UNIVERSIDAD FRANCISCO DE VITORIA

HIGHER POLYTECHNIC SCHOOL

BACHELOR'S DEGREE IN MATHEMATICAL ENGINEERING

FINAL YEAR PROJECT

ENGINEERING MODULE

Regression Models and Machine Learning for Predicting Attractive Cities for Talent

Jaime Mateo López

May, 2026

CALIFICACIÓN DEL PROYECTO FINAL DE GRADO

CUALITATIVA:	
NUMÉRICA:	

Conforme Presidente:	Conforme Secretario:
Fdo.:	Fdo.:

Conforme Vocal:	Conforme Vocal:	Conforme Vocal:
Fdo.:	Fdo.:	Fdo.:

Lugar y fecha: Pozuelo de Alarcón, a ____ de _____ de 202__

"What is the city but the people?" – ***William Shakespeare***

"El talento es la base de la prosperidad de una ciudad, ganarlo significa crecer y multiplicar el desarrollo humano, perderlo conduce a la decadencia y la desaparición." – ***José Antonio Ondiviela García***

To my mother, for being my unwavering guide, for her boundless sacrifice, and for teaching me that with hard work, anything is possible.

To my girlfriend, Blanca, for being my safe haven on stressful days, for her endless patience, and for believing in me even when I doubted myself.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, José Antonio Ondiviela García, for all the help, patience and guidance he has given me. His advice has been essential in enabling me to carry out this project.

I would also like to thank all the lecturers I have had during my degree. Their teaching and knowledge have formed the essential foundation for my education and for the completion of this project.

I would also like to thank my fellow students and friends for making my time at university so much more enjoyable and for the mutual support we have given each other in the face of the challenges we have encountered along the way.

Finally, I would like to express my heartfelt thanks to my family and my girlfriend, Blanca, who have unconditionally supported, encouraged and helped me, and given me the strength I needed to keep going and achieve this goal.

Resumen

Este Proyecto de Fin de Grado aborda el análisis cuantitativo y la predicción del atractivo urbano de 175 ciudades globales frente a la creciente competitividad por el talento y la inversión. Utilizando una base de datos de indicadores socioeconómicos proveniente del Observatorio UFV de Ciudades Atractivas para el Talento, la memoria detalla el diseño e implementación de un Sistema de Soporte a la Decisión (DSS) fundamentado en técnicas de Ingeniería Matemática. Siguiendo la metodología CRISP-DM, el trabajo abarca desde la depuración y segmentación no supervisada mediante *K-Means*, hasta el entrenamiento comparativo de modelos predictivos (Regresión Lineal, *Random Forest, XGBoost* y Redes Neuronales) para estimar el atractivo, el magnetismo y la rentabilidad urbana. Además, se integra Inteligencia Artificial Explicable (SHAP) para conocer los motores causales de las predicciones. El documento detalla la experimentación técnica, el desarrollo de un simulador de proyecciones y pruebas de estrés (*What-If analysis*) y evalúa el impacto ético, legal y social de la solución, garantizando un marco científico transparente que sirva como herramienta estratégica de planificación pública e inversión.

Palabras clave: Atractivo Urbano, Aprendizaje Automático, Modelado Predictivo, Interpretabilidad Algorítmica, Sistema de Soporte a la Decisión (DSS), Simulación de Escenarios

Abstract

This Bachelor's Thesis addresses the quantitative analysis and prediction of urban attractiveness across 175 global cities in the face of increasing competition for talent and investment. Using a socioeconomic indicators database from the *Observatorio UFV de Ciudades Atractivas para el Talento,* the report details the design and implementation of Decision Support System (DSS) based on Mathematical Engineering techniques. Following the CRISP-DM methodology, the project covers everything from data clearing and unsupervised segmentation via K-Means to the comparative training of predictive models (Linear Regression, Random Forest, XGBoost and Neural Networks) to estimate urban attractiveness, magnetism and profitability. Additionally, Explainable Artificial Intelligence (SHAP) is integrated to identify the causal drivers behind the predictions. The document details the technical experimentation, the development of a projection and stress-testing simulator (*What-If analysis*) and evaluates the ethical, legal and social impact of the solution, ensuring a transparent scientific framework to serve as a strategic tool for public planning and investment.

Keywords: Urban Attractiveness, Machine Learning, Predictive Modeling, Algorithmic Interpretability, Decision Support System (DSS), Scenario Simulation

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Meaning
Observatory	Observatorio UFV de Ciudades Atractivas para el Talento
PFG	Final Year Project
RMSE	Root Mean Squared Error o Error Cuadrático Medio
MAE	Mean Absolute Error o Error Absoluto Medio
R^2	Coefficient of determination
SHAP	SHapley Additive exPlanations
DSS	Decision Support System
CRISP-DM	Cross-Industry Standard Process for Data Mining
RGDP	General Data Protection Regulation
LOPDGDD	Organic Law 3/2018 on the Protection of Personal Data and the Safeguarding of Digital Rights
PCA	Principal Component Analysis
ReLU	Ground Linear Unit
Adam	*Adaptive Moment Estimation*
XAI	Explainable Artificial Intelligence

1. INTRODUCTION

1.1. CONTEXT

Cities have taken center stage in recent years on the global stage, having become the world's primary economic and social driving force. Today, more than half of the population lives in urban areas (United Nations, 2019). This leads cities to compete with one another to attract talent and large companies, with the aim of achieving greater prosperity and a better quality of life for their citizens. It is in this context that the concept of urban attractiveness is introduced, understood as the ability of cities to attract and retain talent (Ondiviela, 2020a). Talent signifies prosperity, whilst its absence signifies decline; therefore, assessing a city's attractiveness is vital in the current global climate of fierce competition for talent.

It is essential to bear in mind that urban attractiveness depends on many factors that directly and indirectly affect urban development. These include, notably, the economic sphere, the social sphere, quality of life, employment opportunities, connectivity, culture, etc. All these factors together form what we know as a city, and the study of them has attracted the attention of numerous organizations (de Oliveira et al., 2021).

In this context, the analysis of urban attractiveness can be broken down into different dimensions that allow for a more precise understanding of the complexity of the phenomenon. In particular, this study focuses on three objective variables: Attractiveness, Magnetism and Profitability. The first, attractiveness, refers to a city's overall capacity to attract and retain talent based on factors such as quality of life or career opportunities. The second, magnetism, is more specifically associated with a city's ability to capture the emotional aspect of a person's decision to move there; in other words, it is closely linked to its external image and its physical and cultural characteristics. Finally, profitability refers to the city's performance in terms of the quality of services it provides to its residents, compared with its economic potential, taking into account aspects such as business activity, investment opportunities, the economic return associated with its development, average wages in the city and the cost of living. These three dimensions enable a more comprehensive and structured analysis of urban attractiveness (Ondiviela, 2020a).

Various studies and research projects have been carried out to assess urban attractiveness. Of particular note is the *Observatorio UFV de Ciudades Atractivas para el Talento* (Ondiviela, 2025) – hereinafter referred to as the Observatory – which studies 175 cities worldwide and utilizes over a hundred indicators, having conducted research for several years and already published six annual reports. The parameters studied include most of the factors affecting a city (economic, environmental, social, cultural, mobility, employment, safety, education,

connectivity, gastronomy, quality of life, etc.), thus providing an extremely rich and comprehensive database. For this reason, the Observatory's database has been chosen as the primary source for this project.

1.2. Statement of the Problem

The study of urban attractiveness involves analyzing a wide range of indicators, as mentioned earlier. Cities are not defined by a single, simple variable, but are the result of the interaction of many interrelated indicators, some of which are very diverse. The Observatory's database contains over a hundred indicators covering all areas affecting the city, many of which are interrelated, making it difficult to identify the factors that most influence urban attractiveness. Many studies are based on the analysis of city rankings, but this approach has limitations when it comes to analyzing variables individually and does not allow for the identification of complex patterns or the making of robust predictions of future behavior (de Oliveira et al., 2021). However, the Observatory's dataset focuses on the value of the data itself, rather than its position in the ranking, which greatly facilitates the design of models capable of making predictions.

Nowadays, there is a wide variety of mathematical models that enable complex analyses of many variables simultaneously. Of particular note are regression models and machine learning techniques, which are capable of detecting non-linear patterns in the variables and thus making highly reliable predictions in high-dimensional contexts (Mehrabi et al., 2021). This is where the need arises to determine whether historical data can be used to predict urban attractiveness, and whether it is possible to build models capable of doing so.

This project will focus on developing regression and machine learning models to predict the urban appeal of cities and, in turn, compare the models to determine which is best in each respect. To this end, the Observatory's database will be used, training the models with data from 2020–2024 and validating them with data from 2025.

1.3. Justification

The current global context is fueling fierce competition amongst cities to attract and retain talent (OECD, 2023a). It is therefore crucial to understand what makes a city attractive, with a twofold aim:

- To help city managers carry out effective urban planning and economic and social development, thereby enabling them to make the best possible public policy decisions, using scenario simulators (what-if analysis) (Wexler et al., 2019) to assess the impact of future policies before they are implemented.
- To help talented citizens discover which cities might be the best places for them to realize their full potential.

This understanding is now possible thanks to the data collected by the Observatory, which has a very extensive and comprehensive database, containing numerous indicators and several years' worth of data, enabling the advanced analyses required for this purpose.

In this context, the project aims to quantitatively assess the urban appeal of cities through the development of mathematical models and machine learning techniques. By analyzing the set of indicators, the aim is to assess the predictive power of the various models developed and to identify complex patterns within the dataset. In this way, the work carried out not only enables an analysis of the relationship between the different variables, but also explores the possibility of making predictions based on historical data (OECD, 2023b).

1.4. Scope & Impact

This project focuses on the analysis of 175 cities based on data collected by the Observatory, comprising over a hundred indicators covering the years 2020–2025. Using this data, various mathematical and machine learning models will be developed with the aim of analyzing the relationship between the available variables, assessing their predictive power regarding urban attractiveness, and comparing these models with one another.

This project does not seek to construct a definitive model for urban forecasting or to fully explain how cities function, but rather to assess the potential of the models developed to identify relevant patterns within the dataset. The work aims to contribute to the analysis of urban attractiveness from a quantitative perspective, providing tools that enable a better understanding of the factors affecting cities' competitiveness for talent (OECD, 2023a).

1.5. Motivation

1.5.1. Professional Motivation

This project is driven by professional motivation, as it enables me to apply the knowledge acquired during my degree in Mathematical Engineering, such as data analysis, statistics, mathematical modelling and machine learning. This work bridges the gap between my degree and the real world, by applying these mathematical tools to real-world problems and working with large datasets within a complete analytical cycle (end-to-end pipeline) (Kreuzberger et al., 2023), from statistical cleansing through to deployment.

1.5.2. Personal Motivation

This work is also driven by a strong personal motivation. Having had direct contact with the database over the last two years sparked my interest in analysing it in greater depth. Furthermore, it also heightened my interest in understanding how cities function and in determining, both mathematically and empirically, which factors make them more attractive.

My Final Year Project (FYP) thus presents an opportunity to delve even deeper into the work carried out during my collaboration with the Observatory.

2. STATE OF THE ART

2.1. THEORETICAL FRAMEWORK

2.1.1. Urban Appeal & the Competitiveness of Cities

Urban attractiveness refers to a city's ability to attract and retain talent, investment, economic activity and prosperity, as well as to offer a high quality of life (Glaeser, 2020). In a globalized context, cities compete with one another to position themselves as favorable environments for both professionals and businesses, making this a key concept in urban development (OECD, 2023a).

It is a multidimensional phenomenon, influenced by economic, social, cultural and environmental factors. Owing to this complexity, the analysis of urban attractiveness requires the use of data analysis tools that enable the study of the relationship between multiple variables in order to uncover latent structures or clusters amongst them and to understand the factors influencing their evolution (Jiang, 2025).

2.1.2. Multivariate Data Analysis

Multivariate data analysis focuses on the simultaneous study of multiple variables with the aim of identifying relationships, patterns and structures within a dataset (Hastie et al., 2021). In complex problems, such as the analysis of urban attractiveness, variables do not act in isolation, but rather exhibit interdependencies that must be considered collectively.

This type of analysis enables us to understand how different factors influence a target variable and facilitates the construction of more accurate predictive models. In this context, it is particularly relevant when working with high-dimensional datasets, where the number of variables is high and their interaction can be decisive in the results obtained (Tan et al., 2019).

2.1.3. Regression Models

Regression models are statistical tools used to analyze the relationship between a dependent variable and one or more independent variables. Their main objective is to estimate the value

of the target variable based on the information provided by the other variables in the dataset (Hastie et al., 2021).

In the context of this project, regression models enable the prediction of indicators such as cities' attractiveness, magnetism and profitability based on multiple factors. These models form a fundamental basis for predictive analysis, as they enable both the making of estimates and the interpretation of the influence of variables on the results obtained, serving in practice as a robust baseline model against which to compare more complex algorithmic alternatives.

2.1.4. Supervised Machine Learning

Supervised machine learning is a branch of machine learning in which models are trained using labelled data, that is, datasets in which the value of the target variable is known in advance. Based on this information, the model learns the relationship between the input variables and the output, with the aim of being able to make predictions about new data.

In regression problems, such as the one addressed in this project, supervised learning enables the estimation of continuous values based on multiple explanatory variables. This approach is particularly useful when working with large volumes of data and complex relationships between variables, as it allows patterns to be identified that are not easily detected using traditional methods; in this context, the high performance of tree-based ensemble learning models is particularly noteworthy (Sarker, 2021).

2.1.5. Tree-Based Models: Random Forest & XGBoost

Decision tree-based models are widely used in regression and classification problems due to their ability to model non-linear relationships and handle large numbers of variables. These models divide the data space into regions using decision rules, which allows them to capture complex interactions between variables.

Notable examples of these models include Random Forest and XGBoost. Random Forest is an ensemble method that combines multiple decision trees built on different subsets of the data, thereby improving generalization ability and reducing overfitting (Mohamadi et al., 2025). XGBoost, meanwhile, is a boosting-based algorithm that constructs trees sequentially, correcting the errors of previous models and, in many cases, achieving high predictive performance (Wiens et al., 2025).

Both models are particularly well-suited to analyzing a dataset with multiple variables and complex relationships, as is the case with urban attractiveness; they are therefore key to the development of this project, whilst also offering the added advantage of natively calculating the importance of each variable (Feature Importance) in the predictive process.

2.1.6. Model Evaluation Metrics

The evaluation of predictive models is a fundamental step in determining their quality and ability to generalize. To this end, quantitative metrics are used to measure the error made by the model when comparing predicted values with actual values (Miller et al., 2024).

In regression problems, some of the most commonly used metrics are the root mean square error (RMSE), which penalizes large errors more heavily; the mean absolute error (MAE), which measures the average error in predictions; and the coefficient of determination (R^2), which indicates the proportion of data variability explained by the model. These metrics enable different models to be compared objectively and the one with the best performance to be selected, particularly when evaluating them on future datasets to ensure their robustness against temporal fluctuations (Chicco et al., 2021).

2.1.7. Model Interpretability

Model interpretability refers to the ability to understand how input variables influence the predictions made. In the context of machine learning models, it is not only important to obtain good results, but also to understand which factors are having the greatest impact on the target variable, using methodologies such as SHAP (SHapley Additive exPlanations) values, which are based on cooperative game theory (Hettikankanamage et al., 2025).

This analysis enables the relevance of each variable within the model to be identified and facilitates the drawing of conclusions that are useful from a practical perspective. In problems such as the study of urban attractiveness, interpretability is particularly relevant, as it allows us to understand which factors have the greatest influence and adds value to the analysis beyond the prediction itself (Givisis et al., 2025).

2.2. Background & Related Work

In recent years, the analysis of cities' performance and development has taken on great significance, giving rise to numerous studies focused on assessing urban attractiveness using specific indicators. These studies enable us to understand the city as a complex system in which economic, social, environmental and cultural factors interplay.

In this vein, various studies have examined urban attractiveness using quantitative indicators. These studies focus on identifying the variables that influence urban competitiveness and cities' ability to attract talent and investment (Ondiviela, 2020a). However, in many cases, these approaches are descriptive in nature and do not incorporate predictive models that would enable the future evolution of these indicators to be estimated.

Furthermore, the development of machine learning techniques has enabled progress towards more complex models capable of analyzing large volumes of urban data. In this context, some studies have applied machine learning algorithms to the analysis of cities, demonstrating their usefulness in modelling urban phenomena and improving the accuracy

of predictions (Martínez-Fernández et al., 2020). These approaches make it possible to capture non-linear relationships between variables and to work with heterogeneous datasets, which is particularly relevant in urban environments, where these supervised techniques are frequently complemented by unsupervised learning algorithms (such as clustering) to uncover latent city typologies without prior biases.

With regard to the specific models used in this type of problem, the Random Forest algorithm has been widely used in predictive tasks due to its robustness and generalization ability. This model, based on a combination of multiple decision trees, enables stable results to be obtained and reduces the risk of overfitting. In various studies, Random Forest has proven effective in predicting complex variables in real-world contexts (Biau et al., 2016), making it a suitable tool for urban data analysis.

Meanwhile, the XGBoost algorithm has gained popularity in recent years due to its high performance in regression and classification problems. This method, based on boosting techniques, builds models sequentially by correcting errors from previous iterations, thereby improving the accuracy of predictions. Various studies have shown that XGBoost can outperform other traditional models in certain scenarios, particularly when working with complex, high-dimensional data (Chen & Guestrin, 2016).

Furthermore, various studies have compared different machine learning models with the aim of determining which offers the best performance in predictive tasks. These studies show that there is no universally optimal model, as performance depends largely on the characteristics of the dataset and the problem being analyzed. However, models such as Random Forest and XGBoost tend to stand out for their ability to handle complex relationships and produce accurate results, often outperforming more complex architectures such as deep learning, which tend to suffer from overfitting when applied to smaller datasets (Zhang et al., 2021).

Finally, in recent years, the analysis of the interpretability of machine learning models has taken on particular importance. Beyond the accuracy of predictions, it is essential to understand how variables influence the results obtained. In this context, techniques such as SHAP values (Lundberg & Lee, 2017) enable the contribution of each variable to the model's predictions to be analyzed, facilitating the interpretation of results and the drawing of relevant conclusions.

2.3. Summary Table of Methodological Sources

Below is a summary table setting out the main studies analyzed, highlighting their approach, objectives and relevance to this project.

Source title	Analysis	Aim of the study	Relationship and contribution
Beyond SmartCities: How to create an attractive city for talented citizens (Ondiviela, 2020a).	Quantitative analysis of urban indicators relating to competitiveness and attractiveness.	To identify the factors that influence the attractiveness of cities.	It serves as the conceptual basis for the project. It highlights the importance of the indicators, although it does not include any predictions.
A comparison of machine learning methods for the prediction of traffic speed in urban places (Martínez-Fernández et al., 2020).	Application of machine learning models to complex urban data.	Modelling urban phenomena and improving predictive capabilities.	Explain why machine learning techniques are used in the analysis of cities.
Measuring urban poverty using multisource data and a random forest algorithm: A case study in Guangzhou (Biau et al., 2016).	Using the Random Forest model to predict real-world data.	To assess the ability of Random Forest to model complex relationships.	Explain why you chose the Random Forest model for the project.
An XGBoost-SHAP approach to quantifying morphological impact on urban flooding susceptibility (Chen & Guestrin, 2016).	Application of XGBoost to regression problems involving complex data.	Improving the accuracy of predictions through boosting.	It reinforces the use of XGBoost as an advanced, high-performance model.
Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis (Zhang et al., 2021).	Comparison of different machine learning models using metrics.	To analyze the relative performance of various predictive models.	Explain the rationale behind comparing models in the project and the absence of a universally optimal model.
The use of ICTs and income distribution in Brazil: A machine learning explanation using SHAP values (Lundberg & Lee, 2017).	Interpretability analysis using SHAP values.	Explain the contribution of the variables to the predictions.	It enables the analysis of the significance of variables in the project to be justified.

Tabla 2.1 Summary of relevant sources for the project

3. OBJECTIVES

This section sets out the overall objective and the specific objectives of this project. It also explains the validation methods that will be used to ensure that these objectives are achieved.

3.1. GENERAL OBJETCTIVE

The main aim of this project is to analyze cities from a quantitative perspective in order to predict urban attractiveness using various indicators, thereby understanding which factors have the greatest influence on making a city more attractive.

3.2. LIST OF SPECIFIC OBJECTIVES

1. Explore the dataset to understand the variables that influence urban attractiveness.
2. Prepare the database to obtain a clean dataset suitable for building models, resolving issues of structural multicollinearity through analysis of variance.
3. Develop regression and machine learning models to predict urban attractiveness.
4. Evaluate the performance of the models to measure their predictive capacity.
5. Compare the models developed to identify the best one for predicting urban attractiveness.
6. Interpret the results obtained to analyze the importance of the variables and the most influential factors.
7. Apply unsupervised learning and dimensionality reduction techniques to identify clusters with similar latent characteristics.
8. Design a strategic scenario simulator (what-if analysis) to serve as a decision support system for assessing the impact of future urban policies.

3.3. Validation Methods

The validation process for the methods used in this study aims to assess the predictive ability and reliability of the models developed. To this end, the following steps will be carried out:

1. Splitting the data to validate the results.

The data will be split into two sets: train and test. Data from the years 2020–2024 will be used to train the developed models, and data from the year 2025 will then be used to test these models. This division allows for the reproduction of a future prediction and avoids the use of future information during training (Joeres et al., 2025).

2. Evaluation of models using mathematical metrics.

To assess whether the models built are functioning correctly, various mathematical metrics will be used. These metrics will be the Root Mean Squared Error (RMSE), which penalizes large errors in the result, and the Mean Absolute Error (MAE), which shows the model's average error. A lower error for both metrics indicates a better model. R^2 will also be used, as it measures the model's explanatory power. The closer R^2 is to 1, the better the model fits the data (Miller et al., 2024; Chicco et al., 2021).

3. Comparison of the models developed using mathematical metrics.

The various models developed will be compared using the mathematical metrics mentioned above. One of the criteria for selecting the best model will be to identify the one with the lowest error and the highest R^2.

4. Interpretability analysis of the models.

In addition to basing positive results on the metrics obtained, an interpretability analysis of the developed models will also be carried out. This will enable an analysis of the importance of variables in urban attractiveness and the most influential factors through the extraction of global feature importance and local analysis using SHAP values (Hettikankanamage et al., 2025).

In summary, the validation methods defined will focus on objectively evaluating the performance of the models developed, as well as comparing their predictive capacity. This will make it possible to select the most suitable model and ensure the consistency and reliability of the results obtained.

4. METHODOLOGY & WORK PLAN

4.1. METHODOLOGY

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology has been adopted for the development of this final year project. This methodology provides a robust, structured and iterative framework, and is the standard for projects in data science, machine learning and applied mathematical engineering.

Justification and suitability for the project

The choice of the CRISP-DM methodology over other traditional development methodologies is justified by the analytical, exploratory and quantitative nature of this work. As this is a project based on the construction of mathematical models and predictive algorithms, the development does not follow a linear flow but requires an iterative approach. CRISP-DM allows us to go back and adjust previous phases if, for example, a model's performance during the mathematical development phase requires a different transformation of the input variables.

The suitability of this methodology for the project lies in the fact that it ensures scientific rigor throughout the entire data lifecycle. Its application guarantees that the main objective of the work is addressed in a systematic and structured manner.

The implementation of the six phases of the CRISP-DM methodology has been adapted to the specific modelling requirements of this project as follows:

1. Understanding the Problem: Initial phase in which the mathematical and analytical problem has been conceptualized. The three target variables (Attractiveness, Magnetism and Profitability) have been defined and the general structure of the database provided by the Observatory has been analyzed
2. Understanding the Data: Initial statistical exploration of the dataset (comprising 175 cities, over a hundred indicators and data spanning 2020 to 2025). In this phase, the distributions, variance and linear and non-linear correlations between all available indicators have been analyzed.
3. Data Preparation: A critical stage of feature engineering. This includes cleaning the database, resolving structural multicollinearity issues using statistical techniques,

handling missing values and normalizing variables, ensuring that the data fed into the algorithms is mathematically consistent and free from scale bias.

4. Modelling: A key phase of mathematical engineering, divided into two approaches:
    - Unsupervised learning: The use of clustering techniques to discover latent structures in the data and group them.
    - Supervised learning: Training regression models and advanced tree-based algorithms (Random Forest and XGBoost) to predict the target variables.
5. Evaluation: Rigorous and objective analysis of the performance of the developed models. This was carried out by validating the models against the 2025 dataset, using error metrics (RMSE and MAE) and fit metrics (R^2). Furthermore, model interpretability (XAI) was integrated by extracting SHAP values to assess the quantitative impact of each variable.
6. Implementation: In the context of this applied research, the implementation phase involved the development of a strategic scenario simulator (What-If analysis). This simulator applies the predictive models that have already been trained and evaluated to project the mathematical impact of variations in the indicators on future urban attractiveness (2026).

4.2. TECHNOLOGIES

This project has been developed using the Python programming language (Python Software Foundation, 2024) as its primary foundation. This language was chosen due to its versatility and the extensive ecosystem of libraries specializing in data science and machine learning. This technology has enabled the seamless and consistent integration of all phases of the CRISP-DM methodology employed, from database pre-processing through to predictive modelling and the final visualization of the results.

Jupyter Notebook (Kluyver et al., 2016) was used as the development environment. This is an interactive environment that is ideal for exploratory data analysis, as it allows code blocks to be executed independently, results to be visualized immediately, and technical documentation to be interwoven with algorithmic execution. This flexibility has ensured an agile workflow, facilitating debugging and continuous experimentation with the models.

During the data preparation and processing phase for the Observatory, the Pandas (McKinney, 2010) and NumPy (Harris et al., 2020) libraries were used in conjunction. Pandas has been the key tool for loading, initial processing, cleaning and manipulating the data, enabling the effective handling of missing values and the restructuring of the set of indicators. NumPy, for its part, has provided the necessary support for matrix calculations and the high-level mathematical operations required for the statistical transformations of the variables and the calculation of percentage changes.

For the core mathematical engineering stage, the Scikit-learn library (Pedregosa et al., 2011) formed the cornerstone of the modelling. It was used to implement unsupervised learning

algorithms (using K-Means to identify urban tribes) and supervised learning algorithms (basic regression models and Random Forest). Scikit-learn has also managed the entire validation framework for the project, including the splitting of the dataset (train/test) and the calculation of statistical performance metrics (RMSE, MAE and R^2).

To complement the predictive analysis and maximize performance across profitability metrics, the XGBoost library (Chen & Guestrin, 2016) has been integrated. This environment provides a highly optimized implementation of the eXtreme Gradient Boosting algorithm, demonstrating superior efficiency compared to traditional models when detecting complex patterns in high-dimensional databases, such as the urban indicators studied.

To ensure the mathematical interpretability of the models developed, the SHAP library (Lundberg & Lee, 2017) has been implemented. This tool, based on cooperative game theory, has enabled the calculation of the exact marginal impact of each indicator on a city's attractiveness, providing the model with the interpretability required for decision-making.

The visualization of results and the qualitative validation of the projections were carried out by combining Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021). Matplotlib has served as the base plotting engine, whilst Seaborn has provided a higher level of abstraction that has simplified the creation of complex statistical visualizations, such as correlation matrices, comparative bar charts for predictive scenarios and multivariate scatter plots.

Finally, with regard to computational resources, the entire development, modelling and simulation cycle has been carried out efficiently using local CPU-based processing. Given that tools such as XGBoost and Scikit-learn are highly parallelized and optimized, it has not been necessary to utilize GPU processing or cloud computing infrastructure for the volume of data handled in this project.

Table 4.1 below provides a summary of the technologies used in the development of the project, together with a brief description of their main purpose within it.

Technology	Purpose
Python	Primary programming language for all project development and algorithmic analysis
Jupyter Notebook	An interactive environment used for programming, iterative debugging and code execution
Pandas y NumPy	Structuring and computational libraries used for data cleaning and linear algebra
Scikit-learn	The main machine learning library used for K-Means, Random Forest, error metrics and data separation
XGBoost	Integrated advanced Gradient Boosting framework for optimizing complex regression predictions
SHAP	A library of analytical interpretability metrics used to quantify the importance of variables in models
Matplotlib y Seaborn	Graphical tools used to visually represent distributions, clusters and simulation results

Tabla 4.1 Used Technologies

4.3. Project Development Plan

The project has been structured into three main phases, each focusing on a critical stage in the data lifecycle and predictive modelling. These phases are aligned with the objectives of building a robust analytical foundation, developing competitive predictive models, verifying the results obtained, and validating their usefulness by simulating complex urban scenarios using the CRISP-DM methodology, integrated within a modern MLOps framework (Kreuzberger et al., 2023).

4.3.1. Phase 1 – Exploratory Analysis & Data Architecture

The first stage of the project focuses on defining the problem and preparing the data ecosystem. A thorough cleaning of the historical dataset of urban indicators is carried out, addressing the handling of missing values, the normalization of scales and temporal filtering. This phase includes an in-depth exploratory data analysis to identify correlations between socio-economic variables and the dimensions of urban success defined by Ondiviela (2020a): Magnetism, Profitability and Attractiveness. This variable engineering process, which is essential for the optimal performance of supervised learning algorithms (Sarker, 2021), is based on advanced pre-processing and feature selection techniques (Kuhn & Johnson, 2019).

4.3.2. Phase 2 – Development of Competitive Machine Learning Models

This phase focuses on the design and development of the project's mathematical core. Various architectures are implemented and compared, covering both supervised and unsupervised learning techniques, in order to extract maximum value from urban data:

- Unsupervised learning: Clustering techniques, specifically the K-Means algorithm, are applied to discover underlying structures. This allows cities to be segmented into clusters with similar behavior without relying on prior labels.
- Statistical Baseline Models: Multiple Linear Regression is implemented to establish a statistical baseline and assess the linearity and direct correlation of the predictor variables.
- Ensemble Models: Random Forest algorithms and Gradient Boosting architectures such as XGBoost are applied, due to their efficiency with tabular data, to capture complex non-linear and hierarchical patterns.
- Deep Learning: An Artificial Neural Network (MLP) architecture is developed to assess its generalization ability compared to traditional techniques.
- Implementation and Validation: Development is carried out entirely in Python (Python Software Foundation, 2024) using the Scikit-learn library (Pedregosa et al., 2011). A data splitting strategy (Train/Test Split) is employed to avoid information filtering (Joeres et al., 2025), using R^2 as the primary evaluation metric (Chicco et al., 2021).

4.3.3. Phase 3 – Implementation & Analysis of Results

In the third phase, the selected model (Miller et al., 2024) is applied to generate strategic intelligence using a proprietary simulation engine. Rather than a static analysis, three types of experimental projections for cities are designed:

1. Inertial Trend Projection: Simulation of a city's natural growth based on the extrapolation of its recent historical trends in investment and development.
2. Macroeconomic Shock Simulation: Assessment of urban resilience through the artificial alteration of purchasing power and inflation variables, measuring the impact on the city's profitability.
3. Sensitivity Analysis to Social Crises: A univariate assessment aimed at measuring the vulnerability of the 'Attractiveness' and 'Magnetism' dimensions in the face of a deterioration in critical indicators of security and social cohesion.

4.4. Work Plan

This project has been planned to accommodate an estimated workload of 300 hours of staff time, spread over a 13-week implementation schedule. This plan was designed to cover, in a balanced manner, both the research and theoretical grounding phase and the technical implementation and academic writing. Below is a detailed breakdown of the effort required for each phase and its corresponding representation in a Gantt chart, distinguishing between the initial proposal presented in the preliminary draft and the final implementation.

4.4.1. Initial Planning

The initial planning, defined during the preliminary design phase, envisaged a general structure of work phases with their respective estimated timelines to meet the proposed objectives:

- Literature review and conceptual framework (3 weeks, ~60 hours): Study of the state of the art relating to urban attractiveness and in-depth understanding of each of the indicators.
- Data preparation and exploratory analysis (2 weeks, ~50 hours): Thorough cleaning of the database, normalization of variables, correlation between indicators and detection of outliers.
- Model design and training (3 weeks, ~75 hours): Implementation of the initial predictive models (Random Forest, XGBoost and evaluation of a possible basic neural network) and selection of hyperparameters.
- Evaluation and interpretation of results (2 weeks, ~50 hours): Validation of the models using 2025 data and analysis of the weighting of each indicator.
- Writing the final report (2–3 weeks, ~65 hours): Detailed documentation of all the final work carried out and drawing of conclusions.
- Final review and verification of results (1 week, ~15 hours): Quality control of the document and verification of the consistency of the models.

Figure 4.1 shows the Gantt chart for this initial plan.

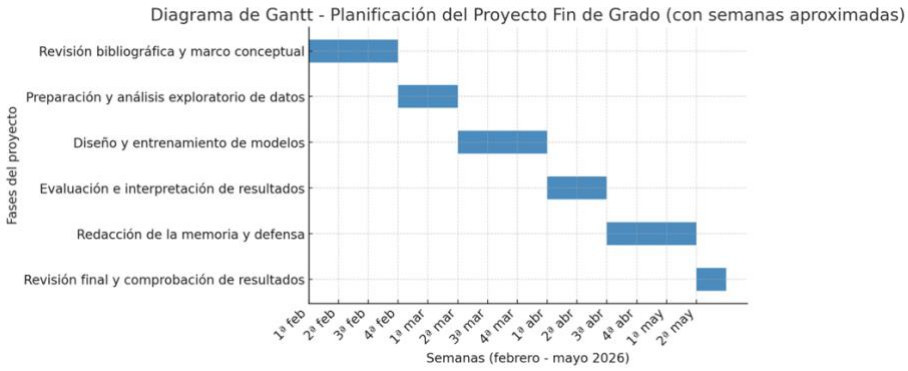

Figura 4.1 Initial Project Planning (Gantt Chart)

4.4.2. Final Planning

During the actual implementation of the project, the planning remained in line with the original deadlines, completing within the planned 13 weeks, but was detailed and broken down into much more specific analytical and predictive subtasks to reflect the advanced technical maturity:

- Data architecture and unsupervised learning: Cleaning the dataset and implementing the K-Means algorithm for the discovery of urban clusters.
- Base and ensemble models: Implementation of Linear Regression as a baseline, followed by the development and advanced optimization of Random Forest and XGBoost.
- Deep learning (RNA): Design, training and fine-tuning of the Artificial Neural Network (MLP) architecture.
- Prospective simulation (2026 scenarios): Development of a proprietary simulation engine and execution of specific scenarios in the cities, such as hypothetical financial crises.

Figure 4.2 shows the Gantt chart for the final plan, which reflects this specialization of tasks whilst maintaining temporal consistency with the initial plan.

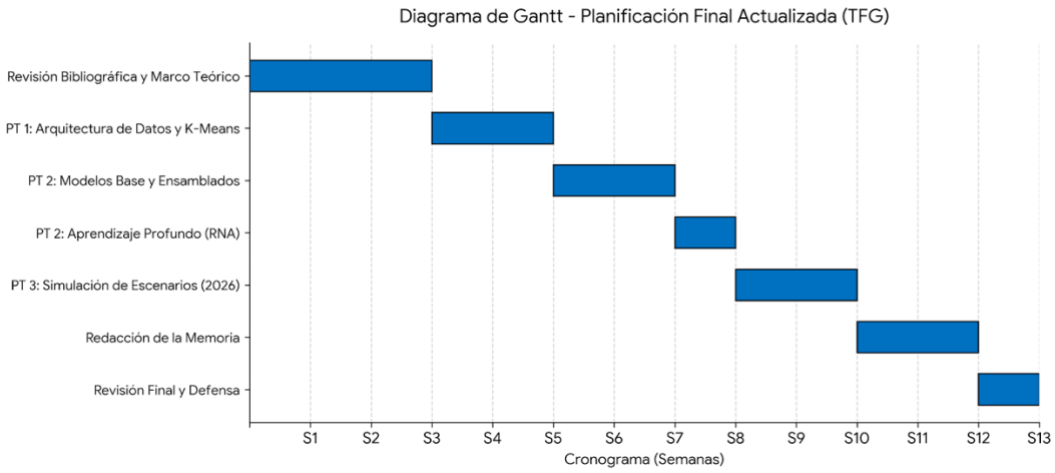

Figura 4.2 Final Project Plan (Gantt Chart)

4.4.3. Deviations from the Initial Plan

The main differences between the initial strategic planning and the final technical implementation of the project are explained below:

- Incorporation of unsupervised learning (K-Means): In the initial approach, data analysis was conceived as purely descriptive. However, during the development of Phase 1, the need to implement clustering techniques to segment cities into groups with similar behavior was identified. This additional task allowed for a deeper understanding of the data before proceeding to the prediction phase.

- Stratification of predictive modelling: Initially, model training was planned as a single block. In the final phase, this block was broken down into three levels of mathematical complexity (linear regression, ensemble models and neural networks) to enable a rigorous technical comparison of the balance between interpretability and accuracy.
- Development of the prospective simulation engine: What was defined in the preliminary draft as a standard validation using data from 2025 evolved into the development of a dynamic scenario simulator for 2026. The creation of specific 'shock' scenarios, such as a security or economic crisis, required greater attention to the design of the variable injection logic in phase 3.
- Drafting of the iterative report: As with the initial planning phase, drafting remained an ongoing task. However, the overlap with the technical phases was intensified in order to document the model architectures and the metrics obtained immediately, thereby preventing the loss of complex technical details.
- Hyperparameter optimization and cross-validation: This sub-task required slightly more time than anticipated due to the computational complexity of the Gradient Boosting (XGBoost) models and the need to ensure that the neural network did not suffer from overfitting.

These adjustments were deemed necessary to enhance the scientific rigor of the work and to transform a static prediction model into a robust and functional urban simulation tool.

4.5. RESOURCES

The development and training of the models presented in this project were carried out in a technical environment with the following hardware and software specifications:

- Operating system: macOS Ventura (version 13.6.7).
- Computer: Apple MacBook Air (Retina, 13-inch, 2019).
- Processor: Dual-core Intel Core i5 at 1.6 GHz.
- RAM: 8 GB 2133 MHz LPDDR3.
- Execution environment: Jupyter Notebooks running in a virtual environment.
- Language: Python, version 3.12.
- Main libraries: Scikit-learn (v. 1.3), XGBoost (v. 2.0), Pandas, NumPy and SHAP, Matplotlib and Seaborn.

Furthermore, all the source code developed – including dataset cleaning, exploratory data analysis, the training of the various architectures (Linear Regression, Random Forest, XGBoost and Neural Network) and the simulation engine for the prospective scenarios for 2026 – has been documented and organized in a structured manner. The code is available for review and execution.

With regard to human resources, this project – from the literature review through to the final draft – has been carried out entirely and independently by the author, with the ongoing support and supervision of the academic supervisor, José Antonio Ondiviela García, to ensure methodological rigor.

4.6. Costs

This section presents the financial estimate for the project as if it had been carried out within a professional technology consultancy environment. A distinction is made between operating expenses (OPEX), which are recurring during development, and capital expenditure (CAPEX), relating to the depreciation or acquisition of fixed assets.

To calculate rates, an industrial margin of 40% has been applied to staff costs, 30% to supplies and miscellaneous expenses, and 20% to hardware. VAT at the current rate of 21% will be added to the total budget.

4.6.1. Personal

Although the project was carried out by a single person, for the purposes of cost estimation, the hours spent have been broken down according to the standard professional roles in the data science sector (CNAE 62) in the Community of Madrid. The hourly cost includes gross salary, social security contributions (34 per cent) and the company's profit margin (40 per cent). A total of 300 hours has been estimated, which is typical for projects of this scale.

Professional Profile	Estimated Hours	Net Cost per Hour (excluding VAT)	Total Net Cost
Project Manager	40	48€	1920€
Data Scientist	140	40€	5600€
Data Engineer	80	35€	2800€
QA and Documentation Analyst	40	30€	1200€
Total	300		11520€

Tabla 4.2 Breakdown of Staff Costs

4.6.2. Software & Data Licenses

The development of this project stands out for its high cost-effectiveness and low budgetary impact in terms of third-party tools. It has not involved any direct financial outlay on licences, as all the computational resources and software tools used have been open source.

Description	Cost (€)
Open-source licenses (Python, XGBoost, …)	0€
Data set (Academic license granted by the Observatory)	0€
Total	0€

Tabla 4.3 Licenses and Subscriptions

The entire technology stack used is available free of charge for both academic and commercial use. However, the dataset has been provided free of charge by the Observatory as part of university research, representing a significant saving compared with the purchase of private datasets.

4.6.3. Supplies & Miscellaneous

The proportionate share of utilities consumed during the months the project was underway is charged. Electricity is estimated based on the consumption of the development team's equipment, and the internet connection is charged on a pro rata basis at a standard fibre-optic rate.

Description	Cost (€)
Electricity supply (proportional share)	120€
Internet connection (pro rata)	100€
Contingencies and office supplies	80€
Total	300€

Tabla 4.4 Costs of Supplies and Materials

4.6.4. Hardware & Infrastructure

The runtime and training environment for the machine learning models has been deployed entirely on-premises, meaning no costs have been incurred for cloud infrastructure (AWS, Google Cloud or Azure). Development requires a computer with sufficient processing power to train the assembled algorithms.

Description	Cost (€)
Workstation / Development laptop (16 GB RAM)	1200€
Total	1200€

Tabla 4.5 Hardware Equipment (CAPEX)

4.6.5. Budget Summary

Below is a summary table showing the total cost of implementing the project, breaking down the base costs, profit margins and the relevant taxes.

Category	Type	Base Cost	Margin	Taxes (21%)	Total
Staff	OPEX	8228€	3292€	2419€	13939€
Licensess	OPEX	0€	0€	0€	0€
Supplies and Miscellaneous	OPEX	230€	70€	63€	363€
Hardware	CAPEX	1000€	200€	252€	1452€
Total OPEX		8458€	3362€	2482€	14302€
Total CAPEX		1000€	200€	252€	1452€
Final Budget		9458€	3562€	2734€	15754€

Tabla 4.6 General Overview of Project Costs

4.7. CONSTRAINTS AND LIMITATIONS

During the technical development of the project, a number of factors have been identified that limit the scope of the study, starting with the complexity involved in the initial phase of consolidating the data architecture. A key challenge lay in the fact that the source data was fragmented across multiple separate Excel files for each year, which required preliminary work to clean, transform and technically merge all the databases into a single consolidated repository in order to carry out the project and apply the machine learning algorithms consistently. Added to this structural constraint is the variability in the availability and consistency of socio-economic indicators, where the lack of updates or the presence of missing values in the historical dataset meant that considerable effort had to be devoted to variable engineering to ensure the integrity of the patterns learnt by the models. Furthermore, a methodological limitation arises concerning the interpretability of deep learning and ensemble architectures. When using models such as artificial neural networks or XGBoost, there is a technical trade-off between accuracy and interpretability, making it difficult to isolate the direct impact of each indicator compared with simpler models such as linear regression. Finally, it should be borne in mind that cities are complex and constantly evolving environments. Consequently, the projections made for the year 2026 have inherent limitations when it comes to purely qualitative variables or exceptional and unforeseeable events (such as COVID-19 in 2020). As the algorithms are based solely on learning from past data, they are unable to anticipate this sort of sudden event that could drastically alter a city's appeal.

5. DEVELOPMENT OF THE TECHNICAL SOLUTION

This section presents the technical procedure followed to achieve the objectives set for this project, omitting the source code (which is provided via a GitHub link within the document itself) in order to focus on design decisions, the analytical methodology, the decisions made, and the architecture of the models. The description is organized sequentially, following the three main phases of work defined in the development plan.

5.1. PHASE-1

The main objective of this first stage was to transform a raw and fragmented dataset into a structured, high-quality, unified analytical ecosystem ready to feed the machine learning algorithms that have been developed (McKinney, 2022). The process was divided into four main phases: database consolidation and architecture, data cleaning and preprocessing, exploratory analysis, and unsupervised learning.

5.1.1. Consolidation and Architecture of the Master Database

As a preliminary and essential step prior to any analysis, the data architecture was consolidated. The initial data provided by the Observatory as part of the research was highly fragmented, scattered across multiple independent Excel files for each year and divided into several sheets within each year. The first major technical task was to design a unification process to create a single “source of truth” (McKinney, 2022). For this unified table, the approach taken was to implement a new “Year” column, which allowed the cities’ historical data to be integrated as a coherent time series, facilitating the detection of patterns over the years and ensuring data integrity (Hyndman & Athanasopoulos, 2021).

During this process, a significant issue was observed: the availability of certain indicators varied by year. The Observatory has continuously refined its measurement methodology, introducing new metrics and discarding others to keep the data as reliable and up-to-date as possible. This is why some columns ended up with certain empty rows (NaNs).

5.1.2. Data Cleaning and Preprocessing

Once the final database to be used—which contained 1,050 rows (175 cities × 6 years of data) and 114 columns—was consolidated, we proceeded to the key phase of data cleaning and algorithmic preparation. The first step involved addressing data leakage. It was determined that certain variables in the dataset (MAG RK, RK PROFIT, and ATTRACTIVENESS RK) directly represented the rankings or results of the target indicators. To prevent the algorithms from "cheating" by memorizing the answer, these three variables were immediately removed. At the same time, a format conversion was performed, transforming text-type variables into numerical format. This was done by replacing the comma (",") with a period (".”). Afterward, out of the 103 text-type columns we had, we found four columns that could not be converted to numeric type: "City," "Country," "GDPPROXIMITY," and "GTCI-COUNTRY." To resolve this, we separated the identifying columns ("Country" and "City," along with "ID" and "Year") and transformed the other two ("GDPPROXIMITY" and "GTCI-COUNTRY") by removing special characters and correcting errors.

The next step was to handle missing values (NaNs). To do this, we checked how many NaNs the database contained (12,064) and identified the columns with the highest number of NaNs, highlighting the top three: Liveable Cities (875 NaNs), Avg. % of Cloudy Daylight Hours/Month (777 NaNs), and Avg. % of Sunny Daylight Hours/Month, 777 NaNs. A strict completeness threshold was set: any variable with more than 30% NaNs would be discarded. This criterion led to the elimination of 19 indicators, whose predictive power was deemed insufficient and whose mass imputation would have introduced unacceptable artificial bias into the models. The 19 indicators are as follows:

- *CRIME Index* (700 nulls)
- *Safe Index* (350 nulls)
- *Avg. % of Sunny Dayligh Hours/Month* (777 nulls)
- *Avg. % of Cloudy Dayligh Hours/Month* (777 nulls)
- *Cost Food USD/m* (525 nulls)
- *Music* (350 nulls)
- *STREETART/10k inh* (525 nulls)
- *CITY BLOGGERS* (525 nulls)
- *Best Cities* (700 nulls)
- *IMD* (525 nulls)
- *Crea Index* (525 nulls)
- *Liveable Cities* (875 nulls)
- *World Best Cities (REV)* (700 nulls)
- *Numbeo Quality of Life* (700 nulls)
- *CITIES NOR* (525 nulls)
- *UNIV BUZ&ECO NOR* (350 nulls)
- *GFI NOR* (525 nulls)
- *TRAFFIC INDEX NOR* (700 nulls)
- *TOMTOM NOR* (350 nulls)

After removing those columns, the number of missing values in the dataset was checked again and had dropped to 1,060. For the remaining variables with fewer missing values, a historical trend imputation technique (Forward & Backward Fill) was applied. Taking

advantage of the time-series structure of the database, records were grouped by city, and missing values were filled in based on the values for that same city in adjacent years, thereby preserving the temporal consistency of each record individually and resulting in a dataset with zero missing values (Hyndman & Athansopoulos, 2021).

Variable scaling was then performed. Since the dataset combines metrics with disparate numerical scales, a Min-Max Scaler transformation was applied, compressing all numerical predictors into a standardized mathematical range between 0 and 10. This procedure is a fundamental requirement for distance-sensitive algorithms such as K-Means or Neural Networks (Géron, 2022).

Finally, to ensure the mathematical robustness of the training, we conducted an analysis of collinearity and dependence. First, we generated a Pearson correlation matrix and removed variables with a correlation coefficient greater than 0.90, as they provided redundant information. Next, an iterative process was carried out to calculate the Variance Inflation Factor (VIF) (Bruce et al., 2020). This statistical indicator quantifies the severity of complex multicollinearity; that is, it measures the extent to which a predictor variable can be mathematically explained by the combination of the other variables in the model. Applying this technique is essential, since retaining overlapping indicators artificially inflates the variance, creates instability in the algorithms, and distorts the actual weight that each metric has on the final prediction. For this project, a maximum penalty threshold of 10.0 was set, leading to the systematic elimination of 17 variables affected by severe multicollinearity, which are:

- *GOV-NOR*
- *SOC ACTIVITY NOR*
- *ETHICS NOR*
- *GENDER*
- *Tolerance*
- *Equality*
- *DYNAMISM INDEX*
- *STRATEGY INDEX*
- *EmpDev NOR*
- *EMPLOYABILITY INDEX*
- *CONNECTED CITY NOR*
- *Public Health Expenditure PPP USD NOR*
- *ENV SUS NOR*
- *CULTURE INDEX NOR*
- *SAFETY NOR*
- *SERVICES PERFORMANCE Index NOR*
- *Monthly Wages (Avg)*

The completion of this entire data cleaning phase resulted in a final, optimized dataset consisting of the four identifying variables, the three target metrics, and a core set of 60 predictor variables that were fully cleaned, scaled, and mathematically independent.

5.1.3. Exploratory Data Analysis

After cleaning and standardizing the database, we proceeded to conduct an exploratory data analysis (EDA). According to Tukey (1977), EDA is a fundamental step for maximizing understanding of the data structure, detecting outliers, and testing underlying hypotheses before applying complex predictive models. In this project, EDA focused on two dimensions: descriptive analysis and interdependence analysis.

- Descriptive statistical analysis

The distribution, measures of central tendency, and dispersion of the predictor variables were evaluated using the statistical description function (describe). This step allowed for an empirical verification of the success of the Min-Max Scaler transformation, confirming that all numerical predictors were constrained within the mathematical range [0, 10], thereby eliminating magnitude biases. Furthermore, the analysis of standard deviations and percentiles yielded relevant structural conclusions about the sample. For example, economic indicators such as Net Purchasing Power and the Cost of Living Index had the lowest standard deviations in the dataset (1.61 and 1.74, respectively), indicating greater relative homogeneity among the cities. In contrast, variables related to tourist appeal (such as Destination NOR) showed strong positive skewness: although their maximum value is 10, their median is only 0.56, which mathematically demonstrates that the status of “major tourist destination” is concentrated in a very small group of cities.

- Correlation and Interdependence Study

After filtering the variables using VIF, the core of the exploratory analysis consisted of quantifying the relationship between the 60 final predictor variables and the project’s three target metrics: Attractiveness, Magnetism, and Profitability. To this end, a selective correlation heat map was generated (Hunter, 2007; Waskom, 2021), designed to isolate the individual impact of each indicator on the desired outcomes.

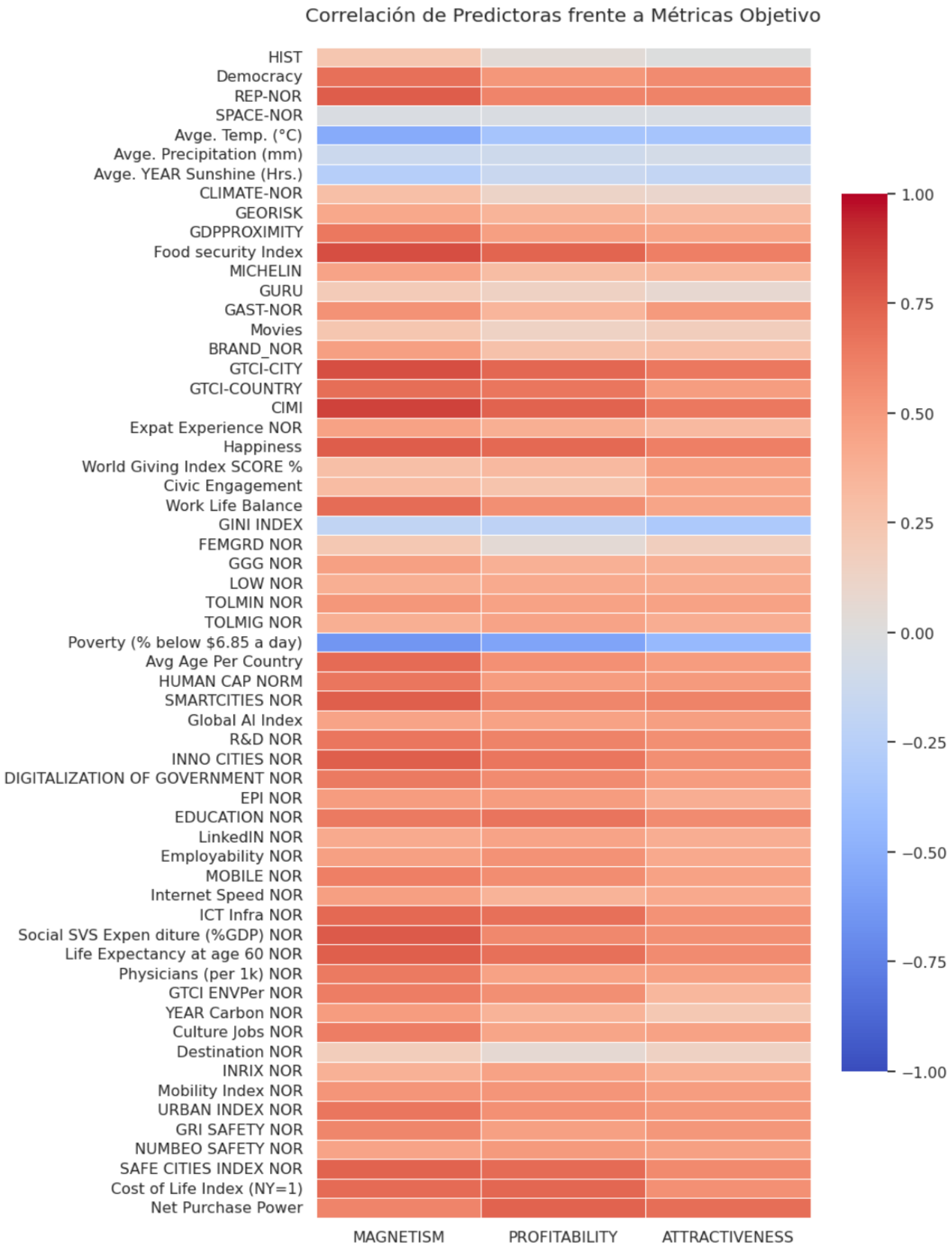

Figura 5.1 Heat map showing the correlation between predictor variables and target metrics.

Looking at *Figure 5.1*, the following empirical conclusions can be drawn, which underpin the logic of subsequent machine learning models:

1. Inverse or negative correlations (Dark blue): The graph reveals that indicators such as poverty (Poverty (% below $6.85 a day)), average temperature (Avg. Temp. (°C)), and inequality (GINI INDEX) are represented by the darkest shades of blue on the map. This empirically demonstrates a strong negative linear correlation: high levels of inequality, poverty, and extreme weather act as the primary drag on or detractors from urban attractiveness.
2. Predictors with a high positive impact (Deep red): Conversely, a strong positive linear correlation is observed in macroeconomic, resilience, and protection variables, such as Net Purchasing Power, the Cities in Motion Index (CIMI), the Food Security Index, and the Safe Cities Index. This indicates that an increase in citizens' purchasing power, overall competitiveness, the guarantee of quality basic supplies, and public safety directly drive the city's indices upward.
3. Absence of a Linear Relationship (White/Neutral Tones): It is noteworthy that variables related to visitor flow and the physical environment—such as the volume as a Tourist Destination (Destination NOR) or the availability of Natural Spaces (SPACE NOR)—show virtually no coloration in relation to the target metrics. This demonstrates that, in isolation and in a strictly linear sense, being a vacation destination or having more green spaces is not a determining factor in ensuring a city's business profitability or macroeconomic appeal.

This analysis of linear interdependence serves as a fundamental diagnostic basis. However, given that urban ecosystems are multidimensional, the subsequent use of nonlinear machine learning models (Random Forest, XGBoost, and Neural Networks) is fully justified, as these models are capable of unraveling complex relationships that Pearson's correlation coefficient fails to capture.

5.1.4. Dataset Partitioning: Temporal Validation Strategy

The first critical step in predictive modeling is splitting the data into training and validation sets for the algorithms. To evaluate the models' generalization ability in a real-world decision-making scenario, we opted for a temporal partition rather than a standard random split.

Following the recommended methodologies for time series (Hyndman & Athanasopoulos, 2021), in contexts where a time component exists, it is essential not to use future information to predict the past, thereby avoiding temporal data leakage. For this reason, the split was structured using a 5-year historical period compared to the most recent year as a blind validation set:

- Training and validation set (Train): We used the consolidated historical data for the cities covering the 2020–2024 period. This dataset contains a total of 875 observations (corresponding to the 175 cities over the 5-year period). It represents

83.3% of the total data volume and is where the algorithms identify patterns and adjust their internal weights.

- Blind Test Set (Test): The year 2025 was reserved as the evaluation set. This block consists of 175 observations (the remaining 16.7%). Since this data is "future" relative to the training data, it allows us to verify whether the model has truly learned to predict urban evolution or has simply memorized the historical data.

This ratio (approx. 83/17) is methodologically robust, as it provides a training dataset broad enough to capture post-pandemic variability in cities while maintaining a significant test set to validate the final results (Géron, 2022).

5.1.5. Mathematics Assessment Criteria and Metrics

Given that the algorithmic nature of this project involves the regression of continuous values (on a standardized scale from 0 to 10), the performance, robustness, and generalization ability of the designed models will be evaluated on the blind test set (year 2025). To establish an objective comparison between the baseline linear regression and the subsequent advanced algorithms, four standard statistical metrics used in the data science industry were employed (Kuhn & Johnson, 2019):

1. Mean Squared Error (MSE) and its Root Mean Square Error (RMSE): MSE calculates the average of the squared errors between the predicted and actual values. Its derivative, RMSE (the square root of MSE), returns the metric to the original scale of the problem. The main mathematical characteristic of this metric is that it penalizes large errors quadratically. In the context of this project (predicting profitability, magnetism, and urban attractiveness), minimizing the RMSE is a critical criterion, since a severe error in predicting a city's strategic positioning could lead to highly flawed corporate investment decisions or public policy choices.
2. Mean Absolute Error (MAE): Provides the average magnitude of errors in predictions by measuring the absolute difference between the prediction and the actual value, without considering the direction of the error. Unlike RMSE, MAE assigns a linear weight to all errors, providing a robust measure against outliers and one that is easily interpretable on a scale from 0 to 10.
3. Coefficient of Determination (Classic R^2): This metric quantifies the proportion of the variance in the dependent variable (the target metrics) that is mathematically predictable based on the independent variables. Its value ranges from 0 to 1 (or is negative if the model performs worse than always predicting the mean). An R^2 value close to 1.0 indicates that the model has successfully captured the underlying structure that defines the attractiveness or profitability of cities.
4. Adjusted Coefficient of Determination (Adjusted R^2): To ensure the scientific rigor of the analysis, an additional evaluation function was implemented algorithmically: the adjusted coefficient of determination. While the classic R^2 tends to increase artificially whenever new predictor variables are added (regardless of whether they provide real value or not), the adjusted R^2 introduces a mathematical penalty based on the

model's complexity (Bruce et al., 2020). The use of this metric is essential in this study due to the high dimensionality of the dataset, which has a core of 60 predictor variables. In models with a high number of dimensions, there is a real risk of "false precision," where a high R^2 could mask an overfitted model. Therefore, adjusted R^2 acts as a detector of overfitting: its value will only increase if the variables included improve the model more than would be expected by pure chance. The difference between R^2 and adjusted R^2 will allow us to determine whether the complexity of the solution is justified by its actual predictive power.

The validation of the developed algorithms will not be based on a single, isolated indicator, but rather on a comprehensive analysis of these four metrics. While RMSE and MAE will quantify the operational margin of error in the predictions (the tool's actual risk), the combination of R^2 and Adjusted R^2 will validate the structural quality of the learning and its resistance to overfitting. This multidimensional evaluation framework constitutes the methodological standard that will be systematically applied during the results phase to determine which predictive architecture is best suited to form the final strategic simulator.

5.1.6. Unsupervised Learning: Clustering and Dimension Reduction

As the final step of Phase 1, an unsupervised learning workflow was implemented using the Scikit-learn library (Pedregosa et al., 2011). The objective was twofold: first, to reduce the dimensional complexity of the dataset to facilitate its visualization; and second, to discover natural clusters of cities that shared similar socioeconomic profiles without relying on predefined labels.

- Principal Component Analysis (PCA): Given the high dimensionality of the final dataset (60 predictor variables), the PCA technique was applied. According to Jolliffe (2002), PCA allows a set of correlated variables to be transformed into a smaller number of uncorrelated variables called principal components, while retaining as much of the original variance as possible. In this project, the first two components (PC1 and PC2) were extracted, with the goal of projecting the urban ecosystem onto a two-dimensional Cartesian plane. To ensure interpretability and avoid treating the model as a "black box," the analysis of the loadings of the original variables on each component was methodologically established, thereby allowing the new X and Y axes generated to be named and given semantic meaning.
- Clustering Using K-Means and the Elbow Method: The K-Means clustering algorithm (MacQueen, 1967) was applied to the two-dimensional transformed space. Since this algorithm requires the user to define the number of clusters in advance, the Elbow Method was used as a validation protocol. This technique evaluates the model's inertia (the sum of the squared distances from each point to the center of its cluster) by iterating over different values of k, allowing the identification of the empirical inflection point that represents the optimal number of clusters for the global ecosystem.

This unsupervised architecture was conceived as a critical milestone in the project methodology. It provides the analytical stratification necessary to assess, after the fact,

whether the mathematically generated clusters converge with the macroeconomic classifications established by experts (Ondiviela, 2025) and will serve as a spatial reference framework for the modeling results of Phase 2.

5.2. PHASE-2

Once the data preparation and structuring were complete, Phase 2 of the project focused on the design and implementation of machine learning algorithms. The technical objective of this phase was to solve a prediction problem involving three continuous variables (Attractiveness, Magnetism, and Profitability). We opted for an independent modeling architecture, specifically a single-target regression approach. This strategic decision involved training separate algorithmic instances dedicated exclusively to each of the three metrics based on the tensor of 60 predictor variables. Working with isolated models allows us to maximize individual accuracy and transparently evaluate which features have the greatest impact on each urban dimension separately (Bramer, 2020).

To ensure a methodologically rigorous evaluation, an incremental pipeline was designed. This strategy involves sequentially implementing algorithms ranging from lower to higher mathematical complexity, starting with linear parametric models and progressing to nonlinear architectures with high predictive capacity (Chollet, 2021).

5.2.1. Multiple Linear Regression: Baseline Model and Diagnostics

The first algorithm implemented was Multiple Linear Regression, the main function of which is to establish a baseline or target performance. According to Géron (2022), this step is essential for determining the minimum accuracy threshold that complex models must exceed.

Following the Single-Target strategy, three independent instances of the Linear Regression class from Scikit-learn were developed. For each target variable (y), the model seeks to fit a hyperplane that minimizes the sum of the squared residuals. The technical process, replicated for Attractiveness, Magnetism and Profitability, followed the implementation flow outlined below:

1. Model fitting: Training using the 'X' tensor comprising 60 predictor variables and the 'y' vector corresponding to the period 2020–2024.
2. Calculation of the dimensionality penalty: Given that the model handles 60 variables (k=60) against a finite test sample (n = 175), the code incorporates a custom function to calculate the adjusted R^2. The 'Complexity Penalty' was technically defined as the arithmetic difference between the conventional R^2 and the adjusted R^2 ($R^2 – R\_adj^2$). This value allows us to quantify how much of the model's accuracy is 'artificial' or due solely to the high number of variables.
3. Diagnostic and assumption validation protocol: Unlike other 'black box' models, linear regression requires the validation of its statistical assumptions. To this end, the

technical implementation includes two visual diagnostic tools run following the prediction:

- Scatter plot (Actual vs. Predicted): To examine the linearity and bias of the model relative to the 45° identity line.
- Homoscedasticity analysis (residual distribution): An analysis of the residuals (e = y_actual – y_predicted) against the predicted values was carried out. This step is critical for verifying that the error variance is constant. A random distribution of points around the zero line confirms that the linear model is a valid approximation for the cities dataset (Hastie et al., 2021).

This comprehensive approach ensures that linear regression serves not only as a benchmark for metrics (MAE, RMSE), but also as a tool for monitoring the quality and reliability of the original data.

5.2.2. Ensemble Algorithms and Non-linearity: Random Forest & XGBoost

After establishing the baseline performance threshold using linear regression, the predictive architecture evolved towards the use of ensemble learning algorithms. The underlying theoretical rationale for this decision lies in the nature of urban ecosystems: a city's success metrics rarely follow a purely linear pattern. Socio-economic and technological variables interact with one another to form complex patterns (for example, the impact of social spending may be influenced by the level of public safety).

To capture these hidden dynamics and non-linear relationships without the need for manual mathematical transformations, two state-of-the-art decision tree-based architectures were implemented (Géron, 2022):

1. Bagging Architecture: Random Forest Regressor

The Random Forest algorithm (Breiman, 2001) works by constructing a multitude of independent decision trees during the training phase and averaging their predictions. This technique, known as Bagging (Bootstrap Aggregation), is highly effective at reducing variance and combating overfitting, a critical risk when dealing with 60 predictor variables.

From a technical perspective, following the Single-Target approach, three independent models were instantiated with the following parametric configuration, specifically designed to balance predictive power and computational cost:

- Number of estimators (n_estimators = 200): This hyperparameter defines the number of individual trees that will make up the ensemble. A value of 200 was selected to ensure that the forest is dense enough to stabilize the predictions and reduce the variance error, whilst reaching an optimal point of convergence without incurring unnecessary computational time.
- Randomization control (random_state = 42): The Random Forest algorithm introduces randomization both in the selection of rows (bootstrapping) and in the

selection of variables for each node split. Setting a 'seed' guarantees the scientific reproducibility of the experiment, ensuring that any researcher or panel member running the code obtains exactly the same partitions and results.

- Parallelization (n_jobs = -1): This allowed all available processor cores to be used to train the 200 trees simultaneously, optimizing execution times.

2. Boosting Algorithms: eXtreme Gradient Boosting (XGBoost)

As a methodological counterpart, XGBoost (Chen & Guestrin, 2016) was implemented. Unlike Random Forest, which creates independent trees, XGBoost utilizes the principle of boosting: it constructs trees sequentially, with each new tree attempting to mathematically correct the errors (the residuals) made by the previous ones. This algorithm was selected because it represents the state of the art in tabular data and features internal regularization mechanisms that penalize excessively complex models.

To control this iterative learning process, the following critical hyperparameters were tuned:

- Iterations of correction (n_estimators = 200): In this context, this indicates the maximum number of rounds of sequential correction (trees added).
- Learning rate (learning_rate = 0.1): This parameter controls the extent to which each new tree contributes to the final model. A conservative rate of 10% (0.1) was chosen, meaning that the algorithm learns 'more slowly' and more cautiously. A low learning rate, combined with an appropriate number of estimators (200), prevents the model from overreacting to data noise, drastically improving its ability to generalize to unseen data (Test 2025).

As with the base model, for each prediction generated by Random Forest and XGBoost, the MAE, RMSE, Classic R^2 and Adjusted R^2 were evaluated using the custom function detailed in previous sections.

The greatest analytical advantage of implementing the XGBoost architecture, beyond its high accuracy, lay in its algorithmic interpretability. To prevent the system from acting as a 'black box', a module for extracting feature importance was programmed.

This mechanism extracts the relative weight (as a percentage) that the algorithm assigns to each predictor when performing its node splits. Mathematically identifying the 'Top 10' metrics with the greatest influence on Attractiveness, Magnetism and Profitability constitutes the final step that will enable, in subsequent stages of the analysis, the provision of business logic and empirical justification for the urban strategic simulator (Kuhn & Johnson, 2019).

5.2.3. Deep Learning: Multilayer Perceptron Neural Network

As the final stage of the modelling process, the use of deep learning techniques was explored through the implementation of a multi-layer perceptron (MLP) neural network. The inclusion of this architecture responds to the need to test whether the extremely high density of

interactions between the 60 urban variables can be captured more efficiently using a structure inspired by biological information processing (Chollet, 2021).

Unlike linear or tree-based models, neural networks operate using layers of interconnected neurons that learn hierarchical representations of the data. For this project, a custom architecture was designed using the MLPRegressor class from Scikit-learn with the following technical parameters:

- Hidden layer topology (hidden_layer_sizes = (128,64)): An architecture with two hidden layers was configured. The first layer has 128 neurons to perform high-dimensional feature extraction, whilst the second layer, with 64 neurons, acts as an abstraction funnel prior to the output layer. This structure allows the model to learn complex patterns without incurring an excessive computational load.
- Activation function (activation = "relu"): The Rectified Linear Unit (ReLU) was used, which is the current industry standard. The ReLU function enables the network to handle the non-linearity of urban data and avoids the problem of 'gradient vanishing', facilitating faster and deeper learning (Goodfellow et al., 2016).
- Optimization algorithm (solver = "adam"): The Adam (Adaptive Moment Estimation) optimizer was selected; this is a stochastic gradient descent algorithm that automatically adjusts the learning rate during training.
- Convergence strategy (max_iter = 2000): The iteration limit was increased to 2000 to ensure that the network has sufficient scope to converge to an optimal solution and minimize the loss function without stopping prematurely.

In line with the project's experimental approach, the Neural Network was not evaluated in isolation, but in a direct 'predictive head-to-head' against the winning model from the previous phase. In other words, for each of the three target variables, the Neural Network was pitted exclusively against the traditional algorithm (linear or non-linear) that had demonstrated the best preliminary metric performance.

This cross-comparison methodology makes it possible to determine whether the increased technical complexity involved in a Neural Network translates into improved generalization ability by the year 2025 or whether, on the contrary, traditional models prove to be more robust for predicting this urban ecosystem.

5.2.4. Temporal Cross-Validation (Walk-Forward)

Once the algorithmic competition phase between the different models had concluded, a definitive 'champion model' was identified for each of the three target urban metrics. This selection was based strictly on the performance achieved on the static test set (year 2025).

However, in time-series-oriented data science, achieving a good result in a single test year is not sufficient guarantee of robustness, as the model may have been fortunate with the

specific partitioning of that data or may have been overfitted to the particular conditions of that year (Hyndman & Athanasopoulos, 2021).

To empirically validate that the top-performing models are structurally robust and consistent over time, a walk-forward validation protocol was implemented.

Unlike standard cross-validation (K-fold cross-validation), which divides the dataset into random subsets whilst ignoring the chronological order and causing a serious problem of temporal data leakage, the walk-forward technique scrupulously respects the arrow of time.

At a programming level, an iterative loop was designed to evaluate each champion model by simulating the passage of the years:

- Iteration 1: The model is trained exclusively on data from the first available year (2020) and is asked to predict the immediately following year (2021).
- Iteration 2: The memory window is expanded. The model is retrained from scratch by combining the years 2020 and 2021 and must predict the year 2022.
- Continuous expansion: This process is repeated iteratively until the final year of the time series is predicted.

For each of these 'prediction windows', the system calculated and stored the MAE and R^2. Applying this protocol to the three winning models demonstrates that their generalization ability is not an isolated phenomenon specific to the year 2025.

5.2.5. Explainable Artificial Intelligence (XAI): Analysis Using SHAP Values

As the final stage of technical development, and with the aim of ensuring complete transparency within the predictive ecosystem, an Explainable AI (XAI) module based on SHAP values was integrated.

The main technical limitation of advanced non-linear models, such as Random Forest or XGBoost, is their 'black box' nature. In these systems, the mathematical relationship between the 60 input variables and the final prediction is algorithmically opaque. To transform this model into a useful strategic decision-making tool in the urban context, it is imperative to understand not only what score the system predicts, but why it predicts it and how it weights each factor (Lundberg & Lee, 2017).

SHAP analysis is based on cooperative game theory, calculating the marginal contribution of each predictor variable to the outcome of a specific prediction (Molnar, 2020). At a programming level, the implementation in this project was dynamically adapted to the underlying architecture of each 'Champion Model":

- For decision tree-based algorithms, the specialized Tree Explainer class was used, which is optimized for navigating the node structure of the ensembles.

- For parametric models (linear regression), the Linear Explainer module was instantiated, which is designed to decompose linear coefficients with high efficiency.

The methodological design of the SHAP analysis was structured to provide answers on two complementary analytical scales:

1. Overall interpretability (Summary/Beeswarm Plots): Density visualizations were designed to illustrate the overall impact of each variable on the target dimensions (Attractiveness, Magnetism, Profitability). This tool enables us to assess whether the model has learnt coherent business logic (for example, verifying that an increase in the cost of living negatively affects the attractiveness score), or whether, on the contrary, its decisions are based on mere statistical coincidences present in the historical data, thereby ensuring the ethical and operational reliability of the simulator (Molnar, 2020).
2. Marginal effects (Dependence Plots): For the variable with the greatest weight or importance in each model, the extraction of its isolated dependence plot was programmed. Unlike the global analysis, this local level reveals the non-linear behavior of the variable, identifying saturation thresholds (points at which continuing to invest a resource no longer generates positive returns for the city).

The extraction of these values and their subsequent graphical representations go beyond mere statistical validation, laying a transparent and explainable foundation for the scenario analysis (What-If) to be addressed in the next chapter.

5.3. PHASE-3

The final phase of the technical development involved the creation of a simulation environment capable of transforming statistical predictions into strategic and business insights. This stage goes beyond conventional predictive modelling to provide a predictive analytics tool or Decision Support System (DSS). The aim is to enable urban planners or investment analysts to assess a city's resilience and the hypothetical impact of specific policies.

To meet the various analytical needs, the simulation was structured in code using three modules of increasing complexity: historical validation, future trend projection and adverse scenario simulation.

5.3.1. Retrospective Evaluation and Historical Consistency

Before proceeding to simulate future or hypothetical scenarios, it was considered methodologically essential to demonstrate, both visually and empirically, the reliability of the system using known data. To this end, an initial back testing module (individualized back testing) focused on trajectory analysis was implemented.

The technical objective of this module is to validate the local temporal sensitivity of the models; in other words, to verify whether the algorithms are capable of tracking the specific evolution of a single city over time, capturing its turning points year on year (Hyndman & Athanasopoulos, 2021).

At the programming level, a dynamic script was designed that takes a target city as an input parameter. The automated execution flow is as follows:

1. Extraction of the historical vector: The system isolates the data matrix corresponding to the city entered by the user and sorts it chronologically for the period 2020–2025. The 60 unaltered predictor variables are extracted for each year.
2. Inference using champion models: The historical dataset is run through the three winning models consolidated in Phase 2. Specifically, for each target variable (Attractiveness, Magnetism and Profitability), the system dynamically invokes the corresponding champion algorithm to generate a continuous retrospective prediction curve across the analyzed period.
3. Generation of the visual comparison: Finally, using the Matplotlib and Seaborn libraries, the code renders a panel of three superimposed subplots. In each subplot, the actual value recorded in the original dataset is plotted against the value predicted by the algorithm.

This functionality provides the user with an essential visual audit tool. Only by verifying that the models' error is minimal when reconstructing the past of a selected city can the mathematical confidence required to run the simulations in the subsequent modules be justified.

5.3.2. Projecting Future Trends (Natural Extrapolation)

Once the historical consistency of the model has been validated, the simulator incorporates a second module designed for short-term probabilistic forecasting and strategic decision support (Provost & Fawcett, 2013). This component does not seek to simulate a sudden disruptive event, but rather to project a city's competitive positioning over a near-future horizon (for example, the year 2026) on the assumption that it will maintain its current trajectory of development.

The distinctive value of this module lies in its ability to generate future (synthetic) data vectors based strictly on the recent historical behavior of the selected city. At a technical level, the algorithmic workflow operates as follows:

1. Calculation of temporal inertia (trend): The algorithm isolates the most recent historical data for the city entered by the user (during code development, this was restricted to the post-pandemic period from 2022 onwards to avoid noise from the health crisis). It then mathematically calculates the average rate of change, or first derivative, of each of the 60 predictor variables—a standard technique for extracting trends from time series (Box et al., 2015).

2. Generation of the synthetic future vector: This 'natural trend' is added arithmetically to the data vector for the most recent consolidated year (2025). To ensure the structural integrity of the data and prevent the mathematical models from extrapolating to impossible values, the code implements a clipping function that ensures no projected variable exceeds the maximum value of 10 or falls below 0, thereby respecting the original normalization scale of the feature space (Kuhn & Johnson, 2013).
3. Growth inference and quantification: The new probability vector is processed through the champion models. The system calculates the percentage difference (Δ%) between the base-year score and the future projection for Attractiveness, Magnetism and Profitability.
4. Visualization: Expected growth or decline trends are automatically plotted on a divergent bar chart, enabling investors or urban planners to immediately visualize the areas of expansion or contraction within the city.

This functionality provides the DSS with a powerful forecasting engine. During the project's technical validation phases, this module was calibrated using cities experiencing rapid urban development and massive capital injections into infrastructure, enabling an assessment of how machine learning algorithms interpret and integrate atypical growth patterns into the urban ecosystem.

5.3.3. Simulation of Adverse Scenarios

To complete the design of the DSS, a third analytical module was integrated, aimed at assessing urban resilience. Unlike natural inertia projections, this component is based on 'what-if' analysis and sensitivity testing (stress testing). Its purpose is to quantify a city's vulnerability to severe disruptions or unexpected macroeconomic events (Saltelli et al., 2008).

The algorithmic engine of this module operates under the ceteris paribus principle (all other things being equal), a fundamental technique in the validation and exploration of predictive models known in the technical literature as scenario profiling or Ceteris Paribus Profiles (Biecek & Burzykowski, 2021). This methodology allows the analyst to isolate a single predictor variable within the urban fabric, introduce a crisis or 'stress' scenario, and observe the knock-on effect on the predictions for Attractiveness, Magnetism and Profitability without interference from the other characteristics.

At the level of code architecture, the system supports two types of mathematical disturbance, which gives the tool a high degree of versatility:

1. Relative stress injection (percentage/linear decrease): The user defines a specific decrease (for example, subtracting 2.5 points from a metric). The system extracts the data vector from the most recent base year available, applies the subtraction to the target variable and, simultaneously, executes a clipping function (numpy.clip) to prevent the new value from falling below the lower bound of 0 on the normalized scale.

2. Absolute stress injection (critical threshold forcing): Alternatively, the system allows the city's historical value to be ignored and the variable to be directly forced to a predefined critical level (for example, collapsing an index to a value of 1.0/10), simulating a total disruption in that sector regardless of the starting point.

Once the modified feature vector (X_future) has been generated, it is subjected to inference using the champion models. The code calculates the net and percentage variation relative to the undisturbed scenario, automatically rendering a comparative bar chart that illustrates the extent of the crisis's impact.

During the methodological validation phase, this module was configured to evaluate two macro-critical scenarios, which will be analyzed in depth in the results section:

- Economic collapse: Using the first method (relative stress), a massive fall in purchasing power (Net Purchase Power) was simulated. This scenario was designed to assess cities with strong links to financial capital and evaluate the consequences this would have for such a city.
- Deterioration of the urban environment: Using the second method (absolute stress), a drastic fall in the Safe Cities Index was simulated. This univariate analysis was designed to measure the fragility of cities whose main pillar is quality of life and to assess how severely they would be affected in the event of a loss of safety.

The implementation of these functions confirms the successful transition from a purely descriptive/predictive model to a prescriptive artificial intelligence framework, fulfilling the primary objective of providing actionable tools for the design of urban strategies (Borgonovo, 2017).

6. RESULTS

This chapter presents, describes, and rigorously interprets the results obtained following the application of the methodology outlined in the previous phases of the project. In line with the principles of applied research in Data Science, the purpose of this section extends beyond the mere quantitative presentation of metrics or graphs; it aims to provide a critical analysis of their significance, their impact within the urban context, and their alignment with the specific objectives established in Chapter 3.

In order to ensure full transparency and reproducibility of the experiments detailed herein, the complete source code (Jupyter Notebook) is hosted and available for consultation in the following GitHub repository: https://github.com/JaimeMLopez/TFG_JaimeMateoLopez_PrediccionAtractivoUrbano

The results are structured into three main sections: the presentation and analysis of the technical experimentation, the discussion regarding the achievement of the objectives, and, finally, the justification of any deviations and methodological limitations encountered during the development of the predictive system.

6.1. PRESENTATION & ANALYSIS OF RESULTS

6.1.1. Results of Unsupervised Learning: Dimensionality Reduction and Segmentation

The first analytical step performed on the static testing dataset (year 2025) consisted of evaluating the internal structure of the data using unsupervised techniques. The objective was to determine whether, by reducing the dimensional complexity of the dataset, natural groupings of cities with similar socioeconomic profiles would emerge without relying on predefined labels.

The application of the PCA algorithm succeeded in compressing the original space of 60 features into only two principal components (PC1 and PC2), capturing 49.2% of the total variance in the data. This percentage represents a notable level of information retention given the high original dimensionality.

The analysis of the variable weights (loadings) of the original features across these two components made it possible to interpret and assign semantic meaning to the resulting Cartesian axes:

- X-axis (Component 1) – "Socioeconomic Development and Welfare State": This axis is strongly defined on its positive side by factors associated with welfare and sustained development, with key contributors including Work–Life Balance (0.217), Social Expenditure (Social SVS Expenditure: 0.214), and economic maturity and innovation (GDPPROXIMITY, URBAN INDEX, INNO CITIES). Conversely, its negative side is driven by indicators of socioeconomic vulnerability, such as extreme poverty (Poverty (% below $6.85 a day): -0.149), extreme temperatures (Average Temp. (°C): -0.135), and social inequality (GINI INDEX: -0.051).
- Y-axis (Component 2) – "Technological-Cultural Power and Soft Power": This axis captures the city's level of global leadership, innovation, and prestige. Its strongest positive drivers are Artificial Intelligence (Global AI Index: 0.433), Educational Excellence (EDUCATION NOR: 0.266), cultural influence (Movies: 0.253), and brand value (BRAND_NOR: 0.187). In contrast, destabilizing factors such as Geopolitical Risk (GEORISK: -0.392) and low migrant tolerance (TOLMIG NOR: -0.227) pull the axis towards negative values.

Within this two-dimensional space, mathematical similarity patterns were identified. To determine the optimal number of clusters, the model's inertia was evaluated using the Elbow Method.

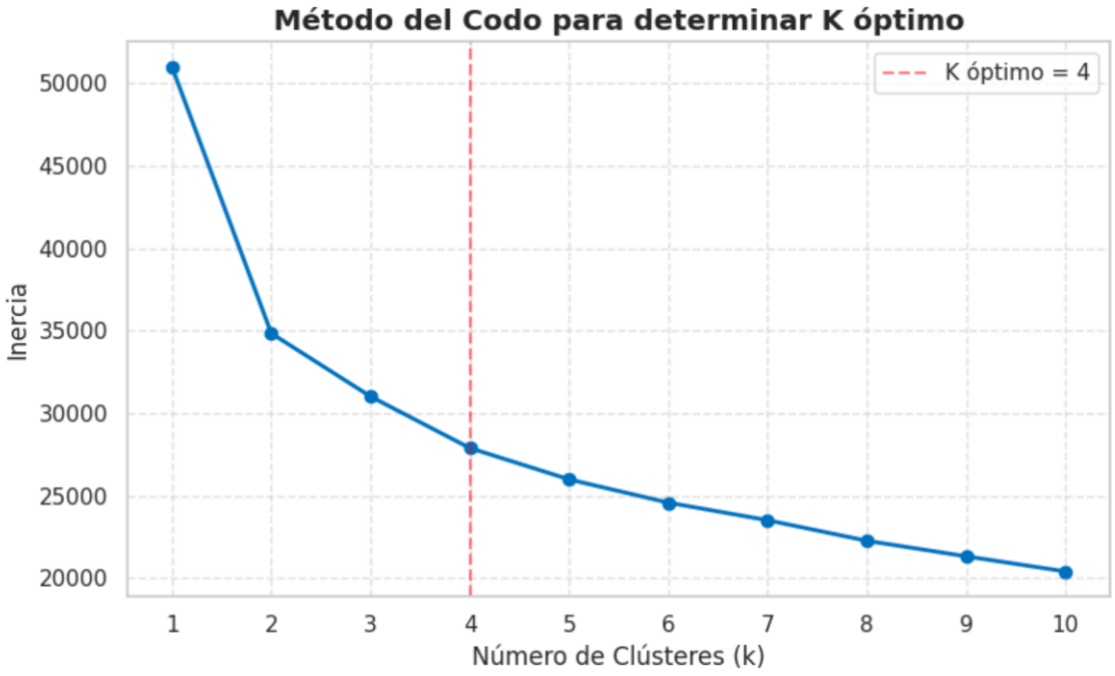

Figura 6.1 Elbow Method Plot

The inflection point indicated an optimal segmentation at k = 4, avoiding both excessive homogenization and over-fragmentation of the analysis. Once this segmentation was applied, the PCA plane visualization classified the 175 cities into four clearly defined profiles, or "urban tribes":

1. Tribe 1: Emerging Global Hubs (32 Cities). This cluster groups major metropolitan areas from Latin America, Africa, and Southeast Asia (e.g., Bogotá, São Paulo, Mumbai, Cairo, and Jakarta). These cities exhibit strong demographic and growth potential, but remain positioned in the lower region of the X-axis, constrained by persistent structural challenges related to poverty and inequality.
2. Tribe 2: Strongholds of Welfare and Quality of Life (88 Cities). This is the largest cluster, comprising the vast majority of European capitals, alongside major cities in Oceania and developed Asian urban centers (e.g., Madrid, Berlin, Copenhagen, Tokyo, Singapore, and Sydney). These cities strongly dominate the X-axis due to robust social safety nets, work–life balance, and sustained innovation.
3. Tribe 3: High-Profit American Model (18 Cities). A geographically homogeneous cluster composed exclusively of cities in the United States (e.g., New York, Los Angeles, Chicago, Miami, and Boston). These cities stand out by scoring exceptionally high on the Y-axis (technological and cultural soft power), as well as in pure profitability metrics. However, their algorithmic separation from the European cluster (Tribe 2) can be explained by a marked structural contrast: they operate under a free-market model that performs significantly worse in variables related to social cohesion, inequality (GINI), and, most notably, public safety indicators (Safe Cities Index, Numbeo Safety), revealing the structural vulnerabilities of this urban ecosystem.
4. Tribe 4: Transitional Powers and Eastern Bloc Cities (37 Cities). This cluster includes cities from China, Eastern Europe, and the Middle East (e.g., Beijing, Shanghai, Dubai, Moscow, Budapest, and Riyadh). These cities demonstrate remarkable growth in infrastructure and technological innovation, yet display a distinct scoring pattern due to institutional variables, geopolitical risk, and governance models.

Once the nature of each cluster and the semantic meaning of the axes had been established, the visualization on the Cartesian plane made it possible to geometrically appreciate how the algorithm successfully segmented the dataset. Figure 6.2 illustrates the distribution of the 175 cities according to their socioeconomic development (X-axis) and their technological-cultural power (Y-axis).

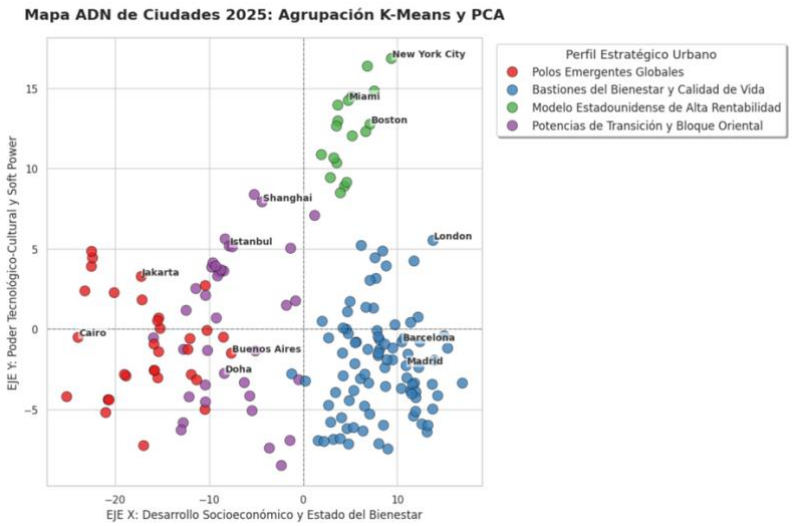

Figura 6.2 Representation of Cities in Principal Component Space and Their Segmentation into Urban Tribes

It is particularly noteworthy that this algorithmic segmentation, generated in a blind and purely mathematical manner, without providing the model with prior geographical or political labels, converges almost identically with the macroeconomic classifications established by experts (Ondiviela, 2025). The fact that the K-Means algorithm organically replicates expert consensus provides strong validation of the objectivity of the dataset. It demonstrates that the predictive variables are capable of capturing the underlying geopolitical, cultural, social, and economic reality while eliminating human bias.

6.1.2. Evaluation of Predictive Modeling: Comparative Analysis

Following the unsupervised exploration of the data, the predictive modelling phase (Phase 2) was carried out with the aim of inferring the three key dimensions of the urban ecosystem: Attractiveness, Magnetism, and Profitability. In order to establish a baseline performance and assess the linearity of the predictive variables, the first algorithm trained and evaluated on the testing dataset (2025) was Multiple Linear Regression.

Table 6.1 presents the error and goodness-of-fit metrics obtained.

Target Dimension	MAE	RMSE	R^2	R^2 Adjusted	Complexity Penalty
Attractiveness	6.9242	8.2974	0.7996	0.6941	-0.1055
Magnetism	0.2466	0.3063	0.9796	0.9689	-0.0107
Profitability	0.7996	1.0451	0.7242	0.5790	-0.1452

Tabla 6.1 Evaluation Metrics for Linear Regression (SStatic Test Set, 2025)

A critical analysis of these results reveals heterogeneous behavior depending on the target variable:

- High Linearity in Magnetism: Linear Regression demonstrates exceptional accuracy in predicting Magnetism, achieving an R^2 of 0.9796. The complexity penalty is virtually negligible (-0.0107), indicating that investment attractiveness and talent attraction capacity respond to highly linear macroeconomic relationships.
- Saturation in Attractiveness and Profitability: By contrast, the baseline model shows clear limitations when inferring Attractiveness and Profitability. In the latter case, the Adjusted R^2 drops sharply to 0.5790, accompanied by the highest complexity penalty (-0.1452). This empirically suggests that a city's profitability is a complex multivariate phenomenon governed by non-linear relationships, requiring algorithms based on decision trees or ensemble methods to capture such complexity.

To audit the behavior of the Linear Regression model in its strongest-performing dimension (Magnetism), both the prediction scatter plot and the distribution of residuals were jointly evaluated (*Figure 6.3*).

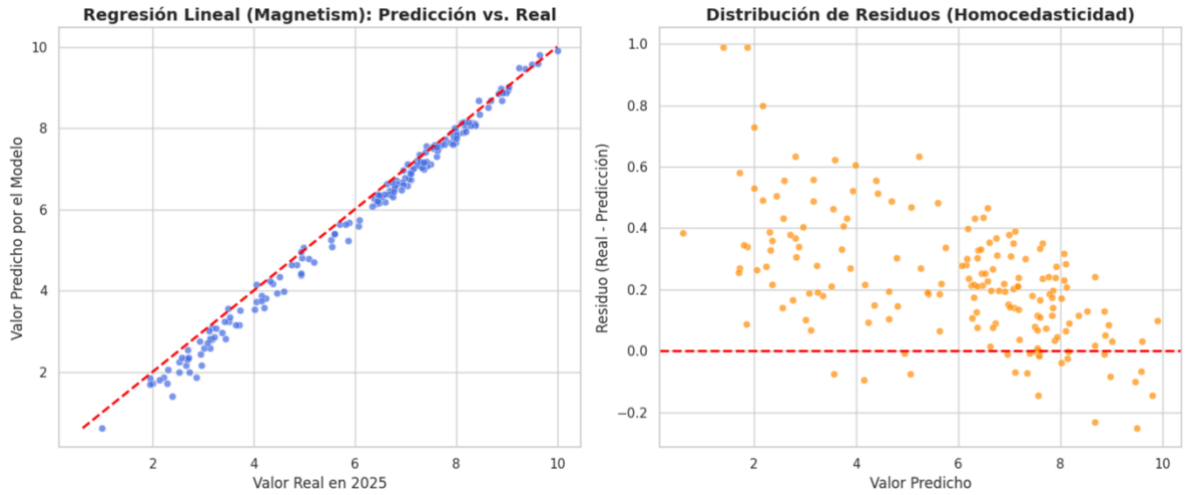

Figura 6.3 Predictive Fit and Residual Dispersion Analysis of the Linear Regression Model for Magnetism

- Linear fit (Prediction vs. Actual): The left-hand panel visually confirms the excellence of the R^2 metric (0.9796). The scatter plot adheres exceptionally closely to the bisector (the red diagonal line representing perfect prediction). This empirically demonstrates that a city's ability to attract investment and talent responds to a linear additive combination of its socio-economic factors.
- Homoscedasticity analysis (Residual distribution): The right-hand panel provides a deeper insight into the model's biases. Although the variance of the errors is very low in absolute terms (the vast majority range between 0 and 0.6 points), a critical analysis reveals a slight systematic pattern (mild heteroscedasticity). In cities with low or medium attractiveness (scores between 2 and 6), the model tends to slightly underestimate the city's potential, producing positive residuals (the actual value is higher than the prediction). Conversely, as we move towards leading cities (scores above 7), the "funnel" of dispersion narrows sharply and errors are distributed symmetrically around zero. This confirms that linear regression is an exceptionally accurate tool for benchmarking the world's most competitive cities, although it shows a slight margin for algorithmic improvement when evaluating cities in development.

Given the linearity limitations detected in the baseline model for the attractiveness and profitability dimensions, ensemble tree-based algorithms were implemented: Random Forest (bagging) and XGBoost (boosting). These models were selected for their ability to capture complex interactions among the 60 predictor variables without assuming a prior linear relationship.

Table 6.2 presents the results obtained by these two algorithms across the three dimensions of the urban ecosystem.

Dimension	Model	MAE	RMSE	R^2	R^2 Adjusted
Attractiveness	*Random Forest*	4.0220	5.8278	0.9011	0.8491
	XGBoost	4.3068	6.8818	0.8621	0.7896
Magnetism	*Random Forest*	0.3259	0.4065	0.9641	0.9453
	XGBoost	0.2888	0.3647	0.9711	0.9559
Profitability	*Random Forest*	0.5934	0.8998	0.7955	0.6879
	XGBoost	0.5204	0.7443	0.8601	0.7865

Tabla 6.2 Performance of Advanced Models (Random Forest & XGBoost) by Target Variable

The analysis of these results shows that the use of ensemble techniques significantly improves generalization performance in the most volatile dimensions. While Random Forest stands out for its robustness in the attractiveness variable, achieving a substantial reduction in mean error, XGBoost demonstrates superior effectiveness when handling profitability, a dimension that requires a higher degree of adaptability to non-linear patterns.

To conclude the evaluation phase, the results of all trained algorithms (Linear Regression vs. Random Forest vs. XGBoost) were compared. This comprehensive comparison makes it possible to determine the winning architecture for each pillar of the final simulator.

Dimension	Model	MAE	RMSE	R^2	R^2 Adjusted
Attractiveness	Lineal Regression	6.9242	8.2974	0.7996	0.6941
	Random Forest	4.0220	5.8278	0.9011	0.8491
	XGBoost	4.3068	6.8818	0.8621	0.7896
Magnetism	Lineal Regression	0.2466	0.3063	0.9796	0.9689
	Random Forest	0.3259	0.4065	0.9641	0.9453
	XGBoost	0.2888	0.3647	0.9711	0.9559
Profitability	Lineal Regression	0.7996	1.0451	0.7242	0.5790
	Random Forest	0.5934	0.8998	0.7955	0.6879
	XGBoost	0.5204	0.7443	0.8601	0.7865

Tabla 6.3 Master Comparative Table: Comprehensive Evaluation of Algorithms

The comprehensive comparison of these metrics empirically justifies the superiority of a hybrid architecture over a “one-size-fits-all” approach. At this stage, the results crown Linear Regression as the optimal model for Magnetism, Random Forest for Attractiveness, and XGBoost for Profitability.

However, before definitively consolidating these three “Champion Models” as the core of the Decision Support System (DSS), it is methodologically essential to evaluate one final, higher-complexity architecture: Deep Learning.

The aim of this phase of the experiment is to determine whether an Artificial Neural Network, designed to extract high-level abstractions and capture extremely complex hidden relationships, is capable of outperforming the benchmarks set by traditional machine learning algorithms.

Table 6.4 and Figure 4 present the performance of the Neural Network compared against the selected "Champion Models" from the previous phase. Training metrics (Train) and testing metrics (Test) have been included.

Dimension	Algorithm	R^2 *Train* (Memorisation)	R^2 *Test* (Generalization)
Attractiveness	*Random Forest*	0.9936	0.9011
	Neural Network (MLP)	0.9939	0.9052
Magnetism	Linear Regression	0.9951	0.9796
	Neural Network (MLP)	0.9982	0.9690
Profitability	*XGBoost*	1.0000	0.8601
	Neural Network (MLP)	0.9938	0.8013

Tabla 6.4 Final R² Comparison: Champion Models vs. Neural Network (MLP)

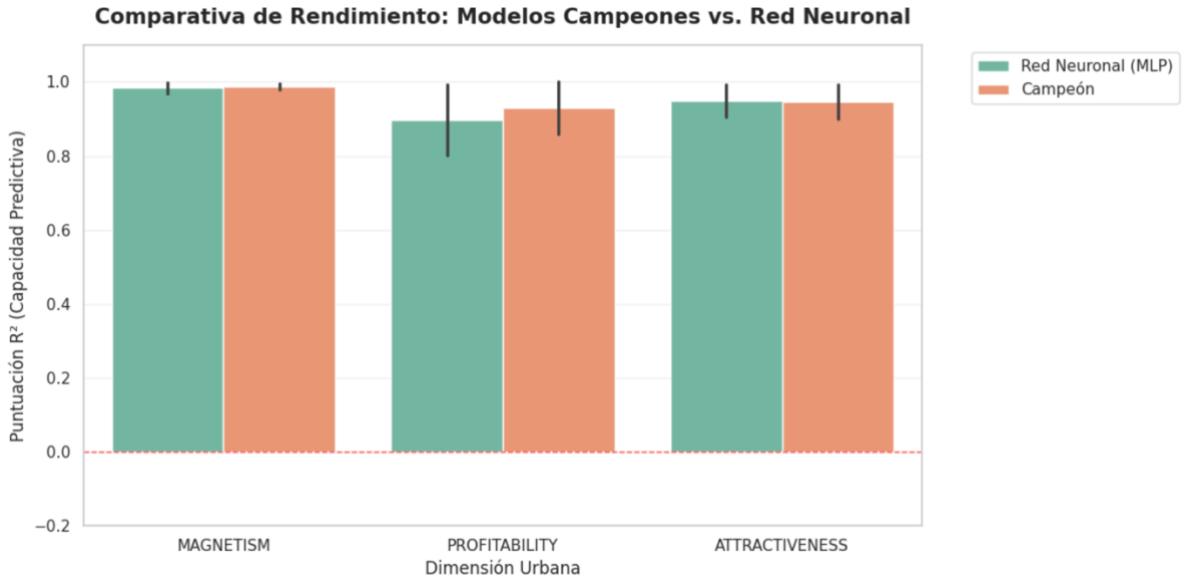

Figura 6.4 Comparative Bar Chart Between: Champion Models and Neural Network (MLP) (Test)

The audit of the metrics reveals highly illustrative behavior in the training phase (Train): all algorithms achieve an almost perfect fit on the known data, reaching R^2 scores between 0.9936 and 1.000. However, the selection of the optimal algorithm does not depend on this memorization, but on the generalization gap (the difference between Train and Test performance).

When the models are exposed to unseen data, the high-complexity architectures clearly exhibit overfitting. For instance, XGBoost and the Neural Network suffer significant drops in their R^2 when predicting profitability. In contrast, this analysis highlights the exceptional

robustness of Linear Regression in Magnetism, as it is the algorithm with the smallest generalization gap.

Focusing exclusively on real predictive performance on the Test set, the following conclusions are drawn:

- Magnetism and Profitability: The Neural Network does not outperform the traditional champion models in these dimensions. Linear Regression continues to dominate Magnetism (0.9796 vs. 0.9690), and XGBoost maintains its absolute superiority in profitability prediction (0.8601 vs. 0.8013).
- Attractiveness: In the attractiveness dimension, the Neural Network slightly outperforms Random Forest on the test set (0.9052 vs. 0.9011). However, from an efficient engineering perspective, this marginal improvement does not empirically justify the substantial increase in computational cost, inference time, and the requirements of modern interpretability guidelines, which favor transparent models when their performance is comparable to that of a more complex architecture (Rudin, 2019). Therefore, Random Forest is identified as the most effective and balanced algorithm for this task.

6.1.3. *Walk-Forward Validation*

To ensure that the Decision Support System (DSS) remains stable over time and does not overfit to a specific year, a progressive temporal validation technique (Walk-Forward Validation) was applied. This methodology preserves the chronological structure of the data. The models are trained exclusively on past historical data to predict the immediate following year, thereby simulating the real-world behavior of the system in production.

Table 6.5 presents the evolution of the R^2 score for the three engines of the simulator.

Year to be Predicted (Test)	Training Windows	*Attractiveness* (*Random Forest*)	*Magnetism* (Linear Regression)	Profitability (*XGBoost*)
2021	2020	0.8872	0.9405	0.6505
2022	2020 a 2021	-197.8119	0.9865	0.3631
2023	2020 a 2022	-1.0184	0.9730	-0.3004
2024	2020 a 2023	0.5332	0.9862	0.8044
2025	2020 a 2024	0.9011	0.9796	0.8601

Tabla 6.5 Predictive Performance Evolution (R^2) using Walk-Forward Validation (2021-2025)

- Impact of the “Black Swan” (COVID-19) on Attractiveness and Profitability

The results reveal the impact of the “Black Swan” event associated with the COVID-19 pandemic. When attempting to predict 2022 and 2023 using training data from 2020 (a global disruption year), the Attractiveness and Profitability models collapse. This does not indicate a failure of the algorithm, but rather the impossibility of forecasting post-lockdown market disruption and the inflationary crisis using data from an exceptional period. The recovery of excellent metrics in 2024 and 2025 demonstrates the system’s ability to absorb shocks once the temporal window includes the “new normal”.

- Magnetism audit: exclusion of data leakage (data leakage)

Given the unusually high accuracy of the Magnetism model (maintaining an $R^2 > 0.94$ even during crisis years), a data leakage test was conducted by analyzing the model coefficients. The objective was to verify that the algorithm was not using "contaminated" variables, i.e., features that act as direct shortcuts to the outcome.

The results of the leakage detector (*Table 6.6*) confirm the validity of the model.

Variable	Absolute Weight
SMARTCITIES NOR	0.198400
BRAND_NOR	0.116164
HIST	0.108806
Expat Experience NOR	0.103049
CLIMATE-NOR	0.094013

Tabla 6.6 Top 5 Variables with the Highest Feature Importance in Predicting Magnetism

As can be observed, the pillars of magnetism are structural and inert variables. Factors such as climate, historical heritage, or a city's digital maturity do not change dramatically in response to a pandemic or a financial crisis. This explains why the model is able to predict this dimension with such high accuracy and temporal stability. A city's magnetism is a long-term construct that does not depend on the volatility of the immediate economic cycle.

6.1.4. Global Interpretability Analysis (SHAP Values)

Once the algorithmic architecture and its temporal robustness have been validated, an empirical audit of the three engines of the simulator is carried out through the extraction of SHAP values. This analysis makes it possible to identify the true causal drivers of the urban ecosystem and to assess the marginal impact of the predictor variables.

To understand the driving forces that define a city's attractiveness, SHAP values were extracted from the Random Forest model on the test set. *Figure 6.5* presents the beeswarm plot, which ranks variables from highest to lowest global impact, while simultaneously showing the direction of their effect.

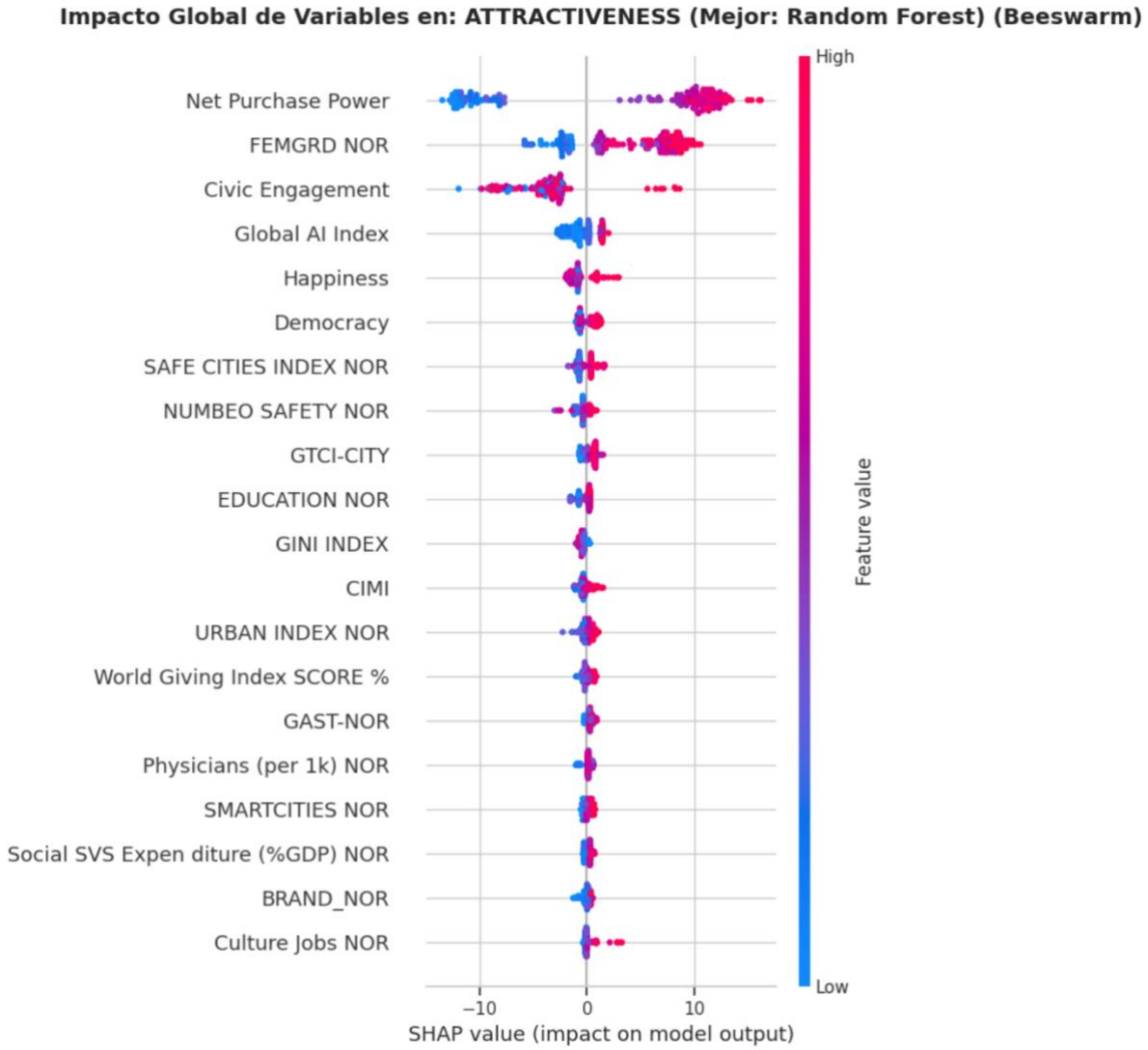

Figura 6.5 Global Impact of Variables on Attractiveness Prediction (Random Forest)

The causal profile analysis reveals highly relevant sociological insights for the urban ecosystem:

1. Hegemony of purchasing power: The Net Purchase Power variable clearly dominates the ranking. As indicated by the colour dispersion, high values of this metric (red points) significantly increase city attractiveness, whereas low values (blue points) penalise it with equal severity.
2. The weight of social progress: Immediately after the purely economic factor, the model identifies socio-demographic variables as the main drivers of attractiveness. Variables such as FEMGRD NOR (associated with gender equality and female graduation rates) and Civic Engagement rank second and third. In both cases, high scores strongly shift the prediction to the right, demonstrating that talent and capital perceive more equal and participatory societies as significantly more attractive cities.

To further audit the behaviour of the leading variable, its marginal effect was isolated in *Figure 6.6*.

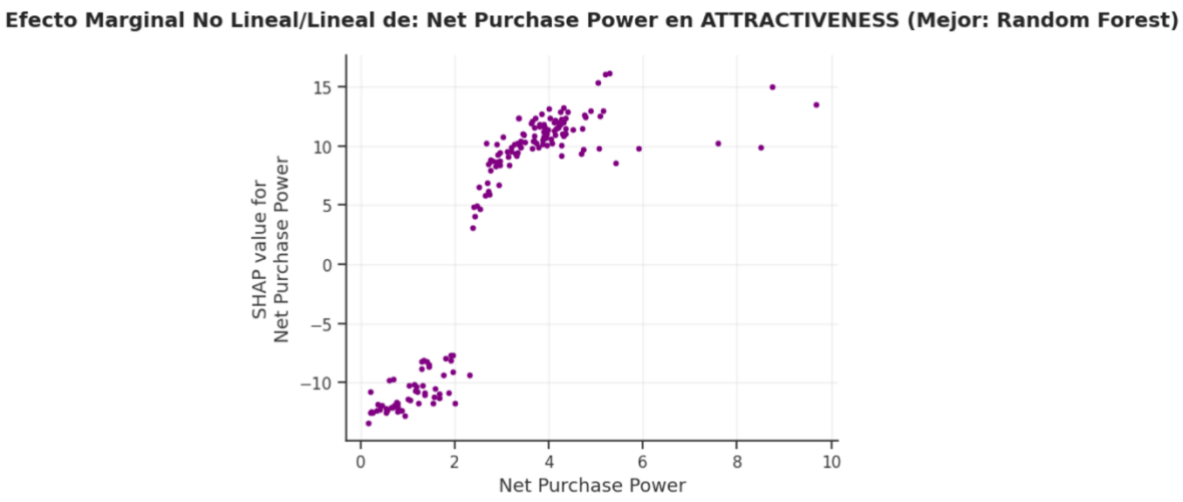

Figura 6.6 Non-Linear Marginal Effect of Net Purchase Power on Urban Attractiveness

The dependence plot empirically justifies why a non-linear algorithm (Random Forest) outperformed classical regression in this dimension. The impact of purchasing power does not follow a straight line, but instead shows a clear S-shaped curve (step-like function) that reflects realistic economic behavior.

- Severe penalty below the threshold: When Net Purchase Power is below 2.5, the city experiences a sharp and constant penalty in attractiveness (SHAP values around -12).
- The turning point: Between values of 2.5 and 4.0, there is a sudden "jump". A slight increase in living standards within this range dramatically boosts city attractiveness into positive values.
- Diminishing returns law: From a purchasing power level of 4.0 onwards, the curve flattens out. This means that once a city guarantees a high level of wealth, becoming "extremely richer" does not increase attractiveness indefinitely. The model captures that, beyond a certain welfare threshold, other factors (such as civic engagement or equality) become more important in making a city desirable.

In contrast to the volatility observed in attractiveness, urban magnetism is governed by highly linear patterns. To visualize the weight of these structural dynamics, SHAP values from the Linear Regression model were extracted and are presented in *Figure 6.7*.

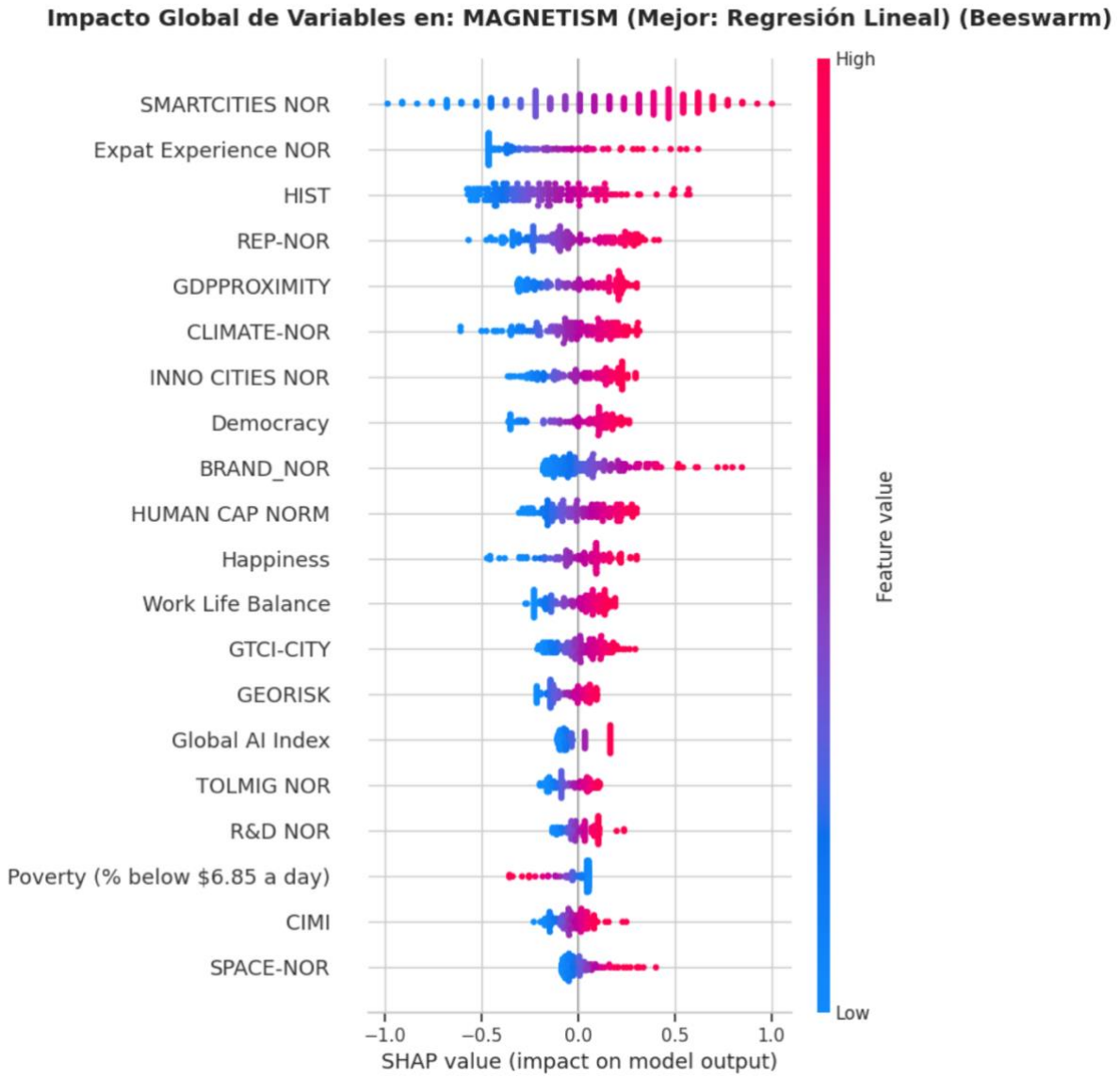

Figura 6.7 Global Impact of Variables on Magnetism Prediction (Linear Regression)

The architecture of this beeswarm plot exhibits perfect symmetry (high values in red cluster strictly to the right, and low values in blue to the left), which is the characteristic and expected behavior of a multiple linear regression model. From this analysis, the following causal drivers are identified:

1. Technological maturity as the main driver: The SMARTCITIES NOR variable (index of smart infrastructure development) is positioned as the most influential factor. Cities with a high level of technological implementation generate a strong positive impact on overall magnetism.
2. Validation of international talent and urban heritage: In second and third place, Expat Experience NOR (related to the experience of foreigners evaluating the best and worst places to live and work) and HIST (historical heritage) emerge as key factors. This empirically confirms the hypothesis that a city's magnetism does not respond to transient trends, but rather to the consolidation of its infrastructure, its history, and the positive perception of the talent already residing in it.

To visualize the behavior of the main variable, its marginal dependence plot was generated (*Figure 6.8*).

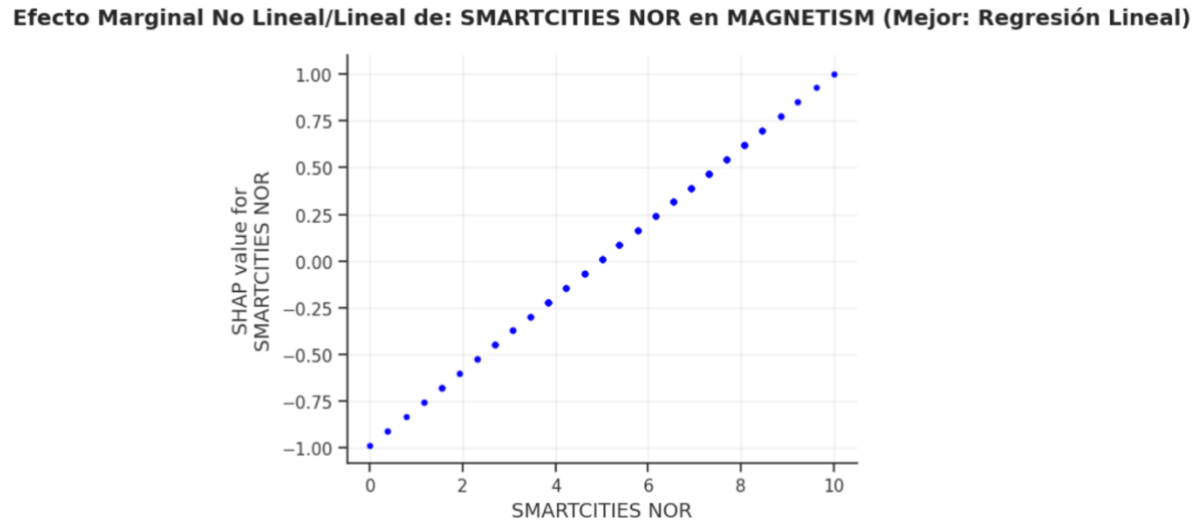

Figura 6.8 Strictly Linear Marginal Effect of Smart City Infrastructure (SMARTCITIES NOR) on Magnetism

Unlike the saturated behaviour (S-shaped curve) observed for net purchasing power in urban attractiveness, the impact of technological infrastructure on magnetism follows a perfect diagonal line. This finding reveals a pattern of constant marginal returns: for each additional point a city invests in or improves its Smart City ecosystem, its overall magnetism increases by a fixed and constant proportion, with no signs of exhaustion or saturation within the observed scale (0 to 10).

Urban profitability proved to be the most volatile dimension during the model comparison, requiring the power of a Gradient Boosting algorithm (XGBoost) to capture its complex business rules. To examine the causal drivers of capital and investment, SHAP values were extracted from the test set, with their impact shown in *Figure 6.9*.

Figura 6.9 Global Impact of Variables on Profitability Prediction (XGBoost)

The analysis of the scatter plot reveals the harsh reality of capital flows, prioritizing risk reduction above all other factors:

1. Safety as a non-negotiable prerequisite: The SAFE CITIES INDEX NOR variable dominates the ranking unequivocally. Markets and investors severely penalize cities with low safety levels (blue points on the left), while high safety (red points on the right) is the strongest driver of urban profitability.
2. Purchasing power as a consumption engine: In second place, Net Purchase Power reappears, confirming that for a city to be profitable for businesses, its citizens must have spending capacity.
3. The cost of the welfare state: Interestingly, XGBoost detects a counterintuitive pattern in social variables such as FEMGRD NOR or Happiness. Unlike in the attractiveness dimension (where they had a positive impact), in profitability these very high social indicators (red points) tend to shift the prediction slightly to the left (negative impact). This reflects a real macroeconomic dynamic: highly developed

welfare societies are often associated with higher taxation and labor costs, which reduce corporate profit margins.

To examine the behavior of urban safety in more detail, the marginal effect of the leading variable was isolated in *Figure 6.10*.

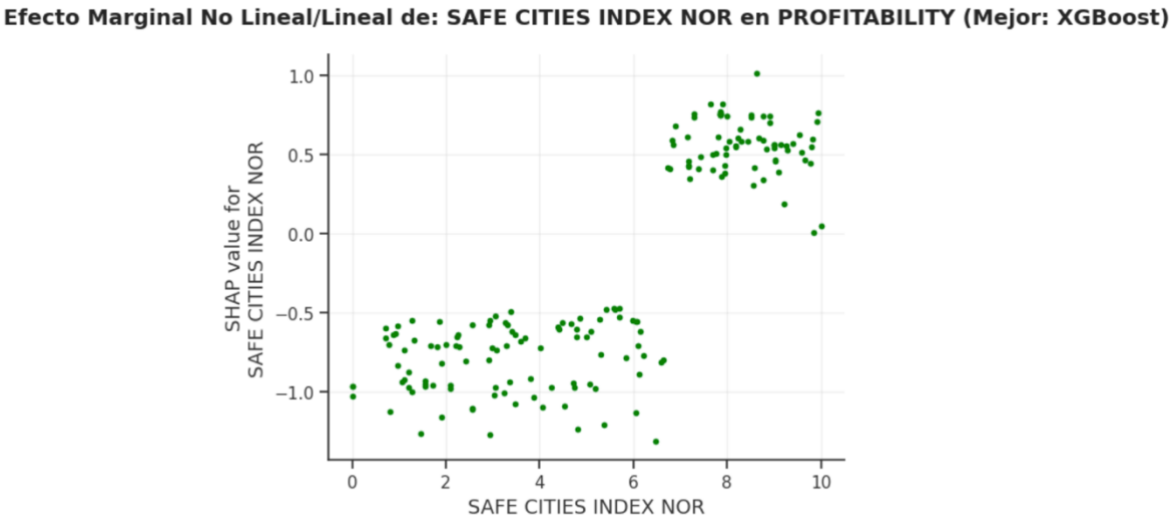

Figura 6.10 Threshold (Bifurcation) Marginal Effect of SAFE CITIES INDEX NOR on Profitability

The dependence plot demonstrates why XGBoost, an algorithm based on strict data partitioning, outperformed in this dimension. The impact of safety on profitability is not gradual, but instead follows a step function or threshold effect (bifurcation):

- Capital exclusion zone: From level 0 up to approximately 6.5 on the safety index, the city suffers a constant flat penalty. For investors, a “somewhat unsafe” city and a “very unsafe” city are equally unviable.
- Stabilization: Once the critical threshold of 6.5 is crossed, there is an abrupt vertical jump. The SHAP value stabilizes into a positive horizontal band. Becoming “infinitely safer” beyond this point does not lead to exponential increases in profitability.

As a conclusion to this interpretability analysis, a synthesis of the variables with the greatest impact on cities is presented. Far from there being a single “master formula”, opening the black box of the algorithms has shown that each pillar of the city requires a different strategic approach.

As summarized in *Table 6.7*, while magnetism depends on structural technological investment, attractiveness is a human-driven factor governed by citizen wealth, and profitability is a market-driven factor strictly conditioned by physical and legal security.

Dimension	Winning Model	Top 1 (Highest Impact)	Top 2	Top 3
Attractiveness	*Random Forest*	*Net Purchase Power* (Poder Adquisitivo Neto)	*FEMGRD NOR* (Graduación Femenina / Igualdad)	*Civic Engagement* (Compromiso Cívico)
Magnetism	Regresión Lineal	*SMARTCITIES NOR* (Infraestructura *Smart City*)	*Expat Experience NOR* (Experiencia de Expatriados)	*HIST* (Patrimonio Histórico)
Profitability	*XGBoost*	*SAFE CITIES INDEX NOR* (Índice de Seguridad Urbana)	*Net Purchase Power* (Poder Adquisitivo Neto)	*FEMGRD NOR* (Graduación Femenina / Igualdad)

Tabla 6.7 Synthesis of the Main Casual Drivers by Urban Dimension according to SHAP values

6.1.5. Practical Application: Results of the Decision Support System (DSS)

Once the predictive engines (Linear Regression, Random Forest, XGBoost, and Neural Network) had been consolidated and interpreted, the final phase of the technical experimentation consisted of executing the simulation environment. The objective of this Decision Support System (DSS) is to go beyond mere statistical prediction and provide analysts and investors with an interactive strategic analysis tool.

The following section presents the empirical validation results of the simulator operating on real urban environments, divided into three functional modules: historical consistency analysis, short-term probabilistic forecasting, and stress testing.

- Module 1: Backtesting and historical consistency

The first step in validating the simulator consisted of verifying its ability to reconstruct the past. It is important to emphasize that the simulation environment has been designed to be fully dynamic; that is, the script is capable of ingesting and processing the data matrix of any city included in the dataset. To illustrate the results of this phase, Madrid has been selected as a demonstrative case study.

When Madrid is introduced as an input parameter, the system extracts its historical vector (2020–2025) and subjects it to inference using the three previously trained champion models. The resulting comparative panel is shown in *Figure 6.11*.

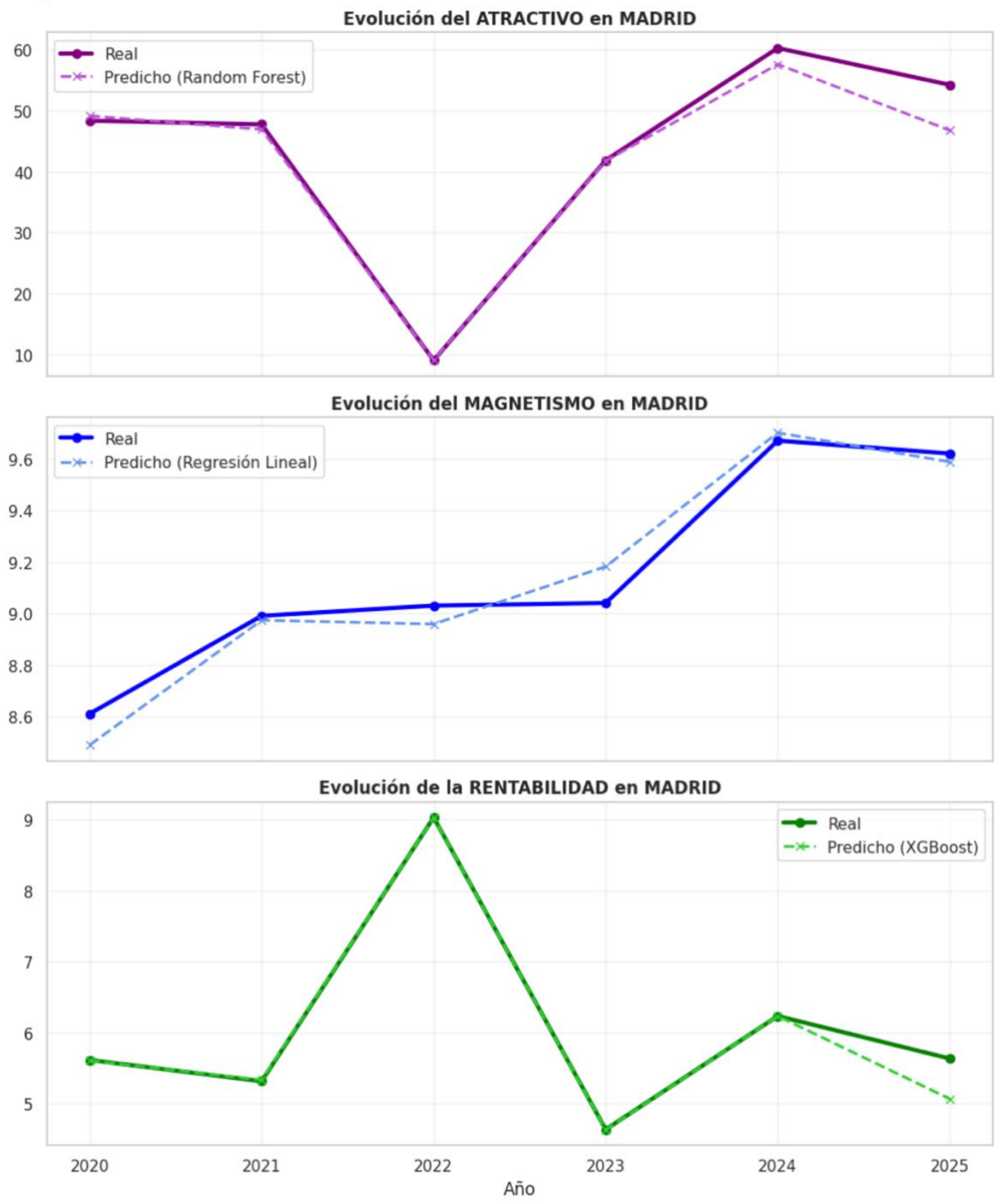

Figura 6.11 Madrid Macktesting: Comparison of Real Values vs. Algorithmic Reconstruction

This module is not intended to evaluate model generalization (an aspect already demonstrated previously), but rather to perform a technical consistency test. The visual analysis of the trajectories confirms the architectural success of the system:

1. High-fidelity algorithmic reconstruction: As observed in the upper and lower panels, the non-linear models (Random Forest for Attractiveness and XGBoost for Profitability) achieve an almost perfect overlap with the real trajectory, closely matching even the extreme peaks and drops of 2022 and 2023. This empirically confirms the structural memorization capability of these algorithms over the training data. The system demonstrates that it has learned Madrid's specific business rules. It

also achieves strong results for 2025, showing that it has not only learned past data, but is also capable of producing reliable predictions for previously unseen years.

2. Linear trend consistency: In the Magnetism dimension (central panel), Linear Regression, due to its inherently less memorization-prone mathematical nature, produces a smoothed trajectory that closely tracks the city's growth inertia, confirming the robustness of this indicator.

By demonstrating that the DSS can reconstruct historical behavior without structural deviations, the integrity of the code is validated. This mathematical foundation allows the next module to be executed with confidence: the extrapolation of these vectors into unobserved future scenarios.

- Module 2: Future Trend Projection (Natural Extrapolation)

Once the structural robustness of the code had been confirmed, the DSS was used for its primary purpose: predictive inference. This second module generates a synthetic data vector for the year 2026 based on the recent temporal inertia of the city (by estimating the first derivative of the post-pandemic period 2022–2025).

To test the algorithm's sensitivity to rapidly expanding urban ecosystems, the city of Riyadh (Saudi Arabia) was selected as a case study. The system generated its 2026 predictive vector (applying a mathematical bounding constraint from 0 to 10 to respect the feature space) and processed it through the champion models.

The expected growth results are detailed in *Table 6.8* and visualized in *Figure 6.12*.

Dimension	2025 Score	2026 Score (Proyected)	Growth (%)
Attractiveness	15.623	22.517	+44.131%
Magnetism	2.965	3.121	+ 5.280%
Profitability	4.997	5.099	+2.039%

Tabla 6.8 Probabilistic Projection of Natural Growth for Riyadh (2025 vs. 2026)

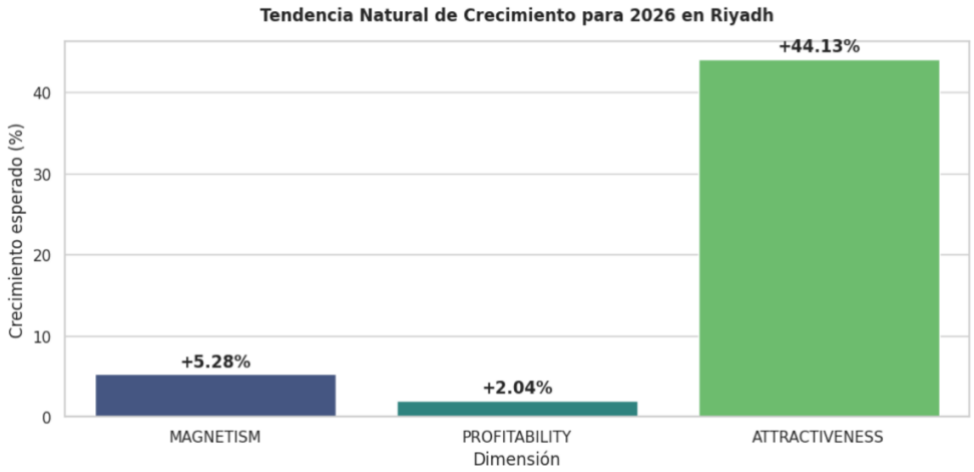

Figura 6.12 Expected Natural Growth Trend for Riyadh in 2026

The extrapolation analysis yields strategically valuable metrics, demonstrating the simulator's ability to capture both linear and non-linear dynamics:

1. Sustained structural growth: In the magnetism and profitability dimensions, the city projects solid and stabilised growth (+5.28% and +2.04%, respectively). This is consistent with the nature of the governing algorithms (Linear Regression and the threshold effect of XGBoost), indicating that Riyadh continues to improve its Smart City infrastructure and consolidate investor confidence at a natural pace.
2. The attractiveness "inflection point" (+44.13%): The most revealing finding of the simulation is the exponential jump projected by the Random Forest model in the human dimension (Attractiveness). Far from being a statistical anomaly, this peak empirically validates the S-shaped curve identified during the SHAP interpretability analysis (see Figure 6.6). The simulator detects that Riyadh's projected growth trajectory for 2026 pushes its purchasing power and socio-economic progress metrics beyond a "critical threshold". Once this positive saturation point is crossed, the non-linear model sharply increases the city's attractiveness, shifting from arithmetic to geometric growth.

This short-term projection demonstrates that the DSS does not simply extrapolate straight lines into the future, but is also capable of alerting investors and urban planners to imminent "quantum leaps" in a city's competitiveness.

- Module 3: Adversarial scenario simulation (Stress Testing)

As a culmination of the descriptive phase, the simulator incorporates a resilience evaluation engine based on the Ceteris Paribus technique. This analysis isolates a single urban variable and injects a "Black Swan" event to observe its cascading effect on predictions without interference from other indicators. To validate the tool, two critical macroeconomic scenarios were simulated.

A) Economic collapse and hyperinflation (case study: London)
The first stress scenario assessed the vulnerability of a major financial capital to a severe economic crisis. Using London as a baseline (2025), a massive drop in citizens' purchasing power was simulated (applying a relative penalty of -3.0 points to the Net Purchase Power variable, while keeping it within the mathematical lower bound of 0).

The inference results, shown in *Table 6.9* and *Figure 6.13*, reveal a radically asymmetric impact across the city's dimensions.

Dimension	Actual Value (2025)	Crisis Projection (2026)	Net Variation (%)
Attractiveness	48.458	20.636	-57.415%
Magnetism	9.795	9.798	+0.022%
Profitability	5.093	4.793	-5.881%

Tabla 6.9 Projected Impacto f a Purchasing Power of Crisis (-3.0 Points) in London

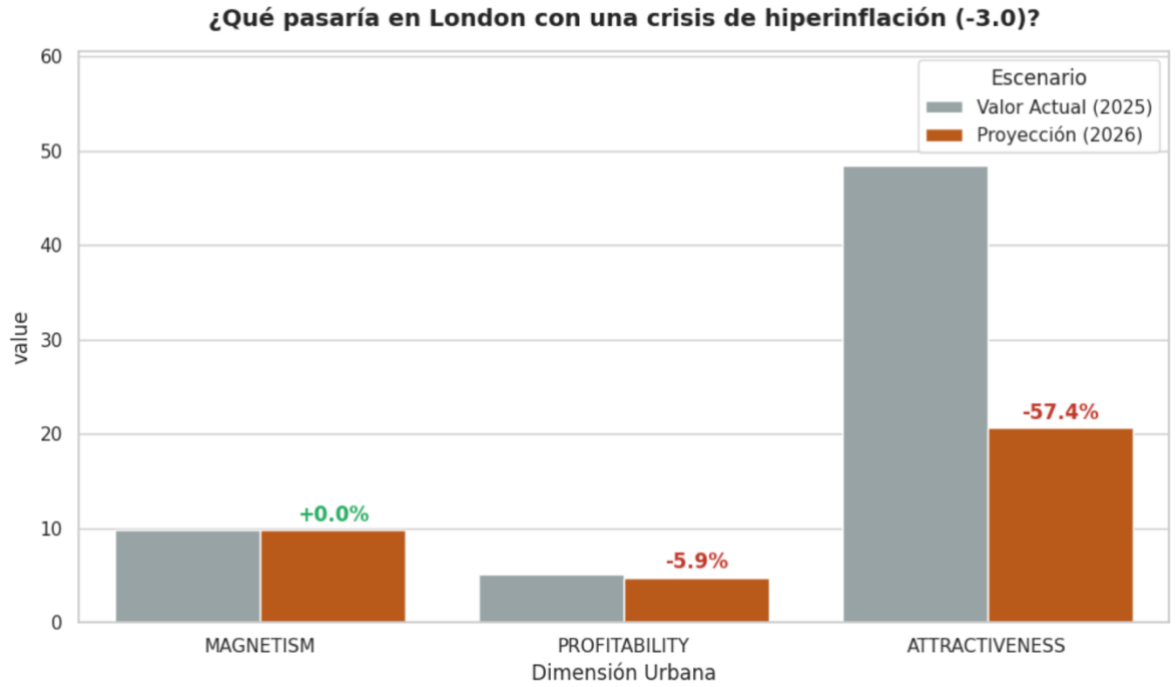

Figura 6.13 Univariable Impact Simulation: Collapse of Purchasing Power in London

The simulator demonstrates outstanding analytical maturity, producing conclusions that cross-validate the algorithmic interpretability audit (SHAP) carried out in the previous phase:

1. Collapse of human capital (Attractiveness): Random Forest penalizes the city with a massive loss of attractiveness (-57.41%). By subtracting 3 points from London's purchasing power, the model detects that the city "falls off the cliff" of the S-shaped curve identified in the marginal analysis (Figure 6.6). Hyperinflation and declining living standards destroy the city's social value proposition.
2. Structural resilience (Magnetism): In a striking result, Linear Regression keeps the city's magnetism completely unchanged (+0.0%). The algorithm mathematically captures urban reality: a temporary financial crisis does not erase London's historical heritage (HIST) or dismantle its Smart City infrastructure overnight. Magnetism acts as an inertial "anchor".
3. Capital pragmatism (Profitability): The XGBoost model predicts a moderate decline in profitability (-5.88%). Although reduced purchasing power lowers local consumption (affecting margins), the algorithm still considers London a relatively profitable market because its main pillar—safety (Safe Cities Index)—has not been altered in this simulation.

The simulator shows that a purely economic crisis devastates talent retention and the social attractiveness of a city, but is not sufficient to destroy its historical magnetism or completely eliminate corporate investment, provided that safety remains intact.

A) Degradation of civic environment and safety (case study: Vienna)

The second univariate stress scenario was designed to assess the vulnerability of a city characterized by extremely high quality-of-life standards. Using Vienna as the baseline, an absolute stress shock was applied: instead of a percentage drop, the main safety variable

(SAFE CITIES INDEX NOR) was deliberately forced to a critical level of 2.0 out of 10, simulating severe civic disruption (disregarding its historically excellent starting point).

The results of this simulation are detailed in *Table 6.20* and *Figure 6.14.*

Dimension	Actual Value (2025)	Crisis Projection (2026)	Net Variation (%)
Attractiveness	48.397	47.801	-1.231
Magnetism	8.868	8.804	-0.723
Profitability	5.115	4.669	-8.713

Tabla 6.10 Projected Impact of a Critical Drop in Urban Safety (Level 2.0) in VIenna

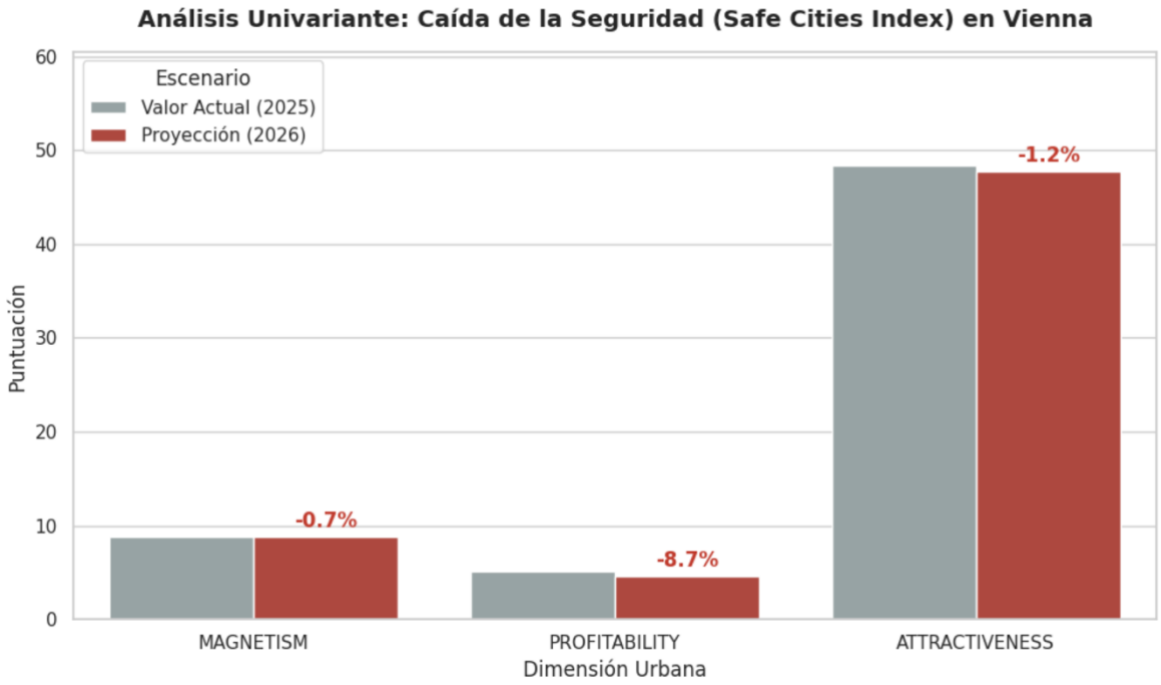

Figura 6.14 Univariate Impact Simulation: Collapse of Safety in Vienna

The analysis of this scenario reveals a different urban behavior dynamic compared to the financial crisis in London, once again validating the rules extracted in the interpretability phase (SHAP):

1. Corporate capital flight (Profitability): The most severe impact is observed in city profitability, with a decline of nearly 9% (-8.71%). This collapse empirically reflects the "threshold effect" identified in the XGBoost model (Figure 6.10). By forcing the safety index to 2.0, Vienna is pushed out of the "investment grade" zone. The algorithm demonstrates that capital is pragmatic and does not tolerate physical or legal risk. Without safety, expected profitability collapses, regardless of how wealthy the city is.
2. Inertia of human quality of life (Attractiveness): Unlike the economic crisis scenario, the collapse in safety has only a marginal impact on social attractiveness (-1.23%). This is explained by the fact that, according to the Random Forest rules, the main drivers of attractiveness are net purchasing power and civic engagement. Under the Ceteris Paribus principle, Vienna remains a wealthy city with a highly engaged society.

The model interprets that a temporary spike in insecurity does not immediately trigger talent outflow if the welfare state remains intact.

3. Structural stability (Magnetism): Once again, Linear Regression shows that magnetism is an inert dimension (-0.72%). A safety crisis does not alter the weight of historical heritage or degrade the existing Smart City infrastructure of the Austrian capital.

The comparison between the London and Vienna scenarios confirms the full success of the DSS. The simulator has successfully isolated the "two souls" of the city:

- A purely economic crisis (London) destroys social fabric and attractiveness, but preserves profitability and business activity if safety remains intact.
- A safety crisis (Vienna) immediately drives corporate profitability down, although the welfare state partially buffers the decline in urban attractiveness.

The DSS thus goes beyond statistical prediction, becoming a prescriptive intelligence tool capable of anticipating second-order effects of public policies and global shocks.

6.2. Discussion on the Fulfilment of Objectives

After the detailed presentation of the technical experimentation and the empirical results obtained from the simulator, it is essential to conduct a retrospective assessment to confirm that the project has fulfilled the fundamental objectives established during its design phase. The following section presents the degree of achievement of the general objective and the specific objectives, as well as the adherence to the validation methods defined in Chapter 3.

6.2.1. Fulfilment of the General Objective

The main objective of the research was to "analyze cities from a quantitative perspective in order to predict urban attractiveness using different indicators, thereby understanding which factors have the greatest influence."

This objective has been fully achieved. To address it with maximum analytical rigor, the macro concept of "urban attractiveness" was broken down and modelled through its three constituent pillars: socio-economic attractiveness (Attractiveness), structural magnetism (Magnetism), and corporate profitability (Profitability). The use of advanced machine learning predictive models has made it possible to translate this wide range of abstract indicators into reliable mathematical predictions, empirically identifying the exact drivers that influence a city's global competitiveness.

6.2.2. Fulfilment of the Specific Objectives

The project roadmap was divided into eight specific objectives, all of which have been successfully addressed and achieved throughout the different methodological phases:

1. Exploring the dataset to understand the variables: An exhaustive exploratory data analysis was carried out in the initial stages, identifying the distribution and nature of urban features.
2. Preparing the database and addressing multicollinearity: Through advanced preprocessing techniques, a robust dataset was generated. The application of variance analysis and correlation checks made it possible to isolate and correct structural multicollinearity issues, ensuring that the models did not suffer from overfitting due to redundant variables.
3. Developing regression and machine learning models: A diverse algorithmic architecture was built, implementing parametric approaches (Multiple Linear Regression), powerful tree-based ensembles (Random Forest and XGBoost), and a Deep Learning architecture (Artificial Neural Network).
4. Evaluating predictive performance: All algorithms were subjected to strict evaluation. Their ability was validated not only in terms of fitting the data, but also in generalizing to unseen future information.
5. Comparing models: A "model tournament" was conducted, allowing the identification of the champion algorithm for each urban dimension based on its mathematical architecture (Random Forest for Attractiveness, Linear Regression for Magnetism, and XGBoost for Profitability).
6. Interpreting results and influential factors: This objective was addressed through the implementation of the SHAP analytical framework. Global (beeswarm plot) and local (dependence plots) analyses unraveled the models' "black box", identifying, for instance, that purchasing power and safety are the main drivers of attractiveness and profitability, respectively.
7. Applying unsupervised learning and dimensionality reduction techniques: During feature engineering phases, techniques were used to cluster variables and extract latent features, reducing noise in the dataset and improving the computational efficiency of subsequent supervised models.
8. Designing a strategic scenario simulator (What-If analysis): As the culmination of the project, a DSS was successfully developed. This simulator demonstrated its ability to project natural trends (such as the upcoming surge of Riyadh in 2026) and to perform univariate stress tests, illustrated through simulations of financial collapse in London and a security crisis in Vienna.

6.2.3. Adherence to Validation Methods

Finally, it is worth highlighting that the achievement of the aforementioned objectives was strictly supported by the proposed methodological validation framework:

- Realistic temporal split: As required, model training was strictly limited to the 2020–2024 period, reserving 2025 exclusively as the test set. This approach ensured the absence of any future data leakage.
- Rigorous mathematical metrics: The selection of the winning models was not based on subjective judgement, but empirically supported by the minimization of absolute and squared errors (MAE and RMSE) and the maximization of explained variance (R^2).
- Cross-interpretability audit: The quantitative validation of errors was complemented by a qualitative validation through SHAP values, demonstrating that the algorithms

were not only accurate but also logically consistent with macroeconomic and sociological theory.

6.3. Justification of Deviations and Methodological Limitations

Despite the high predictive performance achieved and the validity of the DSS, this study is not without limitations. Acknowledging these is essential for the correct interpretation of the results and for establishing the foundations for future lines of research.

6.3.1. Limitations in Data Nature and Temporality

The main methodological constraint lies in the temporal horizon of the dataset used (2020–2025). This period includes the COVID-19 health crisis and the subsequent economic recovery, introducing a level of atypical "noise" and volatility into urban time series.

- Post-pandemic noise: Although data cleaning and normalization techniques were applied, it is possible that some patterns detected by the non-linear models (such as peaks in Attractiveness or Profitability) are influenced by post-crisis technical rebounds that may not repeat under periods of prolonged economic stability.
- Indicator updates: The fidelity of the simulator depends on the frequency with which international indices are updated. Any latency in the publication of real-world data could temporarily misalign the model's predictions with current conditions.

6.3.2. Technical Deviations and the Development of Overfitting

As analyzed in the evaluation phase, the models exhibited an extremely high fitting capacity (particularly during the training stage).

- Risk of memorization: Given that the dataset is temporally balanced, the risk of overfitting arises from potential sample homogeneity. If the training set is predominantly composed of a specific city profile (for example, high-income Western capitals), the algorithm optimizes its rules for that dominant ecosystem. This implies that predictive accuracy could be compromised if, in the future, the simulator is applied to emerging market cities or urban profiles that differ radically from the original dataset distribution.
- Stationarity assumption: The models assume that the "rules of the game" governing urban ecosystems remain constant over time. However, abrupt regulatory changes or disruptive technological innovations could invalidate the coefficients learned by the algorithms in the past.

6.3.3. Limitations of the Simulation Environment (Ceteris Paribus)

The DSS Stress Testing engine operates under the Ceteris Paribus principle (holding all other variables constant while modifying a single one).

- Interdependence of variables: In real urban systems, variables are deeply interconnected. For instance, a collapse in safety (as simulated in Vienna) often triggers a decline in purchasing power or a degradation of infrastructure. The current simulator, being a univariate analysis, does not automatically capture these second-order effects or "cascade effects", requiring expert interpretation from the analyst to complete the scenario.
- Trend simplification: The natural projection module uses the recent average rate of change to extrapolate the year 2026. This is a linear approximation which, although useful for strategic planning, cannot anticipate "Black Swan" events or random shocks that break a city's historical inertia.

7. ETHICAL IMPLICATIONS & SOCIAL IMPACT

This project focuses on the quantitative analysis of urban attractiveness through the use of regression models and machine learning techniques applied to a multivariate international dataset. This data-driven approach, based on predictive modelling, requires careful reflection on the ethical, social, and methodological implications arising from its development and application.

In this context, it is essential to examine aspects such as the privacy of the information used, the potential existence of biases in both the data and the models, and the social impact that may result from the conclusions drawn. This chapter addresses these issues from a critical perspective, considering both the value of the project and the associated risks, in line with current guidelines on ethics in artificial intelligence and data analysis (European Commission, 2019; UNESCO, 2021).

7.1. ETHICAL & LEGAL FRAMEWORK

As previously mentioned, this project complies with the fundamental ethical principles in research established by the European Commission in 2019 (European Commission, 2019), such as transparency, impartiality, scientific rigor, and responsibility in the handling of information.

Firstly, it is important to highlight that the dataset used does not contain personal information or sensitive data, but is composed exclusively of aggregated indicators at city level. Therefore, it does not infringe individual privacy nor violate current data protection regulations, such as the General Data Protection Regulation (GDPR) (European Union, 2016). More specifically, the data used falls outside the definition of “personal data” established in Art. 4.1, while strictly complying with the data minimization principles set out in Art. 5.1.c. At a national level, the project is also aligned with Organic Law 3/2018 on Personal Data Protection and Guarantee of Digital Rights (LOPDGDD) (Organic Law 3/2018, 2018), remaining outside its scope of application as it works exclusively with statistical and anonymized information (Art. 2).

Additionally, as this project is based on the development of Machine Learning algorithms and predictive systems, it is essential to consider the regulations and ethical guidelines governing this field. The work is framed within the principles of the recent European Artificial Intelligence Act (AI Act) (European Parliament, 2024). In particular, the methodology used

for training, validating, and comparing machine learning models adopts as a quality standard the requirements established in Art. 10 (Data and Data Governance). This ensures that the data partitions (train and test) used to evaluate the algorithms were handled under practices designed to mitigate potential biases, preventing algorithmic manipulation of results and ensuring that performance metrics rigorously and objectively reflect statistical reality. Likewise, the integration of techniques such as SHAP values complies with the principles of interpretability and transparency required under Art. 50 of the regulation.

The use of this information to generate knowledge about urban competitiveness is also aligned with the European Data Governance Act (European Parliament, 2022), particularly with regard to the reuse of information for statistical research purposes (Art. 3). Regarding the direct data source, the reuse of the Observatory databases is carried out under the Creative Commons 4.0 BY license, which permits and supports their use for non-commercial academic studies while ensuring proper institutional attribution (Ondiviela, 2025).

At its core, the Observatory explicitly establishes an ethical commitment that guarantees the complete absence of manipulation or favoritism in the collection and processing of indicators. The strongest empirical evidence of this methodological rigor and data impartiality is reflected in the analytical results of this study: the latent clustering ("blind" and unsupervised) obtained through the K-Means model naturally converges with the qualitative city segmentation carried out by the Observatory's experts (Ondiviela, 2025). This mathematical convergence not only validates the integrity and neutrality of the original data source, but also confirms the transparency and technical responsibility applied throughout the project's entire analytical lifecycle.

Finally, the project has not been funded by external entities, which helps to ensure the independence of the analysis and reduces the risk of conflicts of interest. Academic integrity standards have been maintained throughout the development of the study, ensuring the traceability of all methodological decisions adopted (Resnik, 2020).

7.2. Project Value

From a scientific, technical, and social perspective, this work offers several significant contributions. Firstly, it contributes to the study of urban attractiveness through a quantitative, data-driven approach, making it possible to complement traditional analyses with predictive and modelling tools.

The use of machine learning models enables the identification of complex relationships between variables, facilitating a deeper understanding of the factors that influence urban competitiveness (Mehrabi et al., 2021). This type of analysis is particularly relevant in urban contexts, where multiple interrelated dimensions are involved.

From a practical perspective, the results of this project may prove valuable for decision-making in areas such as urban planning and economic development, providing an empirical basis for the ex-ante evaluation of public policies through scenario simulation (What-If analysis) (Wexler et al., 2019). Likewise, the project may help individuals and organizations

access more structured information to support decision-making regarding geographical location.

Finally, the potential integration of the results into open-access tools enhances information accessibility and promotes transparency in data analysis, aligning with the principles of open science and responsible governance of artificial intelligence (OECD, 2022).

7.3. SCOPE

The project is developed within the field of international urban data analysis, considering a set of cities over several years and a large number of indicators of diverse nature.

The main target audiences of the project include:

- Researchers and professionals in fields such as engineering, economics, social sciences, and urban planning.
- Public administrations interested in improving urban planning and management through the adoption of evidence-based Decision Support Systems (DSS) (Shulajkovska et al., 2024).
- Companies making decisions related to investment and location strategy.
- Citizens interested in understanding the factors that influence the attractiveness of cities.

The impact of the project may be manifested both directly, through the generation of knowledge, and indirectly, by influencing the interpretation of the factors that determine urban competitiveness (United Nations, 2019).

7.4. RESPONSIBILITY & IMPACT

From an ethical perspective, it is necessary to consider both the technical risks associated with this work and its social, economic, and environmental implications.

Firstly, one of the main risks lies in the potential presence of biases in the data used. Since the information sources may be predominantly based on Western contexts, there is a possibility that the model reflects only a partial view of urban attractiveness. Likewise, bias may also arise from the origin of the data if certain cities are evaluated under more favorable criteria.

Furthermore, machine learning models may incorporate or amplify patterns already present in the data, which could lead to biased conclusions if the results are not critically analyzed (Mehrabi et al., 2021). The complexity of these models may also hinder interpretability, reinforcing the need for rigorous analysis of feature importance through techniques such as SHAP values, which act not only as explanatory tools but also as empirical mechanisms for algorithmic auditing to detect discriminatory or illogical feature weights (Christoph Molnar, 2022).

From a social perspective, this work may have a positive impact by providing tools that enable a better understanding of the factors that make a city attractive. This can contribute to improved decision-making and urban planning. However, there is also a risk that an oversimplified interpretation of the results could lead to comparisons or rankings that fail to reflect the true complexity of urban environments. This project actively mitigates that risk by proposing latent groupings (clustering) rather than one-dimensional rankings.

In economic terms, this work may support decision-making related to investment and location strategy for both companies and public administrations. Regarding environmental impact, the analysis may indirectly contribute to promoting more sustainable development models by incorporating variables related to quality of life and environmental conditions (United Nations, 2019).

To mitigate the identified risks, several measures have been adopted, including the use of aggregated data, methodological transparency, critical analysis of results, and the treatment of interpretability as a core element of the project. This is further supported by the clustering performed using K-Means, which operates without prior knowledge of city identities, yet produces segmentation results consistent with expert classifications. Consequently, the objectives and results achieved can be considered fair, democratic, transparent, accessible to all, and free from social bias.

Risk	Description	Probability	Impact	Action
R1	Potential bias in the data due to the predominance of Western data sources	High	Medium	Critical use of data and analysis of limitations
R2	Incorrect interpretation of the results by non-specialist users	High	Medium	Clear explanation of the results and their limitations
R3	Amplification of biases by machine learning models	Medium	Medium	Continuous evaluation of results and use of interpretability techniques
R4	Misuse or oversimplified use of data (rankings, misleading comparisons)	Low	High	Contextualization of results and caution regarding their use
R5	Lack of representativeness of the dataset for certain cities or contexts	Low	High	Consideration of dataset scope and its limitations
R6	Dependence of result reliability on data quality	Low	Medium	Rigorous validation and cleaning of the data used

Tabla 7.1 Technical risk register of the project

		Impact		
		Low	Medium	High
Probability	Low		R6	R4, R5
	Medium		R3	
	High		R1, R2	

Tabla 7.2 Risk matrix

7.5. Ethical Synthesis

In relation to the above, and given the value presented by the work, as well as the fact that the identified risks are limited, identifiable, and manageable through responsible data usage and correct interpretation of the results, it can be concluded that the project is not only ethically viable but also advisable to carry out.

The work provides a rigorous quantitative approach to the study of urban attractiveness and contributes to a better understanding of this complex phenomenon, moving from a purely descriptive approach to a predictive and decision-support system (DSS) framework (Shulajkovska et al., 2024). However, it is essential that the results are interpreted with caution, taking into account the inherent limitations of both the data and the models used.

8. CONCLUSIONS

8.1. MAIN CONCLUSIONS

After the development, evaluation, application, and interpretation of the predictive system results, the following key conclusions are drawn in response to the research objectives.

Firstly, the project has empirically demonstrated that urban competitiveness is not a random phenomenon, but rather an ecosystem governed by modifiable mathematical rules. The ability of advanced algorithms such as Random Forest and XGBoost to achieve high explained variance metrics confirms that artificial intelligence is a mature and highly reliable tool for macroeconomic and sociological decision-making in urban environments.

Secondly, through the opening of the algorithmic "black box" using the SHAP analytical framework, the absence of a universal "urban formula" has been confirmed. The analysis revealed that each urban pillar responds to fundamentally different drivers. Investment in infrastructure and technology (Smart Cities) acts as the long-term driver of structural magnetism. Human attractiveness depends on reaching a critical threshold of purchasing power and social progress. Finally, corporate profitability is strictly conditioned by civil security and the classification of cities as low-risk environments for capital.

Finally, the transition from purely descriptive analytics to prescriptive intelligence represents the main technical milestone of this project. The development of the DSS has transformed a static predictive model into a dynamic analytical tool. By incorporating the temporal inertia of growing cities and assessing their resilience under adverse scenarios through stress testing, the simulator provides investors and urban planners with a safe, data-driven environment to experiment with the impact of future public policies before their real-world implementation.

8.2. FUTURE DEVELOPMENT OPPORTUNITIES

This Final Degree Project establishes a robust foundational architecture upon which several research and technical development lines can be opened in order to evolve the product toward a commercial production environment.

A first avenue of evolution would consist of overcoming the limitation of working with a static, annually updated dataset by incorporating real-time data ingestion. The integration of financial APIs or urban mobility metrics extracted from social networks would allow the simulator to dynamically adjust its projections with lower latency and improved responsiveness.

At a purely algorithmic level, the crisis simulation module could be enhanced by integrating Bayesian networks or causal graphs. This would remove the current univariate restriction imposed by the Ceteris Paribus assumption and enable the modelling of second-order or "domino" effects, where, for instance, a simulated collapse in urban safety would automatically degrade purchasing power in subsequent time steps, generating more complex and realistic macroeconomic scenarios.

Furthermore, from a user experience and business applicability perspective, the next logical step is the development of a dashboard-based interactive interface. This graphical layer would allow non-technical analysts to modify variables through visual controls and instantly observe the recalculated resilience profiles of each city.

Finally, in line with emerging global regulations and the increasing concern regarding climate change, the predictive ecosystem could be extended to include environmental sustainability as a fourth target dimension. Training the models to quantify the direct impact of carbon emissions or green transition policies on talent attraction would add an invaluable analytical layer for designing the cities of the future.

9. OTHER MERITS OF THE PROJECT

In addition to achieving the proposed objectives, this project has consistently aimed to be transparent and accessible, particularly for professionals and researchers interested in applying machine learning to urban planning and development. To this end, all source code related to exploratory data analysis, model training and evaluation, and the Decision Support System has been published in a GitHub repository. The repository is available at the following link: https://github.com/JaimeMLopez/TFG_JaimeMateoLopez_PrediccionAtractivoUrbano

Through this open approach and the use of open-source technologies, the knowledge generated is not only documented within the project report but can also be reused and adapted by others to incorporate new cities or macroeconomic variables. Likewise, the commitment to algorithmic interpretability through the use of SHAP values provides a key added value in today's context: ensuring that AI predictions are explainable and understandable to humans.

In this way, the written report not only accompanies the technical tool, but also provides a solid theoretical foundation that facilitates the integration of data science into disciplines such as sociology and urban economics.

10. BIBLIOGRAPHY

Biau, G., & Scornet, E. (2016). A random forest guided tour. Finance and Data Science, 1(1), 197–227. Consultado en marzo de 2026. https://www.sciencedirect.com/science/article/pii/S2210670720300019

Biecek, P., & Burzykowski, T. (2021). Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models. CRC Press. Consultado en mayo de 2026. https://doi.org/10.1201/9780429027192

Borgonovo, E. (2017). Sensitivity analysis: An introduction for the management scientist. Springer. Consultado en mayo de 2026. https://doi.org/10.1007/978-3-319-54672-8

Box, G. E. P., Jenkins, G. M., Reinsel, S. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5.ª ed.). John Wiley & Sons. Consultado en mayo de 2026. https://doi.org/10.1002/9781118619193

Bramer, M. (2020). Principles of data mining (4.ª ed.). Springer. Consultado en mayo de 2026. https://doi.org/10.1007/978-1-4471-7493-6

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. Consultado en mayo de 2026. https://doi.org/10.1023/A:1010933404324

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. O'Reilly Media. Consultado en mayo de 2026. https://www.oreilly.com/library/view/practical-statistics-for/9781492072935/

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). Consultado en marzo de 2026. https://doi.org/10.1145/2939672.2939785

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative. PeerJ Computer Science, 7, e623. Consultado en abril de 2026. https://peerj.com/articles/cs-623/

Chollet, F. (2021). Deep learning with Python (2.ª ed.). Manning Publications. Consultado en mayo de 2026. https://www.manning.com/books/deep-learning-with-python-second-edition

de Oliveira, J. R., de Mello, J. C. C. B. S., & de Sousa, A. F. (2021). Multidimensional sorting framework of cities regarding the smart and sustainable characteristics. Sustainable Cities

and Society, 75, 103296. Consultado en abril de 2026. https://www.sciencedirect.com/science/article/pii/S2210670721004716

European Commission. (2019). Ethics guidelines for trustworthy AI. Consultado en abril de 2026. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Union. (2016). General Data Protection Regulation (EU) 2016/679. Consultado en abril de 2026. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). O'Reilly Media. Consultado en mayo de 2026. https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125646/

Givisis, I., Vasilakos, A., & Theodoridis, S. (2025). Comparing explainable AI models: SHAP, LIME, and their practical differences. Electronics, 14(23), 4766. Consultado en abril de 2026. https://www.mdpi.com/2079-9292/14/23/4766

Glaeser, E. L. (2020). Urbanization and its discontents. Journal of Economic Perspectives, 34(3), 3–24. Consultado en abril de 2026. https://www.aeaweb.org/articles?id=10.1257/jep.34.3.3

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Consultado en mayo de 2026. http://www.deeplearningbook.org

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. Consultado en abril de 2026. https://doi.org/10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. (2021). The elements of statistical learning (2nd ed.). Springer. Consultado en abril de 2026. https://hastie.su.domains/ElemStatLearn/

Hettikankanamage, N., Perera, N., & Jayasinghe, U. (2025). Explainable artificial intelligence (XAI): A systematic review. Sensors, 25(21), 6649. Consultado en abril de 2026. https://www.mdpi.com/1424-8220/25/21/6649

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. Consultado en abril de 2026. https://doi.org/10.1109/MCSE.2007.55

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: principles and practice (3rd ed.). OTexts. Consultado en mayo de 2026. https://otexts.com/fpp3/

Jiang, P. (2025). Configural perspectives on urban talent ecology and competitiveness. Systems, 13(7), 499. Consultado en abril de 2026. https://www.mdpi.com/2079-8954/13/7/499

Joeres, R., Fulle, S., Bietz, S., & Zell, A. (2025). Data splitting to avoid information leakage with DataSAIL. Briefings in Bioinformatics, 26(3), bbaf143. Consultado en abril de 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC11978981/

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer. Consultado en mayo de 2026. https://link.springer.com/book/10.1007/b98835

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... & Jupyter Development Team. (2016). Jupyter Notebooks-a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas, 87–90. Consultado en abril de 2026. https://eprints.soton.ac.uk/403913/

Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. IEEE Access, 11, 31866–31879. Consultado en abril de 2026. https://ieeexplore.ieee.org/document/10081336

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. Consultado en mayo de 2026. https://doi.org/10.1007/978-1-4614-6849-3

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press. Consultado en abril de 2026. https://www.feat.engineering/

Ley Orgánica 3/2018. (2018). Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. Boletín Oficial del Estado, 294, de 6 de diciembre de 2018. Consultado en abril de 2026. https://www.boe.es/eli/es/lo/2018/12/05/3/con

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. Consultado en marzo de 2026. https://arxiv.org/abs/1705.07874

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Consultado en mayo de 2026. https://projecteuclid.org/euclid.bsmsp/1200512992

Martínez-Fernández, C., Audirac, I., Fol, S., & Cunningham-Sabot, E. (2020). Shrinking cities: Urban challenges of globalization. Sustainability, 12(1), 142. Consultado en marzo de 2026. https://www.mdpi.com/2071-1050/12/1/142

McKinney, W. (2010). Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 445, 51–56. Consultado en abril de 2026. https://doi.org/10.25080/Majora-92bf1922-00a

McKinney, W. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media. (Referencia definitiva para la unificación de archivos y limpieza de datos). Consultado en mayo de 2026. https://wesmckinney.com/book/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6). Consultado en abril de 2026. https://dl.acm.org/doi/10.1145/3457607

Miller, C., Varoquaux, G., & Foster, I. (2024). A review of model evaluation metrics for machine learning. PLOS Digital Health, 3(9), e0000532. Consultado en abril de 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC11420621/

Mohamadi, Z., Safari, S., Mirnezami, M., et al. (2025). The application of random forest-based models in clinical decision support: A review. Consultado en abril de 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC12315591/

Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Consultado en mayo de 2026. https://christophm.github.io/interpretable-ml-book/

Molnar, C. (2022). Interpretable machine learning. Consultado en abril de 2026. https://christophm.github.io/interpretable-ml-book/

OECD. (2022). OECD framework for the classification of AI systems. Consultado en abril de 2026. https://oecd.ai/en/classification

OECD. (2023a). Attracting talent: The (no longer) remote option for regions and cities. Consultado en abril de 2026. https://www.oecd.org/en/publications/attracting-talent-the-no-longer-remote-option-for-regions-and-cities_7c0941e1-en.html

OECD. (2023b). Smart city data governance. Consultado en abril de 2026. https://www.oecd.org/en/publications/smart-city-data-governance_e57ce301-en.html

Ondiviela, J. A. (2020a). Beyond SmartCities: How to create an attractive city for talented citizens. Kult-ur, 7(13), 205–231. Consultado en marzo de 2026. https://doi.org/10.6035/Kult-ur.2020.7.13.8

Ondiviela, J. A. (2020b). WorldWide Observatory for Attractive Cities. Universidad Francisco de Vitoria. Consultado en febrero de 2026. https://ddfv.ufv.es/entities/publication/10c837c3-f375-41ad-a936-b6ec60b49c53

Ondiviela, J. A. (2021). WorldWide Observatory for Attractive Cities. Universidad Francisco de Vitoria. Consultado en febrero de 2026. https://ddfv.ufv.es/entities/publication/3b093c1d-bd94-46ae-8b5a-369da8f4c5db

Ondiviela, J. A. (2022). WorldWide Observatory for Attractive Cities. Universidad Francisco de Vitoria. Consultado en febrero de 2026. https://ddfv.ufv.es/entities/publication/379b067b-b2db-4437-a62d-eae57ea7eef2

Ondiviela, J. A. (2023). WorldWide Observatory for Attractive Cities. Universidad Francisco de Vitoria. Consultado en febrero de 2026. https://ddfv.ufv.es/entities/publication/000c2e8b-6e8a-4f7b-a204-42c694d91587

Ondiviela, J. A. (2024). WorldWide Observatory for Attractive Cities. Universidad Francisco de Vitoria. Consultado en febrero de 2026. https://ddfv.ufv.es/entities/publication/adf43c41-c1f2-4953-b663-9bf0ba07a894

Ondiviela, J. A. (2025). WorldWide Observatory for Attractive Cities. Universidad Francisco de Vitoria. Consultado en febrero de 2026. https://ddfv.ufv.es/entities/publication/bbbc7cd0-2c23-4729-be99-538dd73f9aff

Parlamento Europeo y Consejo de la Unión Europea. (2022). Reglamento (UE) 2022/868 relativo a la gobernanza europea de datos y por el que se modifica el Reglamento (UE)

2018/1724 (Ley de Gobernanza de Datos). Diario Oficial de la Unión Europea, L 152. Consultado en abril de 2026. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32022R0868

Parlamento Europeo y Consejo de la Unión Europea. (2024). Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial). Diario Oficial de la Unión Europea, L 2024/1689. Consultado en abril de 2026. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=OJ:L_202401689

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. Consultado en abril de 2026. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media. Consultado en mayo de 2026. https://www.oreilly.com/library/view/data-science-for/9781449374273/

Python Software Foundation. (2024). Python 3 Reference Manual. Consultado en abril de 2026. https://docs.python.org/3/reference/

Resnik, D. B. (2020). What is ethics in research & why is it important? Consultado en abril de 2026. https://www.niehs.nih.gov/research/resources/bioethics/whatis

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. Consultado en mayo de 2026. https://doi.org/10.1038/s42256-019-0048-x

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Global sensitivity analysis: The primer. John Wiley & Sons. Consultado en mayo de 2026. https://doi.org/10.1002/9780470725184

Sarker, I. H. (2021). An overview of supervised machine learning approaches and their applications. Energies, 16(16), 5972. Consultado en abril de 2026. https://www.mdpi.com/1996-1073/16/16/5972

Shulajkovska, M., Gusev, M., & Tashkoski, P. (2024). Artificial intelligence-based decision support system for sustainable urban mobility. Electronics, 13(18), 3655. Consultado en abril de 2026. https://doi.org/10.3390/electronics13183655

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). Introduction to data mining (2nd ed.). Consultado en abril de 2026. https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley. Consultado en mayo de 2026. https://www.researchgate.net/publication/256802405_Exploratory_data_analysis

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. Consultado en abril de 2026. https://unesdoc.unesco.org/ark:/48223/pf0000381137

United Nations. (2019). World urbanization prospects. Consultado en abril de 2026. https://population.un.org/wup/

Waskom, M. L. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. Consultado en abril de 2026. https://doi.org/10.21105/joss.03021

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2019). The What-If Tool: Interactive probing of machine learning models. IEEE Transactions on Visualization and Computer Graphics, 26(1), 56–65. Consultado en abril de 2026. https://ieeexplore.ieee.org/document/8807255

Wiens, M., Kreienkamp, A., & Döring, M. (2025). A tutorial and use case example of the eXtreme Gradient Boosting algorithm: XGBoost. Consultado en abril de 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC11895769/

Zhang, D., & Zhou, L. (2021). Comparison of machine learning models for predictive analytics. Information Sciences. Consultado en marzo de 2026. https://www.sciencedirect.com/science/article/pii/S1386505621001106

ANNEX A: LIST OF ANALYZED CITIES

This annex details the set of cities that make up the Observatory's database used throughout this Final Degree Project. These cities represent the geographical, socio-economic, and cultural diversity on which the regression and machine learning models have been trained, and against which the DSS has been validated. In this way, the reader is provided with a comprehensive overview of the global scope of the research and the real-world environment in which the developed predictive algorithms operate.

City	Country
Buenos Aires	Argentina
Córdoba	Argentina
Sydney	Australia
Melbourne	Australia
Adelaide	Australia
Canberra	Australia
Vienna	Austria
Linz	Austria
Manama	Bahrain
Minsk	Belarus
Brussels	Belgium
Antwerp	Belgium
La Paz	Bolivia
Brasilia	Brazil
Sao Paulo	Brazil
Rio de Janeiro	Brazil
Sofia	Bulgaria
Vancouver	Canada
Toronto	Canada
Ottawa	Canada
Montreal	Canada
Santiago	Chile
Shanghai	China
Beijing	China
Guangzhou	China
Shenzhen	China
Chengdu	China
Chongqing	China
Shenyang	China
Wuhan	China
Suzhou	China
Tianjin	China
Harbin	China
Medellín	Colombia
Bogota	Colombia
San José	Costa Rica
Zagreb	Croatia
Prague	Czech Republic
Copenhagen	Denmark
Aarhus	Denmark
Santo Domingo	Dominican Republic
Quito	Ecuador
Cairo	Egypt
Tallinn	Estonia
Helsinki	Finland
Tampere	Finland
Espoo	Finland
Oulu	Finland
Paris	France
Lyon	France
Marseille	France
Nice	France
Bordeaux	France
Lille	France
Tbilisi	Georgia
Berlin	Germany

Munich	Germany
Dusseldorf	Germany
Frankfurt	Germany
Hamburg	Germany
Stuttgart	Germany
Cologne	Germany
Accra	Ghana
Athens	Greece
Hong Kong	Hong Kong
Budapest	Hungary
Mumbai	India
Bangalore	India
New Delhi	India
Hyderabad	India
Jakarta	Indonesia
Dublin	Ireland
Tel Aviv	Israel
Jerusalem	Israel
Milan	Italy
Rome	Italy
Florence	Italy
Torino	Italy
Tokyo	Japan
Yokohama	Japan
Osaka	Japan
Nagoya	Japan
Kuwait City	Kuwait
Riga	Latvia
Vilnius	Lithuania
Luxembourg	Luxembourg
Kuala Lumpur	Malaysia
Mexico City	Mexico
Monterrey	Mexico
Guadalajara	Mexico
Casablanca	Morocco
Rabat	Morocco
Amsterdam	Netherlands
Eindhoven	Netherlands
Rotterdam	Netherlands
Den Haag	Netherlands
Auckland	New Zealand
Wellington	New Zealand
Oslo	Norway
Bergen	Norway
Stavanger	Norway
Panama City	Panama
Asuncion	Paraguay

Lima	Peru
Manila	Philippines
Warsaw	Poland
Wroclaw	Poland
Lisbon	Portugal
Porto	Portugal
Doha	Qatar
Bucharest	Romania
Moscow	Russia
St Petersburg	Russia
Riyadh	Saudi Arabia
Belgrade	Serbia
Singapore	Singapore
Bratislava	Slovakia
Ljubljana	Slovenia
Cape Town	South Africa
Durban	South Africa
Johannesburg	South Africa
Seoul	South Korea
Barcelona	Spain
Madrid	Spain
Málaga	Spain
Valencia	Spain
Bilbao	Spain
Zaragoza	Spain
Santander	Spain
Seville	Spain
Stockholm	Sweden
Gothenburg	Sweden
Malmo	Sweden
Zurich	Switzerland
Geneva	Switzerland
Bern	Switzerland
Basel	Switzerland
Taipei	Taiwan
Bangkok	Thailand
Tunis	Tunisia
Istanbul	Turkey
Ankara	Turkey
Kiev	Ukraine
Dubai	United Arab Emirates
Abu Dhabi	United Arab Emirates
London	United Kingdom
Edinburgh	United Kingdom
Birmingham	United Kingdom

Liverpool	United Kingdom
Manchester	United Kingdom
Belfast	United Kingdom
Bristol	United Kingdom
Nottingham	United Kingdom
Glasgow	United Kingdom
San Francisco	United States
Boston	United States
New York City	United States
Washington, D.C.	United States
Chicago	United States
Seattle	United States
Los Angeles	United States
Baltimore	United States

Philadelphia	United States
Dallas	United States
Phoenix	United States
Houston	United States
Atlanta	United States
Miami	United States
Denver	United States
Las Vegas	United States
Kansas City	United States
Honolulu	United States
Montevideo	Uruguay
Ho Chi Minh City	Vietnam
Hanoi	Vietnam

ANNEX B: DETAILED COMPOSITION OF URBAN CLUSTERS

This annex presents the complete breakdown of the 175 cities analysed in the project, classified according to the results obtained using the unsupervised learning algorithm K-Means, as detailed in Section 5.1.

The assignment of each city to one of the four "urban tribes" has been determined exclusively through mathematical criteria, after projecting the original 60 predictive variables into a Principal Component Analysis (PCA) feature space. This list allows the geographic and socio-economic cohesion of the clusters to be verified, serving as a reference basis for the comparative analyses of performance and profitability conducted in the subsequent phases of the study.

The cities are presented grouped by cluster:

- **TRIBE 1 (Total: 32 cities):**
  Accra, Asuncion, Bangalore, Bogota, Brasilia, Buenos Aires, Cairo, Cape Town, Casablanca, Córdoba, Durban, Guadalajara, Ho Chi Minh City, Hyderabad, Jakarta, Johannesburg, La Paz, Lima, Manila, Medellín, Mexico City, Monterrey, Mumbai, New Delhi, Panama City, Quito, Rabat, Rio de Janeiro, San José, Santo Domingo, Sao Paulo, Tunis

- **TRIBE 2 (Total: 88 cities):**
  Aarhus, Adelaide, Amsterdam, Antwerp, Athens, Auckland, Barcelona, Basel, Belfast, Bergen, Berlin, Bern, Bilbao, Birmingham, Bordeaux, Bristol, Brussels, Budapest, Canberra, Cologne, Copenhagen, Den Haag, Dubai, Dublin, Dusseldorf, Edinburgh, Eindhoven, Espoo, Florence, Frankfurt, Geneva, Glasgow, Gothenburg, Hamburg, Helsinki, Lille, Linz, Lisbon, Liverpool, Ljubljana, London, Luxembourg, Lyon, Madrid, Malmo, Manchester, Marseille, Melbourne, Milan, Montreal, Munich, Málaga, Nagoya, Nice, Nottingham, Osaka, Oslo, Ottawa, Oulu, Paris, Porto, Prague, Riga, Rome, Rotterdam, Santander, Seoul, Singapore, Stavanger, Stockholm, Stuttgart, Sydney, Tallinn, Tampere, Tokyo, Torino, Toronto, Valencia, Vancouver, Vienna, Vilnius, Warsaw, Wellington, Wroclaw, Yokohama, Zaragoza, Zurich

- **TRIBE 3 (Total: 18 cities):** Atlanta, Baltimore, Boston, Chicago, Dallas, Denver, Honolulu, Houston, Kansas City, Las Vegas, Los Angeles, Miami, New York City, Philadelphia, Phoenix, San Francisco, Seattle, Washington D.C.

- **TRIBE 4 (Total: 37 cities):** Abu Dhabi, Ankara, Bangkok, Beijing, Belgrade, Bratislava, Bucharest, Chengdu, Chongqing, Doha, Guangzhou, Hanoi, Harbin, Hong Kong, Istanbul, Jerusalem, Kiev, Kuala Lumpur, Kuwait City, Manama, Minsk, Montevideo, Moscow, Riyadh, Santiago, Shanghai, Shenyang, Shenzhen, Sofia, St Petersburg, Suzhou, Taipei, Tbilisi, Tel Aviv, Tianjin, Wuhan, Zagreb