## 1. Abstract

Accurate housing price prediction is a critical tool for real estate investors, urban planners, policy makers, and financial institutions (Chen et al., 2017; Glaeser and Nathanson, 2017). This project explores the use of machine learning models to forecast Q3 2024 housing prices across European Union (EU) regions, using a structured dataset comprising quarterly housing indices, macroeconomic indicators, and demographic attributes. The aim is to evaluate and compare various regression models to identify the most accurate and generalizable approach for predicting property values.

The dataset includes housing price indices from Q1 2020 to Q3 2024, along with relevant features such as GDP growth rate, average income, unemployment rate, population growth, and climate zone (Eurostat, 2023). After extensive exploratory data analysis (EDA), several engineered features were introduced to enhance predictive power, including interaction terms and ratio-based indicators. Six machine learning models were implemented: Linear Regression, Random Forest, Gradient Boosting, XGBoost, Support Vector Regression (SVR), and K-Nearest Neighbors (KNN) (Pedregosa et al., 2011; Chen and Guestrin, 2016).

Hyperparameter tuning and cross-validation were applied to optimize model performance, with metrics such as R², Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) used for evaluation. Results show that KNN and SVR outperformed traditional models, achieving R² scores of 0.9970 and 0.9932 respectively, with minimal prediction error. Feature importance analysis highlighted the impact of income, urbanization, and past housing trends on pricing (Lundberg and Lee, 2017).

The findings demonstrate that incorporating both economic indicators and historical price trends significantly improves model accuracy, and that non-linear models offer superior performance in this predictive context.

## 2. Introduction

The real estate sector plays a vital role in economic development, financial markets, and social well-being. Understanding and forecasting housing prices has become increasingly important in the context of market volatility, urban expansion, and post-pandemic economic shifts across the European Union (Gyourko et al., 2021; European Commission, 2022). Accurate prediction of housing prices supports informed decision-making for investors, developers, policy makers, and homebuyers alike. With the growing availability of structured data and the evolution of predictive modeling techniques, machine learning (ML) offers a powerful solution for addressing this complex challenge (Kaufmann and Steinmetz, 2020).

Traditional statistical models, such as linear regression, have been widely used in housing price prediction but often fall short in capturing non-linear relationships and interactions among features (Selim, 2009). Housing prices are influenced by a broad set of factors, including economic indicators (GDP growth, unemployment), demographic trends (urbanization, population growth), and regional attributes such as climate zone (Eurostat, 2023). Moreover, temporal trends embedded in historical housing prices provide additional predictive value, which can be overlooked by simplistic models.

This project leverages machine learning to forecast housing prices for Q3 2024 across EU cities and regions. The goal is not only to build predictive models but to compare a range of algorithms, assess their performance, and understand which features contribute most to price variability. By combining feature engineering, model tuning, and visual diagnostics, the project aims to identify a robust, generalizable solution to housing price forecasting in a dynamic, multi-dimensional environment.

The introduction of historical quarterly housing data as predictive features further enriches the model input space, offering a unique opportunity to capture temporal patterns. The ultimate objective is to deliver a practical, data-driven approach to housing price prediction that is both interpretable and accurate.

## 3. Business Understanding

### 3.1 Objectives

The primary objective of this project is to forecast housing prices across European Union (EU) regions for Q3 2024 using machine learning techniques applied to structured economic, demographic, and historical housing data. The project aims to:

* Evaluate and compare multiple regression models on predictive accuracy
* Identify the most suitable model for generalization and interpretability
* Analyze key factors influencing housing prices across different EU regions
* Use model explainability techniques to support data-driven insights

Success is defined as achieving a high predictive accuracy — ideally **R² ≥ 0.95** — alongside low MAE and RMSE scores, and generating interpretable outputs to inform both academic and applied audiences (Pedregosa et al., 2011; Lundberg and Lee, 2017).

### 3.2 Problem Definition

Housing affordability and regional price volatility have become critical issues in Europe, influenced by macroeconomic forces, urbanization, and supply-demand dynamics (European Commission, 2022). Traditional models, such as linear regression, lack the capacity to account for complex, non-linear patterns and temporal dependencies that are increasingly relevant in real estate forecasting (Selim, 2009).

The core problem addressed by this project is how to build accurate and interpretable predictive models for EU housing prices — accounting not only for cross-sectional features like income and GDP but also for **historical quarterly trends** in housing prices from **Q1 2020 to Q2 2024**.

### 3.3 Scope

This project focuses exclusively on structured tabular data covering EU cities and regions. The dataset includes economic indicators such as GDP growth, unemployment rate, average income, population growth, and urbanization, along with housing price indices for each quarter from 2020 to 2024 (Eurostat, 2023). Climate zones and region classifications are included as categorical variables.

The predictive target is **Q3 2024 housing prices**. The scope excludes external unstructured data sources (e.g., satellite imagery, sentiment analysis, or web scraping). Models are limited to supervised learning techniques using Python-based tools.

### 3.4 Project Plan

The project follows a structured machine learning workflow inspired by CRISP-DM:

1. **Data Collection & Understanding**: Load and explore the EU housing dataset, perform statistical summaries, and visualize trends.
2. **Data Preparation**: Clean data, encode categorical features, scale numeric values, and engineer new features (e.g., ratios and interactions).
3. **Modeling**: Train and tune six models — Linear Regression, Random Forest, Gradient Boosting, XGBoost, SVR, and KNN — using GridSearchCV for optimization.
4. **Evaluation**: Use R², MAE, and RMSE to assess performance. Visualize predictions and residuals.
5. **Comparison & Insights**: Determine the best-performing model and identify key predictive features driving housing prices.
6. **Conclusion**: Summarize findings and propose future work, including time-series modeling or integration of external data sources.

By combining time-aware feature engineering with non-linear models and explainable AI techniques, this project aims to deliver a robust, interpretable solution for housing price forecasting in the EU (Gyourko et al., 2021; Kaufmann and Steinmetz, 2020).

## 4. Technologies & Tools Used

### 4.1 Models and Machine Learning Algorithms

This project applied a diverse set of supervised regression algorithms to predict housing prices for Q3 2024. The models were selected to balance performance, interpretability, and algorithmic variety, ensuring that linear, non-linear, and instance-based approaches were fairly represented.

* **Linear Regression** was used as a baseline model due to its simplicity and interpretability, but its performance was limited by its linear assumptions (Selim, 2009).
* **Random Forest Regressor** and **Gradient Boosting Regressor** were employed as ensemble tree-based models capable of capturing non-linear relationships and feature interactions without requiring feature scaling. These models are known for their robustness and strong performance on structured datasets (Pedregosa et al., 2011).
* **XGBoost Regressor** (Extreme Gradient Boosting) was included for its efficiency and scalability in handling complex feature interactions and sparse data (Chen and Guestrin, 2016).
* **Support Vector Regression (SVR)** was used as a kernel-based model with the radial basis function (RBF) kernel to capture high-dimensional patterns in the data (Kaufmann and Steinmetz, 2020).
* **K-Nearest Neighbors (KNN) Regressor** served as a non-parametric baseline model based on similarity between observations in the feature space.

Each model was trained and evaluated using a consistent workflow, including train/test splitting, hyperparameter tuning via GridSearchCV, and performance assessment using R², MAE, and RMSE metrics.

### 4.2 Libraries and Tools

The implementation was carried out in Python using a variety of open-source libraries commonly used in data science and machine learning:

* **Data manipulation and preprocessing**:
  + pandas for structured data handling
  + numpy for numerical operations
  + sklearn.preprocessing for scaling and encoding features
* **Machine learning modeling**:
  + scikit-learn for regression algorithms, pipeline workflows, cross-validation, and performance metrics (Pedregosa et al., 2011)
  + xgboost for the XGBoost regressor (Chen and Guestrin, 2016)
* **Hyperparameter tuning and validation**:
  + GridSearchCV from sklearn.model\_selection to perform exhaustive grid search over model parameters
* **Visualization and interpretability**:
  + matplotlib and seaborn for visual analytics such as correlation matrices, boxplots, and residual plots
* **Development environment**:
  + All analysis and modeling were conducted in **Jupyter Notebook**, which provided an interactive and modular development environment ideal for data experimentation, visualization, and reproducible workflows.

This combination of models and tools enabled a robust and flexible approach to data exploration, model development, and result interpretation.

## 5. Data Understanding & Preparation

### 5.1 Dataset Overview

The dataset used in this project was obtained from publicly available **Eurostat** sources and structured into a tabular format representing regional housing markets across the European Union (Eurostat, 2023). It includes **310 records** (observations) across **29 features**, covering the period from **Q1 2020 to Q3 2024**. The dataset contains quarterly housing price indices for each city or region, along with associated economic and demographic features.

The key variable to be predicted is the **Q3 2024 housing price index**, treated as a continuous numerical target for regression modeling. Input features include:

* **Macroeconomic indicators**: GDP growth rate, unemployment rate, average income, interest rate
* **Demographic attributes**: population growth rate, urbanization rate
* **Geographic/categorical fields**: country, city, region, climate zone
* **Time-series fields**: quarterly housing prices from Q1 2020 to Q2 2024

The structured format of the dataset enabled feature-level manipulation, model-ready preprocessing, and integration of domain-driven engineered features.