.

1.4 Purpose and Utility of Valuation

The purpose of real estate valuation is to estimate a value associated with a property right, depending on the client's specific needs. It may serve a wide range of objectives—from commercial transactions and banking collateral to taxation and financial reporting.

The valuation purpose is determined by the client and guides informed decision-making regarding the real estate asset. Although the report format may vary by use, the estimated value remains unchanged, as it results from a standardized and objective process.

This thesis details three of the most frequent valuation purposes:

Valuation for Sale/Purchase

The most common form of valuation, used to support the negotiation of a fair price between parties. Automated Valuation Models (AVMs) can streamline this process by offering quick and consistent estimates based on historical and predictive data. They reduce subjectivity and support transparent decisions for both buyers and sellers.

Valuation for Banking Collateral

Used by banks to determine the value of collateral, this form is crucial in the lending process. Integrating machine learning allows for fast, precise, and standardized estimates, reducing the risk of over- or undervaluation. For banks, this improves risk management and decision-making efficiency.

Valuation for Taxation

Used to determine the taxable value of properties, directly impacting tax liabilities. It is legally regulated and must be performed periodically, especially for non-residential buildings. Predictive models can support a more equitable fiscal system through uniform, transparent, and automated estimates, contributing to efficient public revenue collection.

In all three cases, automating the valuation process through machine learning improves accuracy, reduces costs, and provides robust decision-making support for all stakeholders: owners, financial institutions, authorities, and developers.

1.5 Conclusions of Chapter 1

Real estate valuation is a vital pillar of the modern economy, reinforced by professional standards, expert judgment, and solid institutional infrastructure. In Romania, the profession has expanded rapidly since 1990, regulated by ANEVAR and aligned with international organizations such as IVSC and TEGoVA.

The valuation process is not limited to mathematical calculations—it involves professional reasoning and in-depth market knowledge. The three fundamental approaches—market, income, and cost—are applied based on the purpose of the valuation and asset characteristics, with the market approach being dominant in the residential segment.

The estimated value forms the basis for critical economic decisions: sales transactions, banking collateral, taxation, and accounting reporting. Therefore, valuation serves the needs of individuals, companies, financial institutions, tax authorities, and investors alike.

In particular, valuations for sale-purchase, banking collateral, and taxation are of utmost relevance in today's context. These can be significantly optimized through the integration of machine learning, enabling massive data analysis and providing rapid, objective, and scalable estimates. Predictive models reduce uncertainty, eliminate subjectivity, and support transparency in decision-making.

This chapter provides the theoretical framework necessary to understand not only the essence of the valuation process but also the justification for using advanced technologies in this field. Thus, it lays the foundation for the subsequent research direction—automating the real estate valuation process—by clarifying the concepts, purposes, methods, and standards of valuation.

Chapter 2 – Literature Review on Automated Valuation Models (AVMs)

This chapter explores the evolution, applicability, and validation of Automated Valuation Models (AVMs), emphasizing the synergy between traditional valuation methods and machine learning algorithms. AVMs utilize extensive datasets and predictive models to estimate the market value of properties without direct human intervention. This approach enhances the accuracy, efficiency, and objectivity of the valuation process and is increasingly adopted in the context of real estate market digitalization.

2.1 Literature Review Methodology

A total of 203 articles were selected from over 20,000 identified in scientific databases such as Scopus, Web of Science, and Science Direct. The selection was based on strict criteria: thematic relevance, scientific impact, full accessibility, and methodological validation. The focus was placed on studies directly addressing AVMs, using quantitative/statistical methods, and exploring their implications in real estate valuation.

Key Themes Identified in the Literature

The research revealed several main directions:

- Definition and classification of AVMs;

- Types of data and variables used;
- Applied statistical models (regression, neural networks, hybrid methods);
- The utility of AVMs in various geographic and institutional contexts;
- SWOT analysis and the regulatory impact on AVM adoption.

The chapter's central table summarizes 14 representative articles, detailing their research goals, methodologies, variables used, sample sizes, and findings relevant to the present study.

Integration of Quantitative and Qualitative Models

Quantitative studies analyze the performance and accuracy of AVMs on real-world datasets, while qualitative studies provide a conceptual and methodological framework for understanding the context of their application. Together, these approaches support the development of a custom model tailored to the Romanian market.

Contributions and Relevance

This chapter underlines the necessity of applying AVMs in real estate valuation, offering a solid reference framework for defining the research methodology. Through bibliometric analysis and critical evaluation of the literature, the most relevant sources, methodologies, and independent variables are identified for building a robust predictive model.

2.2: Bibliometric Analysis of the Literature

This subchapter presents a detailed bibliometric investigation of the scientific literature on AVMs, aiming to highlight research trends, academic centers involved, and dominant themes. The analysis is based on the Web of Science database and includes 318 articles published between 1998 and 2024. Although the total volume is relatively modest, research activity has clearly intensified, with a peak in publications in 2017 and a citation maximum in 2024.

The studies are primarily distributed across artificial intelligence and computer science fields, but also extend into economics, finance, urban studies, and engineering. Emerging technologies—such as machine learning, regression, decision trees, neural networks, and ensemble methods (e.g., random forest, XGBoost)—are frequently associated with AVMs, reflecting major development directions.

Geographical distribution shows a research concentration in advanced economies (USA, UK, China, Germany), with notable contributions also from emerging European and Asian countries. High-impact journals in this domain come from various disciplines—from real estate valuation and finance to artificial intelligence—suggesting a consolidated interdisciplinary approach.

The terminological analysis using VOSviewer highlights strong links between AVMs and concepts like property prices, mass appraisal, real estate markets, and mortgage lending. These connections validate the relevance of the thesis topic and support the integration of machine learning models in contemporary real estate valuation.

In conclusion, the bibliometric analysis offers a solid foundation for further in-depth research, revealing both methodological advancements and the need for enhancing practical applicability and model transparency. This subchapter thus prepares the ground for a systematic qualitative investigation in the following section.

2.3: Systematic Analysis of Empirical Studies

This subchapter explores the literature through a systematic analysis focused on empirical studies and the methods used to develop AVMs. Unlike the bibliometric approach, this analysis qualitatively examines the most relevant 50 articles from Web of Science and Scopus, selected based on citation count and thematic relevance.

The synthesis highlights the applicability of AVMs in various contexts: residential appraisals, tax assessments, collateral lending, and market analysis. It also reviews the statistical methods and machine learning algorithms frequently used—linear regression, hedonic regression, decision trees, random forests, neural networks, and XGBoost—based on research goals and data quality.

Most analyzed studies noted significant advantages of AVMs in terms of efficiency and accuracy, but also pointed out limitations such as lack of transparency, difficulty in interpreting results, and dependence on data quality. Additionally, some studies emphasized integrating human expertise and hybrid techniques to increase estimation robustness.

The analysis confirms the multidisciplinary nature of AVM research, with contributions from fields such as AI, economics, urban planning, and engineering. It also highlights the need to standardize methodologies and datasets to facilitate result comparability and reproducibility.

In conclusion, the systematic analysis emphasizes the significant progress made in designing and testing AVMs, while also noting the need for continued research into spatial data integration, model explainability, and adaptability to different market conditions.

Conclusions of Chapter 2

Chapter 2 offers a comprehensive overview of how AVMs have been addressed in the scientific literature, emphasizing the transition from traditional to automated methods. The review, based on reputable sources, highlights the frequent use of AVMs in advanced real estate markets and their consolidation as valuable tools in the digitalization of property valuation.

The reviewed studies confirm AVMs' effectiveness in terms of speed, reduced costs, and mass analysis capabilities, but also highlight their limitations: dependence on high-quality data, difficulty adapting to less liquid markets, and challenges related to algorithm transparency. There is also a notable need for hybrid integration between automated models and human expertise to enhance estimation validity and robustness.

The analysis shows that AVMs are most effective in mass valuations, especially for taxation, collateral assessment, and risk analysis. Recent studies focus on optimizing input variables, avoiding overfitting, and continuously calibrating models. At the same time, hybrid approaches and the use of AI support the methodological advancement of AVMs.

The summary table included in the chapter compares the main benefits and limitations of AVMs, reinforcing the conclusion that these models offer valuable support in decision-making but cannot fully replace professional judgment. Thus, AVMs emerge as a complementary solution, with significant expansion potential given the increasing volume of data and the need for efficiency in property valuation.

Chapter 3: Machine Learning Models and Their Applicability in Real Estate Valuation

Chapter 3 explores the applicability of machine learning (ML) models in real estate valuation, highlighting their ability to analyze large volumes of data and identify complex, non-linear relationships often missed by traditional methods. In an increasingly dynamic real estate market, ML emerges as an effective solution for improving accuracy and objectivity in property value estimation.

The main ML models used in property valuation are presented: linear regression, decision trees, random forests, artificial neural networks, and Extreme Gradient Boosting (XGBoost). Each technique is analyzed based on its operating principles, advantages, and limitations, demonstrating their suitability for various market contexts. Model performance is assessed using specific indicators—RMSE, MAE, R^2, F-statistic, and P-value—to ensure prediction validity.

RStudio is mentioned as the preferred environment for implementing and testing these models, offering flexibility in data processing, hyperparameter tuning, and result interpretation.

3.1 analyzes the extension of machine learning into integrated artificial intelligence (AI) systems, classifying its domains (learning, natural interfaces, perception, robotics, etc.) and identifying relevant applications in property valuation—especially expert systems, neural networks, and intelligent agents. AI not only supports but enhances ML's predictive potential, reinforcing the objectivity and efficiency of the valuation process.

The included tables provide an overview of AI domain classifications (Table 7) and the practical implementation of ML models in various industries (Table 8)—ranging from energy and agriculture to financial markets and healthcare. This cross-industry presence confirms the versatility of ML models and their transfer potential to real estate valuation.

In conclusion, machine learning is a scientifically validated and innovative tool for property value estimation, supporting the digital transformation of the valuation process. ML models enable deep data analysis, reduce human error, and facilitate transparent and predictive decision-making—paving the way for a modernized real estate market.

3.2: Machine Learning Models Used in Property Valuation

Subchapter 3.2 provides a comparative analysis of the most commonly used ML models applicable to real estate valuation, focusing on estimation accuracy, robustness, and computational efficiency. The models examined include: linear regression, decision trees, random forests, artificial neural networks, and Extreme Gradient Boosting (XGBoost).

Each model is discussed in terms of its core mechanics, benefits, and specific limitations. Linear regression stands out for its simplicity and interpretability, being suited for direct relationships between variables. Decision trees can model complex relationships and offer clear decision paths but may lack stability. Random forests improve stability and precision by combining multiple trees.

Neural networks adapt well to non-linear patterns but require significant computational resources and a large dataset. XGBoost, an advanced boosting algorithm, is highly efficient in reducing errors and handling outliers and complex variables with strong performance.

Model evaluation is based on performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), providing an objective comparison framework. Case studies and experimental tests indicate that XGBoost offers the best balance between predictive accuracy and data processing efficiency, while random forests deliver robustness and good generalization across diverse market conditions.

In conclusion, selecting the optimal model depends on the real estate market context, the volume and quality of available data, and the valuation purpose. Advanced models like XGBoost and neural networks are recommended for high-precision predictions, while simpler models are preferred when interpretability or limited data is a concern. Thus, ML proves to be a versatile and scalable tool in the modern property valuation process.

3.4: Performance Evaluation of Machine Learning Models

Subchapter 3.4 emphasizes the importance of rigorously evaluating ML models used in property value estimation. Performance analysis relies on a set of key statistical metrics that allow for the objective comparison of algorithm accuracy and support the selection of the most efficient model.

The most relevant validation tools include:

- **RMSE (Root Mean Squared Error)** – penalizes large errors and provides an overall view of deviations from actual values;
- **MAE (Mean Absolute Error)** – expresses the average error in absolute terms and is less sensitive to outliers;
- **R^2 (coefficient of determination)** – shows the proportion of price variability explained by the model;
- **F-statistic and P-value** – essential for testing the statistical significance of relationships between variables.

RMSE is useful when large errors are costly, while MAE offers a more intuitive interpretation closer to operational reality. Comparing RMSE and MAE values can signal the presence of outliers. R^2 helps understand the model's explanatory power but may overestimate performance in

overfitted scenarios. F-statistic and P-value assess the overall validity of the model and the contribution of each variable to prediction accuracy.

These evaluations are performed using specialized software environments such as RStudio, which provides a robust framework for predictive performance analysis. In this research, applying these metrics enabled an objective comparison of tested models and supported the selection of algorithms with the highest accuracy in estimating property values.

3.5: Workflow in RStudio Software – Original Summary

RStudio is an integrated development environment (IDE) based on the R language, used for advanced data processing and analysis, and widely applied in predictive modeling. Its intuitive interface supports script editing, code execution, data visualization, and package management.

Data import is easily achieved using functions like read.csv() or read_excel(), followed by essential cleaning and transformation steps—commonly performed using the dplyr package. Statistical modeling can be conducted with functions like lm() for linear regression, alongside relevant significance tests.

A major advantage of RStudio is its powerful visualization capability through ggplot2, enabling the creation of customizable graphs—useful for exploring relationships between variables. Additionally, RStudio supports the creation of interactive, reproducible reports in PDF/HTML format via rmarkdown, combining code, text, and analytical results in a single document.

In the context of this thesis, RStudio is used to implement and validate ML models for real estate value estimation. It plays a key role in applying statistical methods and evaluating algorithm performance using indicators such as RMSE, MAE, R^2, F-statistic, and P-value.

Chapter 3 Conclusions

This chapter examines the applicability of ML models in estimating property values, emphasizing their advantages over traditional methods—especially in complex and dynamic contexts. Machine learning enables the identification of non-linear relationships between variables and facilitates accurate, objective, and reproducible predictions.

The first part discusses the integration of these technologies into property valuation, including the broader AI framework and relevant subfields (e.g., expert systems, neural networks, intelligent agents). Commonly used models include linear regression, decision trees, random forests, artificial

neural networks, and XGBoost—each offering distinct benefits, but also limitations related to interpretability, computational cost, or risk of overfitting.

The chapter includes an analysis of model performance using statistical indicators such as RMSE, MAE, R^2, F-statistic, and P-value—essential for result validation. The importance of balancing accuracy and robustness is emphasized.

Furthermore, the practical implementation in RStudio is described, the platform used to develop and test models, focusing on data import, cleaning, visualization, and report generation.

In conclusion, no model is universally superior—the optimal choice depends on the valuation context and the nature of the available data. This chapter lays the groundwork for the case study that will demonstrate the practical applicability of these concepts in analyzing real property prices.

Chapter 4 – Evaluating the Performance of Machine Learning Models in Predicting Residential Property Prices

In a context characterized by the increasing volume of data and rapid progress in the field of machine learning, this chapter analyzes the practical application of five predictive techniques for estimating the value of residential apartments in Cluj-Napoca. The main objective is to identify the method with the highest accuracy, adapted to the specific features of the local market.

The comparative study includes: linear regression, regression trees, random forests, artificial neural networks (ANN), and XGBoost. Each technique is evaluated based on predictive performance, implementation complexity, and robustness to data variation. The model-building process involves variable selection, hyperparameter optimization, and rigorous validation through cross-validation. The metrics used – RMSE and MAE – provide a balanced assessment between estimation accuracy and resilience to outliers.

This approach supports the shift from traditional methods to automated solutions, aligned with the dynamics of a complex real estate market. Implementing these predictive models aims to generate fast, objective, and reproducible estimates, contributing to the professionalization of real estate valuation and the optimization of economic decision-making in the residential sector.

4.2 – The Dataset

To support the machine learning models aimed at predicting housing prices, the research utilized a rigorously constructed dataset representative of the real estate market in Cluj-Napoca. This city

was chosen due to its leading position in the residential sector in Romania, known for high prices, strong demand, and average incomes above the national average.

The dataset consists of 773 apartments with one to four rooms, selected from two complementary sources: the Argus platform, intended for real estate professionals, and valuation reports produced by the company Napoca Business S.R.L. The selection process included only listings that provided complete and verifiable information, with incomplete, redundant, or unreliable entries being eliminated.

The analysis covers the years 2019–2020, and the data was manually completed and validated by the author, a certified appraiser, including phone calls to sellers. This process ensured a high degree of accuracy and market fidelity. Over 90% of the initial listings were excluded to guarantee the integrity and relevance of the final database.

This dataset provides a solid foundation for statistical calibration and testing of predictive model performance, allowing a nuanced exploration of the relationships between variables and property prices. The following subsection details these variables and how they contribute to predictive model construction.

4.3 – The Variables

The study relies on a complex set of 36 variables, divided into two main categories: quantitative (e.g., usable area, number of rooms, distance from the center) and qualitative (categorical) (e.g., building type, amenities, neighborhood). These variables reflect current practices in traditional real estate valuation and were properly encoded in RStudio for optimal integration into machine learning models.

For the analyzed area – Cluj-Napoca – variable selection aimed to cover essential apartment characteristics: location, structure, layout, amenities, and urban accessibility. The information was collected from the Argus platform listings, processed, and aggregated by neighborhood, with variables also validated using tools such as Google Maps.

Key variables include: neighborhood, area, construction type, price, distance from the center, and proximity to urban points of interest (institutions, green spaces, shopping centers). Comfort-related variables (balcony, elevator, parking space, storage room), as well as those concerning energy efficiency and pollution levels, were also included.

Special attention was given to the "concentration" variable, which quantifies building density within a 500 m² radius and was transformed into an ordinal factor (low, medium, high). Additionally, buyer preferences regarding the number of rooms, layout type, construction type (new/old), and availability of parking were analyzed using graphical representations generated in RStudio.

The collected data and variable structure decisively support the development of robust predictive models, reflecting the real dynamics of Cluj-Napoca's real estate market. This framework allows for identifying the most influential factors in determining apartment value and supports decision-making in real estate valuation and investment.

4.4 – Results and Interpretation

The comparative analysis of the five machine learning models applied to apartment price prediction in Cluj-Napoca highlighted significant performance differences. The tested models – linear regression, decision trees, random forests, artificial neural networks, and XGBoost – were evaluated using RMSE and MAE metrics, employing cross-validation for robustness.

The XGBoost model stood out as the best-performing, achieving the lowest prediction error values. It demonstrated superior capacity to capture complex, nonlinear relationships between the analyzed variables. Artificial neural networks (ANN) and random forests followed, both showing good results but requiring more computational resources.

While linear regression was interpretable and easy to implement, it delivered the weakest results, confirming its limitations in nonlinear and complex market contexts. Decision trees, though simple and explainable, had moderate performance but were prone to overfitting.

These results support the idea that advanced models like boosting and neural networks offer more accurate property price estimations, especially in dynamic markets such as Cluj-Napoca. Additionally, the analysis shows that model selection should consider evaluation goals, variable complexity, and available resources.

4.4.2 – Decision Trees

The decision tree model (CART) was implemented for apartment price estimation, calibrated using grid search for hyperparameter optimization. Over 21,000 combinations were tested, each evaluated through 5-fold cross-validation, using RMSE, MAE, and R^2 metrics.

The key adjusted hyperparameters included tree depth, minimum terminal node size (MinBucket), minimum observations required for a split (MinSplit), and complexity parameter (CP). The optimal result was obtained for a tree with depth 6, MinSplit = 10, MinBucket = 19, and CP = 0.001, producing an RMSE of 18,052.18 and R^2 of 0.732.

The decision tree diagram showed that apartment area was the most influential splitting factor, followed by neighborhood, number of rooms, thermal insulation, and parking availability. A variable importance ranking confirmed the essential role of size and layout in determining price.

The model effectively captured nonlinear relationships but required careful hyperparameter tuning to avoid overfitting. Compared to linear regression, decision trees offered clear visual interpretation and greater flexibility for real estate data analysis. However, limitations in robustness led to the exploration of a superior method – random forests, analyzed in the next subsection.

4.4.3 – Random Forests

The random forest method, proposed as the third technique in the case study, aggregates predictions from an ensemble of decision trees to produce a robust apartment price estimate. Implementation was performed in RStudio using the caret, randomForest, and dplyr libraries, with 5-fold cross-validation and hyperparameter tuning via grid search.

Tuned parameters included the number of trees (Ntree), the number of variables selected at each split (Mtry), and minimum node size (Nodesize). After testing 345 combinations, the optimal configuration was: • Ntree = 179 • Mtry = 22 • Nodesize = 9 • RMSE = 17,532.62 • R^2 = 81.96%

These metrics indicate high accuracy and strong generalization capacity, superior to previously tested models. Variable importance was evaluated using the Mean Decrease in Accuracy method, highlighting area, number of rooms, number of floors, distance from center, and neighborhood as key price predictors.

The learning curve confirmed that after about 200 trees, model performance stabilized, indicating the optimal point was reached without overfitting. Variable importance analysis also showed that urban infrastructure (public transport, healthcare institutions) had moderate impact, while proximity to shopping centers or religious institutions had negligible influence.

Compared to previous models, random forests outperformed linear regression and decision trees, providing a stable, interpretable, and efficient estimate of apartment value in Cluj-Napoca.

4.4.4 Artificial Neural Networks (ANNs)

In the case study, dense artificial neural networks (ANNs) were employed to model the complex and nonlinear relationships between apartment features and their prices. The implementation was carried out in Google Colab using the Keras library in R. Data preparation included normalization of numeric variables and one-hot encoding of categorical variables.

The model was trained through a rigorous optimization process involving:

- selecting the optimal number of epochs (identified as 173),
- using the RMSprop algorithm with fine-tuned hyperparameters for learning rate and momentum,
- testing various network architectures with one or two hidden layers, each containing between 32 and 512 neurons.

Model performance was evaluated through 5-fold cross-validation on the training set using RMSE and MAE metrics. The optimal architecture consisted of two hidden layers (64 and 32 neurons), and the chosen combination of hyperparameters (learning rate = 0.005, momentum = 0.95) resulted in an RMSE of 16,940.07—the lowest among all analyzed models.

The learning curve demonstrated a gradual reduction in error on both training and validation sets, confirming the network's efficiency in the learning process. Divergence observed at higher epoch counts signaled overfitting risks, which were mitigated through careful parameter tuning.

The ANN model outperformed linear regression, decision trees, and random forests, offering superior accuracy in estimating real estate prices. This outcome confirms the high potential of dense neural networks in capturing complex patterns in the real estate market and supporting automated valuation decisions.

4.4.5 Extreme Gradient Boosting (XGBoost)

The XGBoost algorithm, renowned for its efficiency in modeling complex relationships, was implemented to estimate apartment prices using decision trees and boosting techniques. Model

training involved a rigorous hyperparameter optimization process conducted solely on the training set validated through 5-fold cross-validation.

The extensive hyperparameter set included:

- number of trees (nrounds),
- tree depth (max_depth),
- learning rate (eta),
- regularization coefficients (gamma, min_child_weight),
- percentage of features selected (colsample_bytree).

Over 5,000 parameter configurations were tested, and model performance was evaluated using RMSE, R^2, and MAE. The best configuration included:

- eta = 0.05,
- max_depth = 2,
- gamma = 10,
- colsample_bytree = 0.75,
- min_child_weight = 1,
- 750 iterations,
- resulting in an RMSE of 16,706.96, R^2 = 82.3%, and MAE = 11,551.2.

The model demonstrated excellent generalization ability, balancing complexity and accuracy through advanced regularization mechanisms and iterative error adjustment. Additionally, variable importance analysis via xgb.importance highlighted the most influential factors in determining final price.

XGBoost proved to be the most accurate model tested in this research, surpassing previous methods through a robust, scalable approach to handling nonlinear relationships.

4.5 Results Obtained

The case study applied five machine learning methods to estimate apartment prices in Cluj-Napoca using the same dataset and validation method (5-fold cross-validation). Model performance was compared based on RMSE values.

4.5.1 Comparative Analysis of Applied Methods

The models tested included linear regression, decision trees, random forests, artificial neural networks (ANN), and XGBoost. RMSE values revealed a clear performance ranking:

- XGBoost: 16,706.96
- ANN: 16,940.07
- Random Forest: 17,533.62
- Linear Regression: 17,780.00
- Decision Trees: 18,052.18

These results validate the research hypothesis: complex methods like XGBoost and ANN offer superior predictive accuracy by better capturing the relationships between explanatory variables and apartment prices.

4.5.2 Performance of the XGBoost Model in Price Estimation

The XGBoost model, trained on 80% of the data and tested on the remaining 20%, achieved an RMSE of 16,706.96 and an MAE of 11,551.20. The differences between actual and estimated values were moderate, with a notable example being a prediction deviating by only 44 EUR from the real price.

The associated plot highlighted a strong correlation between estimated and actual values, confirming the model's ability to capture general market trends. However, in some extreme cases, deviations were noted, indicating potential areas for improvement.

4.5.3 Final Evaluation of the XGBoost Model on the Test Set

In the final test using a completely separate set (20% of the data), XGBoost achieved:

- RMSE = 20,457.68
- 95% RMSE Confidence Interval = 17,260.76 – 23,654.61

These results validate the model's robustness but also highlight limitations due to factors such as dataset size, geographic imbalance, and missing explanatory variables (e.g., economic or subjective factors).

4.6 Discussions, Interpretations, Limitations, and Final Conclusions

Subchapter 4.6 synthesizes the theoretical and practical implications of the research, offering a critical interpretation of the results, identifying study limitations, and presenting final conclusions on the performance of machine learning methods applied to real estate price evaluation.

4.6.1 Discussions and Limitations

The analysis was based on a manually collected dataset from Cluj-Napoca and tested five predictive models, from traditional techniques (linear regression, decision trees) to advanced models (Random Forest, ANN, XGBoost). Results confirmed the superiority of modern methods, particularly XGBoost. However, performance was affected by factors such as small sample size, uneven distribution across neighborhoods, and imbalance in categorical variables.

Identified limitations include:

- limited database size due to manual collection,
- geographic and categorical imbalances in apartment distribution,
- underrepresentation of certain subcategories, reducing model generalizability.

Future recommendations include automating data collection and expanding datasets to ensure balance and estimation accuracy.

4.6.2 Final Conclusions

Model comparison revealed the superior performance of XGBoost, which achieved the lowest RMSE (16,706.96) and robust statistical explanation ($R^2 = 0.823$). The comparative table of significant variables showed recurring attributes—such as area, number of rooms, floor, and neighborhood—across all models, confirming their importance in real estate evaluation.

Additionally, advanced methods demonstrated greater ability to capture complex relationships between variables, identifying relevant factors overlooked by traditional models. Using multiple algorithms in parallel can provide more robust and comprehensive predictions.

In conclusion, the study validates the usefulness of machine learning in estimating the value of residential properties, emphasizing the importance of variable selection, hyperparameter optimization, and data expansion. The research provides a solid foundation for developing automated valuation systems tailored to Romania’s real estate market.

Conclusions and Future Research Directions

This study has demonstrated that machine learning methods can significantly enhance the real estate property valuation process, offering a predictive framework that is faster, more objective, and more adaptable than traditional approaches. Building on the theoretical and methodological foundations of valuation (Chapter 1), the research supports the integration of Automated Valuation Models (AVMs) as an innovative solution in today's real estate market.

The literature review (Chapter 2) confirmed the growing academic interest in AVMs and highlighted both their advantages in mature markets and the challenges associated with implementing them in local contexts such as Romania. Chapter 3 explored the applicability of modern methods—regression, decision trees, random forests, neural networks, and XGBoost—emphasizing criteria for selecting the optimal model based on the purpose and quality of available data.

The case study conducted in Cluj-Napoca (Chapter 4) validated the hypothesis that complex models, particularly XGBoost, offer superior accuracy in estimating residential prices. The XGBoost model achieved the lowest RMSE and MAE values, confirming its superiority over the other analyzed methods. However, limitations such as the small dataset size and imbalances in categorical variables require caution when generalizing the results.

The conclusions support the use of a hybrid framework: combining artificial intelligence with human expertise in the valuation process. Proposed future research directions include expanding the study to various regional and temporal datasets, integrating macroeconomic factors, and exploring emerging technologies such as IoT or blockchain. Additionally, future studies should focus on model interpretability and the influence of sustainability on real estate values.

Overall, this dissertation provides a relevant contribution both theoretically and practically, confirming the real potential of machine learning in redefining real estate valuation standards.