Giani-Ionel GRĂDINARU, PhD
giani.gradinaru@csie.ase.ro
Bucharest University of Economic Studies, Bucharest, Romania
National Institute of Economics Research, Bucharest, Romania

Daniela-Ioana MANEA, PhD (corresponding author)
daniela.manea@csie.ase.ro
Bucharest University of Economic Studies, Bucharest, Romania
National Institute of Economics Research, Bucharest, Romania

Florin ANDREESCU, PhD Candidate
florin.siriusresearch@gmail.com
Bucharest University of Economic Studies, Bucharest, Romania

Darius-Andrei TOMA
toma.darius20@stud.ase.ro
Bucharest University of Economic Studies, Bucharest, Romania

Liana Ioana PARASCHIV, PhD Candidate
lianaioanaparaschiv@gmail.com
Bucharest University of Economic Studies, Bucharest, Romania

Identifying the Main Factors of Elaborating "Smart City" Strategy Using Machine Learning. A Comparative Study Among Romanian Cities

Abstract. *The paper aims to identify the essential factors for developing a "smart-city" strategy at the level of local public administration of county residences. Furthermore, this research also identifies structural differences between Bucharest and county residences in Romania. The main method used was the Random Forest classification algorithm, which is based on the idea of building decision trees for classifying statistical units. Also, in order to identify the structural differences between Bucharest and the county residences, exploratory analysis was used. The results of this research show that large cities tend to develop the "smart city" strategy more easily, the algorithm classifies with a low error rate the cities that have developed a "smart city" strategy, which means that essential factors in developing this type of strategy can be identified very easily. Furthermore, major differences were identified between Bucharest and the county residences; Bucharest is better performing on almost all areas of interest analysed in the research. The originality of the article lies in the use of a machine learning algorithm to identify influencing factors in the process of developing a strategy at the level of local public administration.*

Keywords: *smart city, urban development, sustainability, machine learning*, *Deep Learning model, quantitative research*.

JEL Classification: O18, R20, R23.

DOI: 10.24818/18423264/58.3.24.04

1. Introduction

Urban development is one of the determining factors in ensuring sustainable development, the Sustainable Development Strategy having as one of the objectives SDG 11 Sustainable cities and communities. Within objective 11 of the Sustainable Development Strategy, an important element in reaching the target is represented by the “smart-city”.

According to the European Commission, by 2050 the percentage of people in Europe living in urban areas is estimated to reach around 84%, which represents an increase from the current level of over 70% (European Commission, 2019). While urban areas can serve as centres for innovation and contribution to regional sustainability, there are also new, interconnected issues that challenge the resilience of urban socio-environmental systems (Erős et al., 2022).

Therefore, the research aims to identify the factors for the development of a “smart-city” strategy at the level of county seat cities in Romania. In addition, this research also analyses the structural differences between the Municipality of Bucharest and the other 41 county seat cities. To achieve the main objectives of the document, the Citadini data source was identified, an online platform developed by the Ministry of Development, where different state institutions report statistical indicators at the city level of cities in Romania. The variable of interest used is the existence of a “smart-city” policy at the local administration level (1-Yes/0-No).

Subsequently, the factors that significantly influence the development of a “smart-city” policy were analysed. In order to identify factors that influence the development of a “smart-city” policy, a classification method based on the concept of supervised learning, Random Forest, was used. The original element of the work is represented by the use of this machine learning algorithm in identifying the determining factors for the development of the „smart-city” strategy.

In the Literature Review section, the theoretical framework from which the research starts is outlined, being a defining element of the development of quantitative analysis. Later, in the Methodology and Data section, the methods used are briefly described, as well as the robustness of their results and the entire data collection process, as well as their processing.

Furthermore, in the Results and Discussions section, the main results of using the Random Forest algorithm are analysed from an economic and social point of view. Also, in this section, the comparative analysis between Bucharest and the other county residences is carried out. Finally, in the section related to the conclusions, the main conclusions of the study, the limits of the research, and the directions for the development of the research are presented.

2. Literature review

The quality of the urban environment can be negatively affected by environmental pollution and waste that can have a negative impact on public health (Meerow et al., 2016). Furthermore, Lewis et al. (2022) argue that the green infrastructure of urban and peri-urban areas is capitalised for its essential role in stimulating social and environmental resilience, by ameliorating the effects of extreme climate variations on urban society for its contributions to stimulating social capital by providing an authentic space for human and human-nature interactions, and for its role in biodiversity by providing permanent or transient habitats and connecting green corridors for wildlife (Xu et al., 2022).

To address these various issues, urban systems must undergo sustainability reforms. Both at the level of governing policy and in a more general social context, this presents a dilemma (Song et al., 2019). Urban transformations, which are fundamental and irreversible changes in various dimensions such as agency configurations, infrastructure, lifestyles, innovation, governance, and ecosystems, are more precisely described as sustainability transformations, which refer to a series of changes in social-technological-environmental interactions and feedback to promote a resilient and secure system (Hölscher et al., 2018). The incorporation of ideas that address various aspects of sustainability in the socio-medio-social context into significant formal documents of urban governance represents a significant lever for changes in urban sustainability, because as a result of this integration, sustainability ideas can acquire strategic importance for the development of the regional urban environment (Erős et al., 2022). Also, Eros et al. (2022) argues that urban areas in eastern European countries share many features with Western European urbanisation styles, but also show symptoms of "path dependency", i.e., they still strongly reflect the spatial planning and governing culture of the communist-socialist regime. These differences result from the economic development model of the countries from which the cities originate, the countries of Western Europe, which were not part of the communist bloc, being more concerned with the impact generated by economic activity than the countries of the former communist bloc (Gradinaru and Maricut, 2023).

Regarding Romania, "How to transform a city in Romania into a regional model of sustainable development or, better said, into a sustainable city?" is a question that has become increasingly popular. Thus, Romania has not had a sustainable city concept for a very long time due to the fact that scientists and administrative entities have only recently begun to explore this idea. However, there have been smart initiatives that have expanded rapidly, and even though the first projects were only launched in 2014, there are now hundreds spread across the nation (Zaman et al., 2021). Most of these initiatives were launched in the large regional nodes and in the capital of Romania, Bucharest, but several rural communities, such as Ciugud and Luncăvița, came up with creative ideas that quickly attracted the attention of the media. From the beginning, these projects were criticised because they failed to affect in any way the economy of the neighbourhoods, being often difunctional, and require a much too long period of implementation (Ibanescu et al., 2022).

Along with the notion of "sustainable city", the subject of "intelligent city" or "smart city" can also be approached. According to Chourabi and Nam (2012) but also Yu and Xu (2018), the application of technology-based solutions to improve citizens' interaction with government and to promote sustainable development is what is commonly understood by “smart cities”. When social, environmental and economic development factors are balanced and connected through decentralised mechanisms to more efficiently manage important urban assets, resources, and flows for real-time activities, a city is said to be smart (Ismagilova et al., 2020). To support social and urban interconnectedness through increased citizen interaction and government efficiency, smart cities are designed around an ICT-based infrastructure with sensor technology compatible with the “Internet of Things” (IoT) (Kumar et al., 2020).

Numerous cities around the world have embraced the smart mindset and are either actively pursuing plans to modify their current assets and networks, or have built their infrastructure to support this new status. Yan et al. (2023) argues that to enable the various elements of smart cities to cooperate and interact with the network architecture, smart city designers use contemporary technologies such as mobile cloud computing, electronic objects, networks, sensors, and machine learning technologies. Governments and regional authorities face major political, regulatory, and technical hurdles as a result of the changing inherent complexity of current infrastructure and the new forms of citizen participation that are required (Gkontzis et al., 2024). Data processing and management is one of the major obstacles to creating smart cities. This refers to the data that already exists in the city's databases, as well as their connection with the new technologies and sensors that are part of the smart city, which has an impact on security and privacy (Van Zoonen, 2016). Threats to information security, data privacy, and cyber issues, where unauthorised access to information can have negative effects, highlight the importance of addressing these issues at an early stage in the design and development of smart cities (Elmaghraby and Losavio, 2014).

3. Data & Methodology

In order to identify the factors influencing the development of a "smart-city" policy, the Random Forest method is used. According to James et al. (2014), random forests provide an improvement over bagged trees by means of a random small tweak that decorrelates the trees. As in bagging, it builds a number forest of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$ that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (Scornet et al., 2015). The main difference between bagging and random forests is the choice of predictor subset size m. For instance, if a random forest is built using m = p, then this amounts simply to bagging. Random forests using $m = \sqrt{p}$ leads to a reduction in both test error and OOB error over bagging. Using a small value of m in building a random forest will typically be helpful when we have a large number of correlated predictors (Scornet et al., 2015).

In order to make a detailed analysis of significant cities in Romania, out of a number of 319 cities, only 41 cities were included in the analysis, these being the residences for each county in Romania, together with the country's capital, Bucharest. Following the exploratory analysis of the data, it was noticed that Bucharest is a very large and developed city, with a vast population; its corresponding values are much different from those of the rest of the cities. Thus, Bucharest was treated as an atypical case, being analysed separately from the rest of the 41 county residences.

The data were extracted from the Citadini database, having as a moment of reference the year 2018.

Table 1. The description of variables

Acronym	Definition	Unit of measure
Nmtc	The number of public transport vehicles	Number
Lpb	Bicycle track length	Km
Chelt_Loc_Sociale	Housing expenses, services and public development	RON per capita
Tot_Chelt_Dez	Total expenses section development	RON
Unemploymnet	Unemployment rate	%
Ncipfx≥30Mbps	Number of internet connections fixed points ≥30Mbps	Thousands
Regen	Power consumption of renewable electricity	MWh
Cpmed	Environmental protection expenditure	RON per capita
Green_Space	Green space areas	Square meters
AQI	Index of air quality	%
Total_Pop	Population	number
Waste_per_capita	Total amount of waste per capita	Tonnes per capita
Water	Share of dwellings equipped with water supply installations	%
Sewerage	Share of dwellings connected to the sewerage system	%
Vtgd	Total volume of distributed gases	Thousand cubic meters
Smart_City	Is there in place a smart-city strategy?	1 - yes / 0 – nu

Source: own work using Citadini data.

Apart from the variable "Smart-city", which is a dichotomous variable and answers the question "Does this city have a strategy on the smart-city area?" (Yes/No)", the rest of the variables are quantitative and can be used as predictor variables in the Random Forest algorithm, while the "Smart-City" variable will be used as a variable to classify cities into cities that have developed a "smart city" strategy (1) and cities that have not adopted a "smart city" strategy (0) (Table 1).

Next phase: The Deep Learning analysis

Here, we extend the first set of variables used in the Random Forest algorithm, with the other 29 additional variables extracted from Citadini. In this way, the research was developed by increasing the number of statistical units included in the sample, but also increasing the number of variables used in the classification of cities to identify the areas with the greatest impact in the process of developing "smart-city" strategies (Table 2).

Table 2. The description of variables

Acronym	Definition	Unit of measure
Pop_Density	Population density = The ratio between the number of inhabitants and the area of the locality	Number
Employee	Number of Employee	Number
Empl_rate_20_64	Employment rate of population (20 to 64 years)	%
Pgdi	Share of industrial gases	%
Rate_02_19	The Dynamics of the number of Inhabitants 2002-2019	%
Constr_permits_per_1000	Construction Permits Per 1000 Inhabitants	Number
Pmod_str	Share of Modernized Streets	%
D_urban_str	Density of Urban Streets	Number
IDX_Tourism	Tourist Attraction Index	Number
Nr_of_turists_per_UAT	Intensity of Tourist Traffic	Number
T_to_european_road	Time to European Road	Minutes
T_Hospital	Time to the Emergency Hospital	Minutes
T_industrial_park	Time to Industrial Park	Minutes
Ppriv_mtra	Percentage of Private Motorized Transport	%
Pnmtra_mq	Non-Motorized Transport Modal Quota	%
C_mq	Car Modal Quota	%
PT_mq	Public Transport Modal Quota	%
Eptr_bin	Existence of Public Transport	1 - yes / 0 – nu
Erail_st_bin	Existence of Railway Transport Station	1 - yes / 0 – nu
UMP_av_bin	Urban Mobility Plan availability	1 - yes / 0 – nu
P_homes_electr	Percentage of Homes with Electricity	%
Nr_sewerage	Number of homes with sewage system	Number
Nr_water	Number of dwellings with water supply	Number
Nr_pers_per_room	Housing density = the number of people per living room	Number
Water_cons_per_cap	Water Consumption Per Inhabitant	Cubic meters/capita
NrB_per_1000	Entrepreneurial Density = Number of businesses/1,000 inhabitants	Number
GDP_RON_per_capita	GDP per capita	RON per capita
SEP_av_bin	Availability of the Sustainable Energy Plan	1 - yes / 0 – nu
GS_sqm_per_cap	Green space square meters per inhabitant	Square meters / inhabitant

Source: own work using Citadini data.

In this phase, we used the total set consisting now of 46 variables to train a Deep Learning model and to evaluate the factors influencing the development of a "smart-city" policy. We also extended the number of cities analysed from 41 to all 319 cities. Bucharest was considered even here as an atypical case and it was excluded from the training data.

The model was defined with a 3 layers architecture:

Layer 1: 46 neurons and the activation function: ReLU

Layer 2: 92 neurons and the activation function: ReLU

Layer 3: 41 neurons and the activation function: Sigmoid

Optimiser: ADAM

Loss: ‘binary crossentropy’

Metris: 'accuracy'

From the training data set, 20% of the records were reserved for the test set and another 20% for the validation set.

Deep Learning presentation - the neural networks architecture and the learning process.

ANN architecture contains neurons, layers, and weights.

Neuron: receives input data and applies a weighted sum and an activation function. Produces an output value; 3 types of layers are defined: input, output, and hidden.

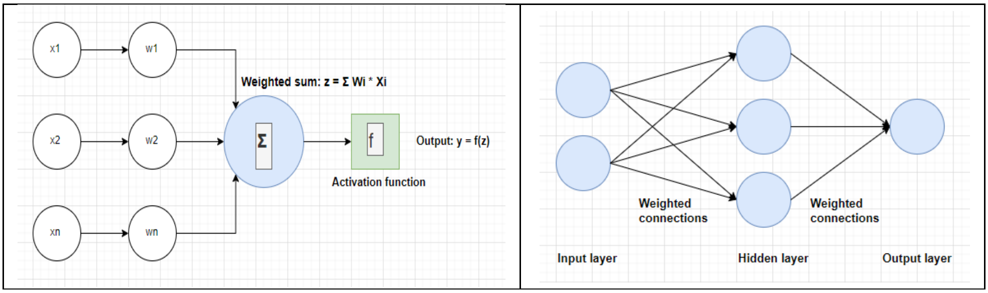


Figure 1. Neuron & Layers

Source: own work using draw.io.

The learning process can be defined as iteratively adjusting of all network weights:

Forward propagation – an input set produces an output value (predicted value)

Backpropagation – you calculate the error of the network – real output vs expected output using cost or loss function.

The cost function depends on the network. We need to optimise it to minimize the cost using a method called ‘gradient descent’. This uses small increments towards the negative gradient to reach the minimum cost value. Hyperparameters are important fixed variables in the learning process that can be tuned to improve network performance. Those values are changed, then the learning process is done again, for example:

- Number of layers and the neuron count.
- Learning rate

- Activation functions (Linear, Sigmoid, Tanh, ReLU)
- Cost functions: Mean Squared Error (MSE) – used for regressions and Cross-entropy (or log loss) – used for classification tasks.
- Epocs - the number of learning passes used to optimise the function
- Batch size – number of samples from the training set used in a complete forward/backward cycle.

Optimisers, those update the value of the weights to minimise the loss value. Examples: Momentum, AdaGrad (Adaptive Gradient Algorithm), RMSProp (Root Mean Square Propagation), and Adam (Adaptive Moment Estimation).

SHAP presentation

Starting from Shapley values in cooperative game theory, the SHAP explanation framework is used to compute and draw those values for further analysis. Finally, the SHAP explanation measures the influence feature X has on the outcome by computing a score for each feature averaged over all possible contexts.

SHAP (Shapley Additive Explanations) is a framework used for explainable AI. Shapley values measure a feature's contribution to a machine learning model's output. We imagine the features as players cooperating to make a prediction. The value computed for a feature is the difference between the model's prediction with and without that feature, averaged across all possible feature combinations.

If we consider all features subsets $S \subseteq FS \subseteq F$, where F is the set of all features, this method involves trying different combinations of features to see which ones contribute the most. For each feature, you train two models: one with that feature and one without. Then you compare their predictions to see how much difference that feature makes.

The model trained with the feature number i is: $f_{S\cup\{i\}} fS \cup \{i\}$ and the model trained without that feature will be: $f_S fS$.

Then, predictions from the two models are compared for a set S and S $\cup$ $\{i\}S \cup \{i\}$ of input features, $_{xSx}S$ and $_{xS\cup\{i\}x}S \cup \{i\}$ are the values of features in the two sets. The difference is:

$$fS\cup\{i\}(xS\cup\{i\}) - fS(xS) fS\cup\{i\}(xS\cup\{i\}) - fS(xS) \quad (1)$$

By summing up these differences across all possible feature combinations, you can assess the overall importance of each feature. This helps you decide which features are most relevant for your model as follows:

$$\Phi_t = \frac{1}{|F|} \sum_{S\subseteq F\setminus\{t\}} \binom{|F|-1}{|S|}^{-1} \left[f_{S\cup\{t\}}\left(x_{S\cup\{t\}}\right) - f_S(x_S)\right], \quad (2)$$

with $|S||S|$ and $|F||F|$ being the cardinalities of S and F.

Shapley values consider all possible feature combinations to assess each feature's importance. To account for varying group sizes, the contribution is divided by the number of features in each group and the total number of features. This ensures a fair comparison regardless of model complexity. These contributions can be used to explain the model's output. A linear function of the feature contributions will define the *Explanation Model* g:

$$g(z) = \phi_0 + \sum_{t=1}^{M} \phi_t z_t \quad (3)$$

Here $\phi 0 \phi 0$ is the SHAP value equal to the average of the samples' outcomes, $z_i z_i$, are binary variables with $z_i \in \{0,1\}, z_i \in \{0,1\}$, M is the number of input features.

SHAP Values

SHAP values are based on the classical Shapley values and work even with missing data (unlike most models). They estimate missingness by averaging over random replacements (Lundberg and Lee). Each SHAP value (a single number per feature) tells us the feature's influence on a specific sample's output. Positive values mean that the feature pushes the output in that direction, while the magnitude reflects its impact. The model's prediction will always equal the sum of SHAP values for a sample. The Explain Like I'm Five (ELI5) library for Python empowers researchers and developers to delve into the inner workings of diverse machine learning models (Korobov and Lopuhin, 2021). It offers a standardised interface (API) to visualise and debug these models, simplifying the process. This allows users to understand how models make predictions based on the data they are trained on. ELI5 provides valuable information for both model refinement and clear communication of model behaviour to a broader audience.

4. Results and comments

To answer the first research question "What are the factors that influence the development of a smart city strategy at the level of the local administration of the cities included in the sample?" the results of the Random Forest algorithm were analysed. To minimise the classification error, the algorithm was optimised by running a number of 500 decision trees.

Table 3. Classification matrix

Category	0	1	Classification Error
0	1	7	0.875
1	1	32	0.030

Source: own work using RStudio.

Following the application of the classification algorithm, it can be observed that 33 statistical units are correctly classified, while eight of them are incorrectly classified, resulting in a model classification error of 19.51%. Interestingly, the classification error for cities that have not developed a "smart-city" type strategy is high, approximately 87.5%, while for cities classified in category 1, which have formulated a "smart-city" strategy, which means that the results are robust for the analysis of influencing factors in the strategy development process in the "smart-city" area (Table 3).

To better understand the classification matrix, the main results of the classification can be summarized as follows:

- True positive values: The model predicted 32 cities as having a "smart-city" strategy (1), they applying a "smart-city" strategy (1);

- False positives: The model predicted 7 cities as having a "smart-city" strategy (1), but they do not adopt a "smart-city" area strategy (0);
- False negative values: The model predicted a single city as not applying a "smart-city" type strategy (0), but it was using this type of strategy (1);
- True negative values: The model predicted only one city as not using the "smart-city" type strategy (0), as it did not develop a strategy for this area of interest (0).

Considering the previously discussed elements, it can be stated that the model has a very high accuracy in predicting cities that are sustainable and a very low accuracy in predicting cities that are not sustainable. Further, to answer the first question of the research, the value of the "Mean Decrease Gini" attribute will be analyzed, which presents us with the importance of each predictor in the classification of statistical units through the Random Forest algorithm. Thus, the higher the value of the Mean Decrease Gini attribute, the more important that variable is in developing the "smart-city" strategy.

Table 4. Mean Decrease Gini values

Variable	Mean Decrease Gini
Nmtc	0,6462934
Lpb	1,206981
Chelt_Loc_Sociale	0,8015832
Tot_Chelt_Dez	0,6127595
Pond_Someri	0,3822373
Ncipfx≥30Mbps	1,2381306
Regen	0,368126
Cpmed	0,6161484
Green space	0,4678207
AQI	1,4479003
Total_Pop	0,5557452
Waste/Capita	0,5995595
Water	1,0622164
Sewerage	1,2907537
Vtgd	0,9126902

Source: own work using Rstudio.

Interpreting the results obtained, it is observed that among the most important variables for analysis are the specific air quality index (*AQI* with a value of approximately 1.45), conventional homes that have a sewage system in the home (*Sewerage* with a value of 1.29), the number of internet connections with a speed higher than 30Mbps (*Ncipfx≥30 Mbps* with a value of 1.24), the length of cycling tracks (*Lpb* with 1.21) and homes equipped with water supply facilities (*Water* with a value of 1.06) (Tabel 4).

Moreover, the least important variables that contribute to a very small extent to the classification of the trees are the share of the unemployed (*Pond_Someri* with a value of approximately 0.38) and the consumption of renewable energy (*Regen* with a value of 0.37).

Thus, from the results obtained from the Random Forest algorithm, we can propose some sustainability policies to further develop the county residences in Romania. Some of these would be: improving internet connectivity, improving air quality, infrastructure development, and sustainable management of hydrological resources. These apples involve investments in the area of digitalization, the development of public policies for the large-scale use of energy from renewable sources, the increase of investments and the inclusion of the population in the sphere of the sewage system and water purification.

The differences between Bucharest and the other cities

The city of Bucharest in the context of urban development and sustainable cities in Romania, represents a special case due to the difference in urban development between it and the rest of the cities in Romania, in this case the county residences. Being the capital of Romania, a large city with an equally large population, it had a "special treatment" when it comes to the urban development of cities. Thus, the main desire was the transformation of Bucharest into a sustainable city, being the city with the highest probability of achieving this goal given the fact that it has an infrastructure specific to the capital of a country, much more developed technologies that are not found in the same preponderance in other cities, and different living conditions.

Table 5. The differences between Bucharest and the Average value of smart-city

Variable	Bucharest	Average value for smart-city
Nmtc	1.899,00	115,97
Lpb	177,13	16,79
CLS	366,67	236,59
TCD	860.592.050,86	47.604.415,98
Unemployment rate	1,10	0,70
Ncipfx≥30Mbps	718,20	44,98
Regen	754,341	53,63
Cpmed	396,525	120,04
Green Space	45.060.000,00	3.387.575,76
AQI	50,58	39,21
Total_Pop	2.131.034,00	160.217,12
Waste/Capita	0,77	0,53
Water	96,80	95,99
Sewerage	96,60	95,47
Vtgd	2.937.904,00	98.697,52

Source: own work using RStudio.

In the table 5, very large differences can be observed between the average values of the county residences, considered sustainable, in Romania, and the values of the variables for Bucharest. Thus, the Municipality of Bucharest is very different from the rest of the cities and has already reached a high level of development and urbanization compared to the rest. However, the capital of Romania has only one deficiency, represented by the high percentage of unemployed with a value of 1.10% compared to only 0.70%, on average, for the other county residences.

This fact is also given by the capital's very large population of approximately two million people residing in Bucharest, but a discussion can arise regarding the type of unemployment existing in the capital, as it is very likely that the active population in Bucharest will requests a higher salary level compared to the rest of the county residences (Table 5).

Although the Municipality of Bucharest surpasses any other city in this analysis, the capital of Romania could be a good example to follow and study in the years to come in order to observe the specific behaviours and models of a prosperous city, because currently, according to the analyzed data, Bucharest it is the best example of a city that follows the models and ideals of sustainability in Romania.

In order to identify more clearly the differences between the capital of Romania Bucharest and the other cities, both those that currently adopt a smart-city policy and those that do not adopt a smart-city policy, the 42 cities for each variable that obtained a Mean Decrease Gini score with a value above 1 (Table 4) will be analyzed.

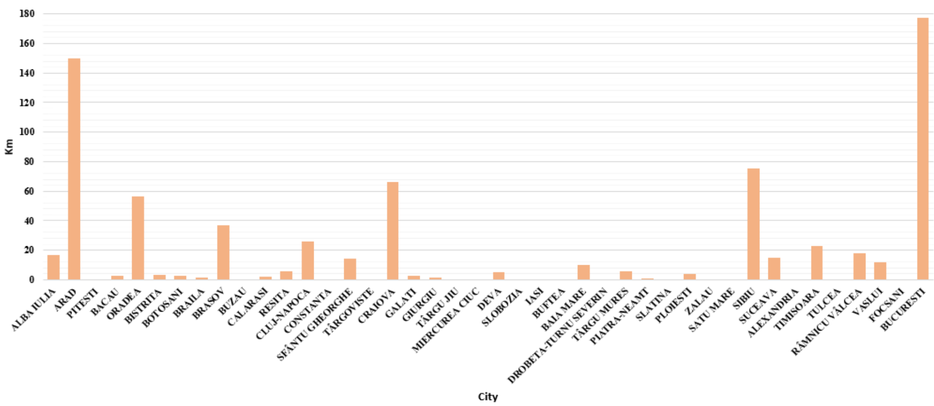


Figure 2. Bicycle Track Length

Source: own work using Excel.

By far, Bucharest is the city with the longest length of bike lanes, followed in this ranking by cities such as Arad, Craiova, Sibiu or Oradea. At the opposite pole, there are cities that do not register a single km of bike lanes. Among these cities, there are also cities that approach a "smart-city" strategy, such as: Satu Mare, Tulcea, Targu-Jiu, Iasi, Constanta, Slatina or Targoviste, but also cities that do not approach a "smart-city" strategy at the level of administration, public, such as: Zalau, Focsani, Buzau, Pitesti, Buftea, Drobeta Turnu-Severin or Miercurea Ciuc. Also, another city that does not approach a "smart city" strategy, Calarasi, has a very low level of this indicator (Figure 2).

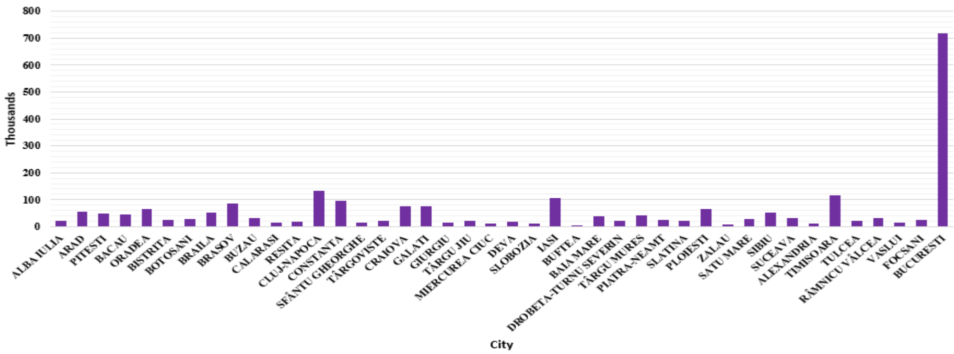


Figure 3. Number of internet connections fixed points ≥30Mbps (Ncipfx)
Source: own work using Excel.

In terms of digital infrastructure, Bucharest is much more developed compared to the rest of the cities, county capitals. The second-highest-ranked city is Cluj-Napoca, but at a great distance from Romania's capital Bucharest. However, it is noticed that large urban agglomerations (Bucharest, Cluj-Napoca, Iasi, Constanta, Timisoara) are more developed from the perspective of digitalisation, to the detriment of cities with a smaller population, such as: Miercurea Ciuc, Zalau, Alba-Iulia, Alexandria, or Buftea. As expected, cities that have not adopted a smart-city policy register low values, this result confirming the idea of the direct association relationship between the concept of smart city and the field of digitisation (Figure 3).

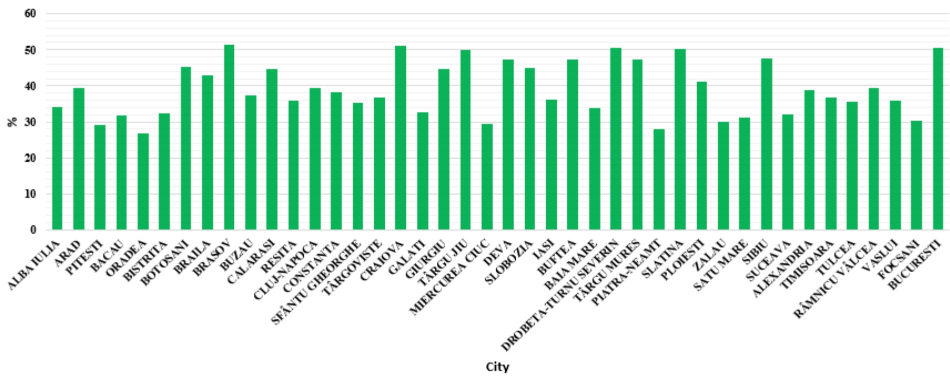


Figure 4. Air quality index (AQI)
Source: own work using Excel.

If for the previous variables, the differences between Bucharest and the rest of the cities were significant, in terms of air quality, Bucharest is in the top, along with cities such as Brasov, Craiova, Sibiu, Slatina, or Drobeta Turnu-Severin. Interestingly, the city of Drobeta Turnu-Severin has a similar value to most cities that have adopted a "smart-city" strategy, confirming once again the idea that the Random Forest algorithm performs very well in identifying the factors that led to the development of such a strategy, having a lower degree of accuracy in classifying cities that have not adopted a smart-city strategy. It can also be seen that the amplitude of the series is relatively small (Figure 4).

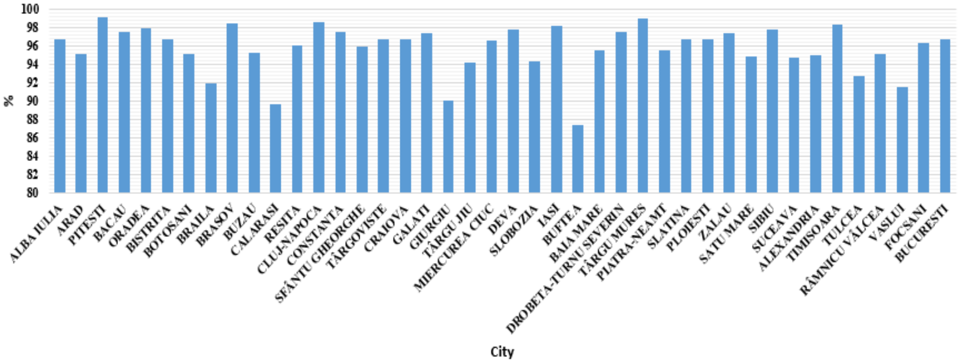


Figure 5. Share of dwellings equipped with water supply installations (Water)
Source: own work using Excel.

As with the Air Quality Index variable, the amplitude of the series is relatively small. Again, Bucharest is overtaken by cities such as Brasov, Cluj-Napoca, Constanta, Galati, Targu-Mures, Sibiu, Timisoara. Interestingly, some cities that have not adopted a "smart city" strategy, such as Pitesti, outperform many cities that have approached a "smart-city" policy, including Bucharest. On the other hand, Buftea and Calarasi are two cities that have low values of these indicators and have not approached a "smart-city" strategy (Figure 5).

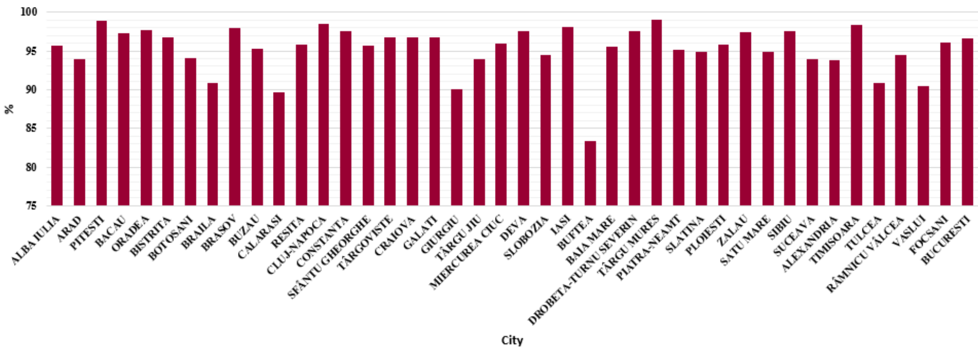


Figure 6. Share of dwellings connected to the sewerage system (Sewerage)
Source: own work using Excel.

The results of this variable present the same picture as for the variable Share of dwellings equipped with water supply installations, indicating a direct and strong association between the two variables. The lowest values are found in Buftea, Giurgiu and Calarasi, while Pitesti, Brasov, Timisoara, Cluj-Napoca, Targu Mures, Constanta, and Deva are cities ahead of Bucharest, which in turn register a value of more than 95% (Figure 6).

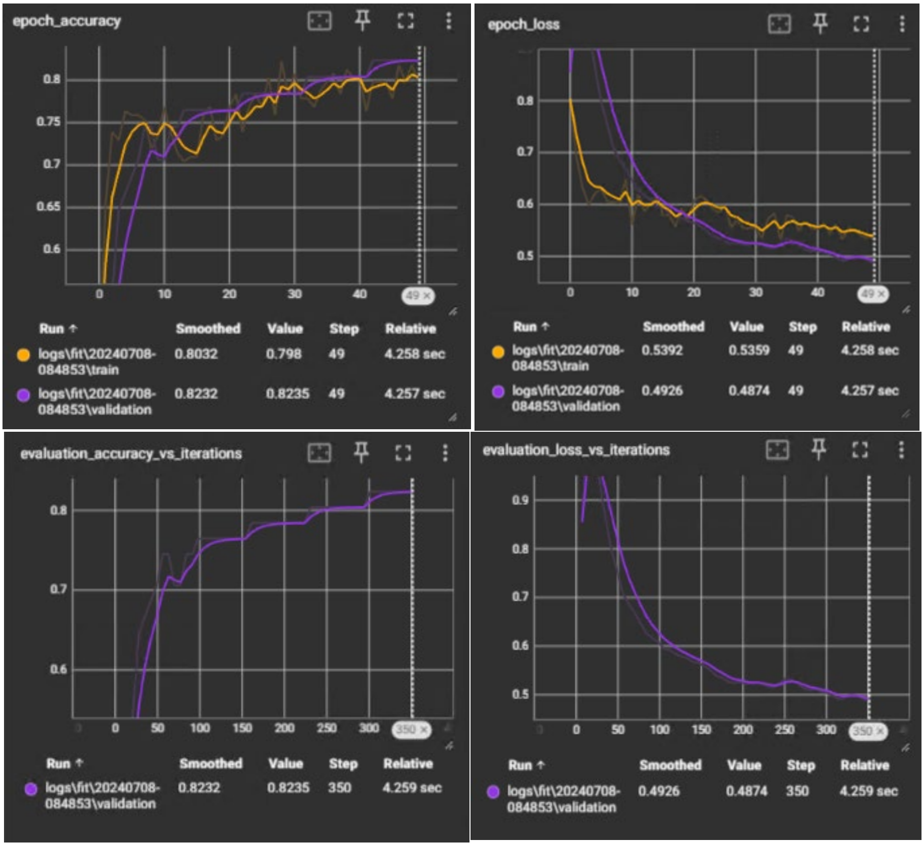


Figure 7. Epoch Accuracy; Epoch Loss; Accuracy v.s. Iterations; Loss v.s. Iterations
Source: own work using TensorBoard.

The training and validation phases of Deep Learning obtained the following results: test Accuracy: 0.921875; validation Accuracy: 0.8235294222831726 (Figure 7).

The evolution of the training and validation phases is (Figure 7):

Correct count: **267** (real = prediction)
Accuracy: 0.839622641509434 (Correct count / All records)
Pessimism count: **38** (real=1 / prediction = 0)
Optimism count: **13** (real=0 / prediction = 1)

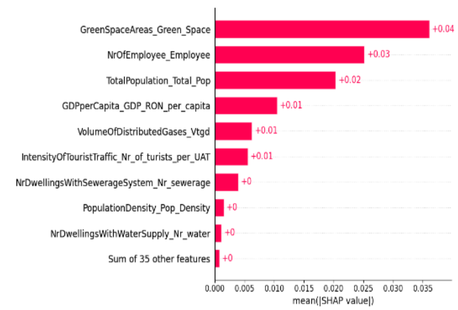
Eli5 results for the feature importance:

Eli5 shows green spaces, total population and number of employees as the main contributing factors for a smart city. This is also in line with previous findings.

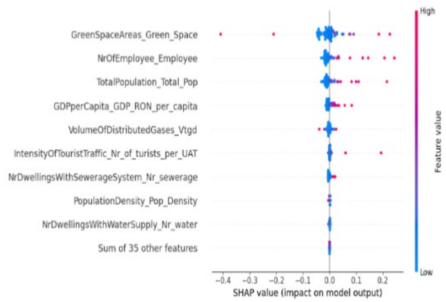
0.0424 ± 0.0257 GreenSpaceAreas_Green_Space
0.0252 ± 0.0057 TotalPopulation_Total_Pop
0.0074 ± 0.0044 NrOfEmployee_Employee
0.0015 ± 0.0019 IntensityOfTouristTraffic_Nr_of_turists_per_UAT
0.0005 ± 0.0010 GDPperCapita_GDP_RON_per_capita
0.0000 ± 0.0000 GreenSpacePerInhabitant_GS_sqm_per_cap
0.0000 ± 0.0001 PopulationDensity_Pop_Density
0.0000 ± 0.0000 TimeToEmergencyHospital_T_Hospital
0.0000 ± 0.0000 DevelopmentExpenses_Tot_Chelt_Dez
0.0000 ± 0.0000 TimeToIndustrialPark_T_industrial_park
0.0000 ± 0.0000 HousingExpenses_Chelt_Loc_Sociale
0.0000 ± 0.0000 TouristAttractionIndex_IDX_Tourism
0.0000 ± 0.0000 RenewableElectricity_Regen
0.0000 ± 0.0000 PercentOfHomesWithSewerage_Sewerage
0.0000 ± 0.0000 EmploymentRate_20_64_Empl_rate_20_64
0.0000 ± 0.0000 NrInternetConnOver30_Ncipfx
0.0000 ± 0.0000 NonMotorizedTrModalQuota_Pnmtra_mq
0.0000 ± 0.0000 TimeToEuropeanRoad_T_to_european_road
0.0000 ± 0.0000 EnvironmentalProtectionExp_Cpmed
0.0000 ± 0.0000 DensityOfUrbanStreets_D_urban_str
… 24 more …

Figure 8. Eli5 results
Source: own work using Eli5 algorithm.

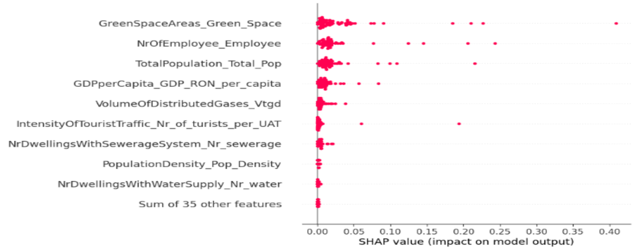
According to SHAP, the most relevant metric in determining a smart city characteristic is the presence of green space. We can note that the second metric is the total population, and the 3rd is the number of employees. This confirms the previous findings and validates the model.



SHAP bar plot - the mean absolute value of each feature over all the rows of the dataset



SHAP beeswarm chart



SHAP beeswarm chart – Absolute values

Figure 9. SHAP results for the feature importance
Source: own work using SHAP algorithm.

Both interpretability methods used offer similar results in respect to metrics importance. Due to the benefits of smart cities to life quality, those findings can help authorities focus on the development of strategies in this field.

5. Conclusions

Cities such as Bucharest, Cluj-Napoca, Timisoara, and Iasi are part of the best-developed county residences that support sustainable development by arranging large area of green spaces and using renewable energy.

At the same time, there are also less developed cities, such as Slatina, Alba-Iulia, Targu-Mures. However, it is worth noting that although they do not excel in certain areas, they show some strong points, as we noted earlier, when it comes to each county's spending on housing and public development, meaning that the cities show a desire to achieve the goal of a sustainable city.

Regarding the most important factors that influence whether a city applies a "smart city" policy or not, we can find the provision of housing with water and sewage systems, air quality, bike lanes, and the number of Internet connections.

It is true that for a city to be sustainable, it should also be smart, as mentioned in the specialised literature, a fact confirmed by the increased importance of the predictor coming from the digitalisation sphere, more precisely from the internal connection area, in the Random Analysis Forest.

Moreover, cities must have the minimum necessary for living in residential housing, thus the water distribution network and sewerage are important factors that must be considered, although it has been demonstrated that most county residences have a high level of inclusion of the population from the perspective of the two utilities.

Recommendations for cities in Romania would be to focus on technological development to be able to implement more and more approaches to sustainable development, encouraging the use of renewable energy, especially in small cities, and maintaining and developing green spaces to encourage environmental protection.

One limit of the research is represented by the reference moment of the data, represented by the year 2018. In the Random Forest study, another limit is represented by the relatively small number of cities selected in the sample, compared to the total number of urban settlements in Romania.

Interpretability techniques such as Shap and Eli5 were used to find which metrics have more importance in the determination of a smart city attribute. Both showed green space, population, and number of employees as the main contributing factors towards a smart city characteristic.

Therefore, in the future, the research will be developed by updating the data to the last available reference year and by training other types of models to extract other insights from the available data.

References

[1] Chourabi, H., Nam, T., Walker, S., Mellouli, S., Nahon, K., Pardo, T.A., Scholl, H.J. (2012), *Understanding Smart Cities: An Integrative Framework. 2012 45th Hawaii International Conference on System Sciences.*

[2] Elmaghraby, A.S., Losavio, M.M. (2014), *Cyber security challenges in Smart Cities: Safety, security and privacy. Journal of Advanced Research*, 5(4), 491-497.

[3] Erős, N., Török, Z., Hossu, C.-A., Olga Réti, K., Maloș, C., Kecskés, P., Morariu, S.-D., Benedek, J., Hartel, T. (2022), *Assessing the sustainability related concepts of urban development plans in Eastern Europe: A case study of Romania. Sustainable Cities and Society*, 85, 104070.

[4] European Commission. (2019), *European Green Deal*. Brussels: European Commission.

[5] Fetting, C. (2020). *The European green deal. ESDN Report*, 2(9).

[6] Gkontzis, A.F., Kotsiantis, S., Feretzakis, G., Verykios, V.S. (2024), *Enhancing urban resilience: smart city data analyses, forecasts, and digital twin techniques at the neighborhood level. Future Internet*, 16(2), 47.

[7] Goodfellow, I., Bengio, Y., Courville, A. (2016), *Deep Learning*. MIT Press.

[8] Gradinaru, G.I., Maricut, A.C. (2023), *The Impact of Economic Behaviour Environment: A Comparative Study Among EU's Countries and China. Journal of the Knowledge Economy*, 1-15.

[9] Hölscher, K., Wittmayer, J.M., Loorbach, D. (2018), *Transition versus transformation: What's the difference? Environmental Innovation and Societal Transitions*, 27, 1-3.

[10] Ibănescu, B.C., Pascariu, G.C., Bănică, A., Bejenaru, I. (2022), *Smart city: A critical assessment of the concept and its implementation in Romanian urban strategies. Journal of Urban Management*, 11(2), 246-255.

[11] Ismagilova, E., Hughes, L., Rana, N.P., Dwivedi, Y.K. (2020), *Security, Privacy and Risks Within Smart Cities: Literature Review and Development of a Smart City Interaction Framework. Information Systems Frontiers*, 24(2), 393-414.

[12] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R.,* Springer Science & Business Media.

[13] Korobov, M., Lopuhin, K. (2021), ELI5 Documentation Release 0.11.0, https://eli5.readthedocs.io/en/0.11.0/.

[14] Kumar, H., Singh, K., Gupta, M., Madaan, J. (2020), *Moving towards smart cities: Solutions that lead to the Smart City Transformation Framework. Technological forecasting and social change*, 153, 119281.

[15] Lewis, O., Sousa, S., Pinho, P. (2022), *Multifunctional Green Infrastructure in Shrinking Cities: How Does Urban Shrinkage Affect Green Space Planning? Urban Planing*, 7(2S1), 186-202.

[16] Meerow, S., Newell, J.P., Stults, M. (2016), *Defining urban resilience: A review. Landscape and Urban Planning*, 147, 38-49.

[17] Scheda R, Diciotti S. (2022), *Explanations of Machine Learning Models in Repeated Nested Cross-Validation: An Application in Age Prediction Using Brain Complexity Features*. *Applied Sciences*, 12(13), 6681.

[18] Scornet, E., Biau, G., Vert, J.P. (2015), *Consistency of random forests. Annals of statistics*, 43(4), 1716-1741.

[19] Song, Y., van Timmeren, A., Wandl, A. (2019), *A literature review and categorisation of sustainability-aimed urban metabolism indicators: a context, indicator, mechanism, outcome analysis. Regional Statistics*, 9(1), 54-71.

[20] Tatsat, H., Puri, S., Lookabaugh, B., (2021), *Machine Learning and Data Science Blueprints for Finance.* O'Reilly Media.

[21] Van den Broeck G., Lykov, A., Schleich, M., Suciu D. (2022), *On the Tractability of SHAP Explanations. Journal of Artificial Intelligence Research*, 74, 851-886.

[22] Van Zoonen, L. (2016), *Privacy concerns in smart cities. Government Information Quarterly*, 33(3), 472-480.

[23] Xu, C., Chen, G., Huang, Q., Su, M., Rong, Q., Yue, W., Haase, D. (2022), *Can improving the spatial equity of urban green space mitigate the effect of urban heat islands? An empirical study. Science of The Total Environment*, 841, 156687.

[24] Yan, Z., Sun, Z., Shi, R., Zhao, M. (2023), *Smart city and green development: Empirical evidence from the perspective of green technological innovation. Technological Forecasting and Social Change*, 191, 122507.

[25] Yu, W., Xu, C. (2018), *Developing Smart Cities in China. International Journal of Public Administration in the Digital Age,* 5(3), 76-91.

[26] Zaman, G., Grădinaru, G.I., Maricuț, A.C. (2021), *Maintaining and Rehabilitating Urban Ecosystems, in the Context of European Green Deal. Romanian Statistical Review*, 2, 3-13.