Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 345-353
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Deciphering Urban Happiness: Analysis of Machine Learning Approaches for Comprehensive Urban Planning

Gregorius Airlangga
Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
Email: gregorius.airlangga@atmajaya.ac.id

Abstract

This study embarks on an analytical journey to decode the complexities of urban happiness, leveraging a suite of advanced machine learning models. At the heart of our methodology is the innovative application of CatBoost, Random Forest, Gradient Boosting, and Linear Regression models, each chosen for its distinct ability to navigate the multifaceted nature of our urban dataset. CatBoost is highlighted for its proficiency in managing categorical data, essential in reflecting the diverse elements of urban environments. Concurrently, Random Forest's capability in reducing variance and overfitting, along with Gradient Boosting's precision in optimizing across various loss functions, plays a pivotal role in the accuracy of our predictions. Linear Regression serves as a baseline, offering simplicity and interpretability for comparative analysis. Central to our evaluation is the Root Mean Squared Error (RMSE) metric, providing a quantitative measure of our models' accuracy. This approach is instrumental in translating the intricate relationships within urban data into actionable insights for urban planning and policy-making. Our study not only demonstrates the effectiveness of an ensemble approach in machine learning but also emphasizes the importance of interpretability in model selection and evaluation. The findings offer a comprehensive understanding of urban happiness, serving as a valuable resource for stakeholders in urban development and policy formulation. This research marks a significant stride in harnessing machine learning's potential to enrich urban life quality.

Keywords: *CatBoost, Random Forest, Gradient Boosting, Linear Regression, Urban Hapiness*

1. INTRODUCTION

In the realm of urban studies, the concept of 'urban happiness' has emerged as a focal point of interdisciplinary research, reflecting the complexities and intricacies of life in modern cities [1]–[3]. This burgeoning field, situated at the intersection of sociology, urban planning, environmental science, and data science, seeks to unravel the various factors that contribute to the well-being of urban dwellers [4]–[6]. With the United Nations projecting that 68% of the world's population will reside in urban areas by 2050, understanding the dynamics of urban happiness is not only academically intriguing but also of paramount importance for effective urban planning and policy formulation [7]–[9]. The study of urban happiness traditionally relied on survey-based methods and statistical analyses, which primarily focused on correlating demographic and socio-economic factors with self-reported levels of happiness [10]–[12]. While these approaches have provided valuable insights, they often fall short in capturing the dynamic and multifaceted nature of urban ecosystems. Urban happiness is influenced by a myriad of factors, including but not limited to environmental quality, healthcare infrastructure, traffic conditions, cost of living, and green spaces [13]–[15]. These elements interact in complex ways, and their impact on urban dwellers' happiness can vary significantly across different contexts and cultures.

Recognizing these limitations, recent research in the field has turned to more sophisticated analytical methods. The advent of big data and advanced computational techniques has particularly revolutionized this domain [16]–[18]. Machine learning algorithms, with their ability to handle large datasets and uncover complex, non-linear relationships, have opened new avenues for research. For example, studies have employed various machine learning models to assess the impact of environmental factors, urban design, and social dynamics on urban happiness [19]–[21]. These models range from traditional regression analyses to more advanced algorithms like Random Forest, Gradient Boosting, and neural networks. However, despite the growing body of research employing machine learning in urban happiness studies, there remains a significant gap [19], [22], [23]. Much of the existing literature focuses on the application of single models or a limited set of models, with little comparative analysis of their efficacy. This lack of comprehensive comparative studies is a notable omission, given the diverse nature of machine learning algorithms and their differing abilities to capture various aspects of urban happiness data [24]–[26].

Our research aims to address this gap by conducting an extensive comparative analysis of several advanced machine learning models in predicting urban happiness. This study includes CatBoost, Random Forest, Gradient Boosting, and Linear Regression. By comparing these models in the same experimental setting, we aim to provide a clearer understanding of their respective strengths and weaknesses in the context of urban happiness prediction. The methodology adopted in this research is marked by its comprehensive approach to data preprocessing, feature engineering, and model optimization. We utilize state-of-the-art tools such as AutoViz for automated data visualization. This approach is critical, as it moves beyond mere prediction accuracy, offering insights that are valuable for urban planners and policymakers.

Furthermore, our study is distinguished by its rigorous approach to data handling and analysis. The dataset employed encompasses various aspects of urban life, including environmental quality indicators, traffic density, green space availability, air quality indices, and healthcare infrastructure. This rich dataset enables a holistic analysis of the factors contributing to urban happiness. Additionally, advanced encoding methods are employed to handle categorical data effectively, ensuring that the models effectively capture the nuances of the data. The results of this study are presented in a detailed comparative analysis, focusing on the Root Mean Squared Error (RMSE) metric for both training and testing datasets. This metric is chosen for its ability to quantify the difference between the predicted and actual values, providing a clear measure of model accuracy. The analysis goes beyond mere RMSE scores, delving into the models' performance nuances and discussing their implications in the broader context of urban studies.

In the discussion section, we interpret the results, drawing connections to existing literature and highlighting the practical implications of our findings. This includes an examination of how different models capture various data aspects and the implications of these findings for urban planners and policymakers. We also discuss the limitations of our study and suggest avenues for future research.

Finally, the paper concludes by summarizing the key takeaways and contributions of the study. We emphasize the importance of using advanced machine learning techniques in urban studies and the potential of these methods to provide deeper insights into the complex dynamics of urban happiness. The paper also outlines potential areas for future research, suggesting that further studies could explore additional models, incorporate different types of data, or apply the methodology to different urban contexts..

2. RESEARCH METHODOLOGY

2.1. Data Collection and Preparation

The foundation of our research rests on a comprehensive dataset focused on various aspects of urban life, including environmental quality, healthcare, traffic density, green spaces, air quality indices, and cost of living. The data were sourced from a combination of urban wellbeing surveys and publicly available urban metrics, encompassing a diverse range of cities across different geographical and socio-economic contexts. The dataset can be downloaded in [27]. The dataset was subjected to a rigorous preparation process to ensure its suitability for machine learning analysis. Initially, the data were cleansed to remove any inconsistencies and missing values. This involved techniques like imputation for handling missing data and anomaly detection to identify and correct any outliers or data entry errors. Following the cleaning process, the dataset underwent a transformation phase, where categorical variables were encoded using advanced encoding techniques. This was essential for converting non-numeric data into a format suitable for machine learning models. The encoding was carefully chosen to preserve the inherent characteristics of each variable, with techniques such as one-hot encoding for nominal variables and ordinal encoding for ordinal variables. Lastly, we split data into 50% training and 50% testing.

2.2. Feature Engineering

In the pursuit of predicting urban happiness, the role of feature engineering emerged as a cornerstone in our methodology, transcending mere data analysis to craft a narrative that intertwines various urban elements into a coherent story. The process of feature engineering was not just a technical exercise; it was an endeavor to unearth and articulate the subtle nuances and intricate relationships embedded within the urban fabric. This phase was particularly crucial given the multidimensional nature of our dataset, encompassing diverse aspects such as traffic density, air quality, cost of living, and healthcare quality. The genesis of our feature engineering process was rooted in a deep understanding of urban dynamics. It involved an iterative process of hypothesis formulation, testing, and refinement. Central to this was the recognition that certain aspects of urban living do not exist in isolation but rather in a symbiotic relationship with one another. For instance, we hypothesized that traffic density and air quality are intrinsically linked - an increase in traffic density often leads to deterioration in air quality as presented in the equation 1. To capture this interaction, we engineered a feature

that encapsulates the interplay between these two variables, providing our models with a more nuanced view of urban environmental quality.

Similarly, the relationship between the cost of living and healthcare quality was another focal point of our feature engineering as presented in the equation 2. Intuitively, cities with a higher cost of living often have better healthcare services, but this is not a universal truth. By creating a feature that explores this relationship, we aimed to offer a more refined lens through which the models could assess the overall quality of urban life. Moreover, our approach to feature engineering was informed by a blend of empirical insights derived from the data and theoretical frameworks from urban studies. This dual approach ensured that the new features were not only statistically sound but also held practical significance in the context of urban planning and policy-making. The engineered features were designed to be interpretable, ensuring that the insights gleaned from the models could be translated into actionable strategies for enhancing urban happiness.

In our experimental setup, as reflected in the code, these engineered features played a pivotal role. The process started with the selection of relevant columns from our dataset, including 'Happiness_Score', 'Month', 'Year', 'Decibel_Level', 'Traffic_Density', 'Green_Space_Area', 'Air_Quality_Index', 'Cost_of_Living_Index', and 'Healthcare_Index'. These columns not only provided a broad overview of the various factors influencing urban happiness but also served as the raw materials for our feature engineering process as presented in the equation 3. The creation of these new features necessitated careful preprocessing and transformation of the data. This included handling missing values, encoding categorical variables, and normalizing the data to ensure consistency and comparability across different scales. Such meticulous preprocessing provided a robust foundation for the subsequent modeling phase, ensuring that the input data effectively represented the complex realities of urban environments.

$$\text{Traffic}_{\text{AirQuality}_{\text{Feature}}} = \alpha \cdot \text{Traffic}_{\text{Density}} + \beta \cdot \text{Air}_{\text{Quality}_{\text{Index}}} + \gamma \cdot \text{Traffic}_{\text{Density}} \times \text{Air}_{\text{Quality}_{\text{Index}}} \quad (1)$$

$$\text{CostHealthcare\_Feature} = \delta \cdot \text{Cost\_of\_Living\_Index}^{n} + \epsilon \cdot \text{Healthcare\_Index}^{m} \quad (2)$$

$$\text{Happiness\_Score} = w_1 \times X_1 + w_2 \times X_2 + \cdots + w_n \times X_n \quad (3)$$

2.3. Model Selection and Training

In our quest to unravel the complexities of urban happiness, we embarked on an analytical journey with a suite of diverse machine learning models, each chosen for its unique strengths and suitability for various aspects of our multifaceted dataset. This selection encompassed CatBoost, Random Forest, Gradient Boosting, and Linear Regression as presented in the equation (4) – (7) respectively, forming a comprehensive ensemble that addresses different dimensions of data complexity and model interpretability. CatBoost emerged as a frontrunner in our modeling arsenal, primarily for its exceptional capability in managing categorical data, which forms a significant portion of our dataset. Its prowess in reducing overfitting, a common pitfall in machine learning endeavors, further solidified its position in our

study. The robustness of CatBoost is particularly advantageous in scenarios where the dataset is replete with categorical variables, making it a tool of choice for datasets reflecting the variegated nature of urban environments.

Parallelly, we employed Random Forest, an ensemble learning method celebrated for its effectiveness in diminishing variance and overfitting. The inherent structure of Random Forest, building a multitude of decision trees and merging them for a more accurate and stable prediction, makes it particularly suitable for our complex dataset that includes both numerical and categorical variables. Its ability to handle such a mix with finesse is instrumental in dissecting the intricacies of urban happiness. Further augmenting our model suite is Gradient Boosting, another ensemble technique renowned for its precision and ability to optimize across various loss functions. The iterative nature of Gradient Boosting, where each new model incrementally improves upon its predecessor, lends itself to scenarios where accuracy is of utmost significance. This model's inclusion is pivotal in our pursuit of the most nuanced understanding of the factors contributing to urban happiness. Complementing these advanced models is Linear Regression, a fundamental statistical approach revered for its simplicity and interpretability. Its role in our study extends beyond its predictive capability; it serves as a baseline for comparison against the more complex models. The juxtaposition of Linear Regression's results with those of the advanced models provides a frame of reference to gauge the added complexity's effectiveness and necessity. The training process was meticulously designed to ensure the integrity and robustness of our findings.

We commenced by partitioning the dataset into training and testing sets, a critical step that guarantees both sets are representative of the overall data spectrum as presented in the equation (8). The training phase was not merely an execution of predefined steps; it was an iterative process of learning and adaptation. Each model was attentively trained on the training set, with hyperparameters being meticulously fine-tuned through a series of experiments. This fine-tuning was not an exercise in trial and error but a deliberate strategy to optimize the models' performance, ensuring they are attuned to the subtleties of our data. To fortify the training process against overfitting and to validate the models' generalizability, we employed cross-validation as presented in the equation (9). This technique, pivotal in the realm of machine learning, entails dividing the dataset into several smaller sets and using these sets in rotation as training and validation data. Through cross-validation, we ensured that each model's performance was not a fluke of particular data quirks but a true representation of its predictive prowess.

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \tag{4}$$

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} T_i(x) \tag{5}$$

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 345-353
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

$$\widehat{y_t} = \widehat{y_{t-1}} + \nu \cdot h_t(x) \quad (6)$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (7)$$

$$\text{Dataset} = \text{TrainingSet} \cup \text{TestingSet} \quad (8)$$

$$\text{CV}_{\text{score}} = \frac{1}{k}\sum_{i=1}^{k} \text{ModelScore}(\text{Train}_i, \text{Validate}_i) \quad (9)$$

2.4. Model Evaluation

Model performance was primarily evaluated using the Root Mean Squared Error (RMSE) metric as presented in the equation (10). RMSE provides a clear measure of the model's accuracy by quantifying the difference between the predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad (10)$$

2.5. Ethical Considerations and Limitations

Throughout the research process, ethical considerations were taken into account, particularly regarding data privacy and the use of publicly available data. The limitations of the study were also acknowledged, including the scope of the dataset and the inherent biases in machine learning models.

3. RESULTS AND DISCUSSION

The deployment of our selected machine learning models on the test dataset yielded insightful results, primarily evaluated through the lens of the Root Mean Squared Error (RMSE) as presented in the table 1. The RMSE scores for each model on the test data were as follows: CatBoost registered an RMSE of 0.932, Random Forest achieved an RMSE of 0.887, Gradient Boosting recorded an RMSE of 0.986, and Linear Regression presented an RMSE of 0.906. These results offer a rich ground for discussion, not only in terms of comparative model performance but also in their implications for urban happiness prediction and the underlying dynamics of the dataset. The Random Forest model emerged as the top performer with the lowest RMSE score of 0.887, indicating its superior predictive accuracy in our urban happiness dataset. This model's success can be attributed to its ability to handle the complex and non-linear relationships inherent in the multifaceted urban data. Random Forest's ensemble approach, which builds numerous decision trees and merges their results, provides a robust mechanism for capturing the diverse influences on urban happiness.

CatBoost, with an RMSE of 0.932, also demonstrated commendable performance. Its specialization in handling categorical data likely contributed to its effectiveness, considering the significant presence of categorical variables in our dataset. The slightly higher RMSE compared to Random Forest could be indicative of CatBoost's sensitivity to certain features or a requirement for further

hyperparameter tuning. Linear Regression, with an RMSE of 0.906, surprisingly outperformed the more complex Gradient Boosting model (which had an RMSE of 0.986). This outcome suggests that, despite the high dimensionality and complexity of the data, a simpler model like Linear Regression can still capture significant variance in urban happiness scores. It underscores the notion that more complex models do not always guarantee better performance, especially in datasets where linear relationships are prominent.

Table 1. Machine Learning Results (RMSE)

Methods	RMSE
CatBoost	0.932
Random Forest	0.887
Gradient Boosting	0.986
Linear Regression	0.906

The Gradient Boosting model, while generally highly regarded for its predictive accuracy, scored the highest RMSE in this study. This could be due to overfitting during training, where the model becomes too finely tuned to the training data, losing its generalizability to new data. Alternatively, it might indicate a mismatch between the model's complexity and the data's structure, suggesting a need for further refinement of the model's configuration. The results from our study have significant implications for urban happiness prediction. The effectiveness of Random Forest highlights the importance of ensemble methods in capturing the complex, layered nature of urban data. These methods, by aggregating multiple learning algorithms, provide a more nuanced understanding of the factors influencing urban happiness.

The performance of Linear Regression, being close to that of more complex models, raises critical questions about model selection in urban studies. It suggests that while advanced machine learning models are powerful tools, simpler models should not be overlooked, as they can offer comparable accuracy with the added benefit of interpretability. This is particularly relevant for policymakers and urban planners who require clear, understandable models to inform their decisions. Moreover, the varying performances of the models underscore the need for a thoughtful approach to model selection, considering both the characteristics of the data and the specific research questions at hand. It also emphasizes the importance of comprehensive data preprocessing and feature engineering in enhancing model performance.

4. CONCLUSION

The exploration into predicting urban happiness through various machine learning models has culminated in a comprehensive understanding of the intricacies involved in this domain. Our study, centered around the application and comparative analysis of CatBoost, Random Forest, Gradient Boosting, and Linear Regression models, has not only shed light on the predictive capabilities of these methods but also illuminated the complex nature of urban happiness itself. The findings of our research reveal a significant insight: the effectiveness of a machine

learning model in predicting urban happiness is profoundly influenced by its ability to handle the multifaceted and often non-linear relationships inherent in urban data. Random Forest emerged as the most effective model in our study, indicating the strength of ensemble methods in managing complex datasets. However, the competitive performance of the simpler Linear Regression model underscores an essential lesson in predictive modeling: complexity does not always equate to superiority. In certain scenarios, simpler models can offer comparable accuracy, with the added advantage of interpretability and ease of use.

This research contributes to the growing body of knowledge in urban studies, particularly in the application of machine learning to urban planning and policy-making. The nuanced understanding gained through this comparative analysis provides valuable guidance for urban planners and policymakers. It underscores the potential of data-driven approaches in enhancing urban life, emphasizing that the selection of appropriate predictive models is crucial in accurately capturing the dynamics of urban happiness. Furthermore, our study opens avenues for future research. The exploration of additional variables, alternative modeling techniques, and the application of these findings to different urban contexts could enrich the understanding of urban happiness. The potential of machine learning in this field is vast, and its full exploration requires continual and rigorous investigation.

REFERENCES

[1] J. Pykett, T. Osborne, and B. Resch, "From urban stress to neurourbanism: how should we research city well-being?," *Ann. Am. Assoc. Geogr.*, vol. 110, no. 6, pp. 1936–1951, 2020.

[2] J. Finch, *Literary Urban Studies and how to Practice it*. Routledge, 2021.

[3] M. J. Alfaro Muñoz, "Youth Happiness in the City: Children's and Adolescents' Experiences of Happiness in the Urban Environment of Lima, Peru," 2021.

[4] A. Khan and A. Ali, "The Interplay Between Urban Environment and Mental Health: A Comprehensive Examination and Policy Roadmap," *J. Hum. Behav. Soc. Sci.*, vol. 7, no. 7, pp. 1–18, 2023.

[5] R. Wang, X. Zhang, and N. Li, "Zooming into mobility to understand cities: A review of mobility-driven urban studies," *Cities*, vol. 130, p. 103939, 2022.

[6] Y. L. R. G. J. L. R. Zhang L. Cao and P. Shu, "Decoding Spontaneous Informal Spaces in Old Residential Communities: A Drone and Space Syntax Perspective," *ISPRS Int. J. Geo-Information*, vol. 12, no. 11, p. 452, 2023.

[7] M. Reeves *et al.*, "Visions for a flourishing society under demographic aging."

[8] V. Narain and D. Roth, "Introduction: Peri-Urban Water Security in South Asia," *Water Secur. Confl. Coop. Peri-Urban South Asia Flows across Boundaries*, pp. 1–26, 2022.

[9] A. Va\vnová, K. Vitálišová, and D. Roj\'\iková, "Prospects of Systems of Megacities and Individual Megacities with Respect to Regional Economy," in *Indo-Pacific Smart Megacity System: Emerging Architecture and Megacity Studies*, Springer, 2023, pp. 163–206.

[10] A. Sinha, B. Chandra, A. K. Mishra, and S. Goswami, "An Assessment on Quality of Life and Happiness Indices of Project Affected People in Indian Coalfields," *Sustainability*, vol. 15, no. 12, p. 9634, 2023.

[11] J. Brzozowski and N. Coniglio, "International migration and the (un) happiness

push: Evidence from Polish longitudinal data," *Int. Migr. Rev.*, vol. 55, no. 4, pp. 1089–1120, 2021.
[12] Á. Fernández-Pérez and Á. Sánchez, "Non-clinical factors and citizens' satisfaction: A way to improve the quality of health systems," *Int. J. Healthc. Manag.*, pp. 1–9, 2023.
[13] R. leBrasseur, "Citizen sensing within urban greenspaces: Exploring human wellbeing interactions in deprived communities of Glasgow," *Land*, vol. 12, no. 7, p. 1391, 2023.
[14] M. Juntti, H. Costa, and N. Nascimento, "Urban environmental quality and wellbeing in the context of incomplete urbanisation in Brazil: Integrating directly experienced ecosystem services into planning," *Prog. Plann.*, vol. 143, p. 100433, 2021.
[15] A. Addas, "The importance of urban green spaces in the development of smart cities," *Front. Environ. Sci.*, vol. 11, p. 1206372, 2023.
[16] I. H. Sarker, "Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective," *SN Comput. Sci.*, vol. 2, no. 5, p. 377, 2021.
[17] J. M. Górriz *et al.*, "Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications," *Neurocomputing*, vol. 410, pp. 237–270, 2020.
[18] G. Aceto, V. Persico, and A. Pescapé, "Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0," *J. Ind. Inf. Integr.*, vol. 18, p. 100129, 2020.
[19] L. Xiang, M. Cai, C. Ren, and E. Ng, "Modeling pedestrian emotion in high-density cities using visual exposure and machine learning: Tracking real-time physiology and psychology in Hong Kong," *Build. Environ.*, vol. 205, p. 108273, 2021.
[20] C. Yin and C. Shao, "Revisiting commuting, built environment and happiness: New evidence on a nonlinear relationship," *Transp. Res. Part D Transp. Environ.*, vol. 100, p. 103043, 2021.
[21] M. Helbich, J. Hagenauer, and H. Roberts, "Relative importance of perceived physical and social neighborhood characteristics for depression: a machine learning approach," *Soc. Psychiatry Psychiatr. Epidemiol.*, vol. 55, pp. 599–610, 2020.
[22] Y. Chen, K. Sherren, M. Smit, and K. Y. Lee, "Using social media images as data in social science research," *New Media \& Soc.*, vol. 25, no. 4, pp. 849–871, 2023.
[23] E. F. Fang *et al.*, "A research agenda for ageing in China in the 21st century: Focusing on basic and translational research, long-term care, policy and social networks," *Ageing Res. Rev.*, vol. 64, p. 101174, 2020.
[24] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, 2021.
[25] Z. Jiang, J. Liu, and L. Yang, "Comparison Analysis of Stock Price Prediction Based on Different Machine Learning Methods," in *Proceedings of the 2nd International Academic Conference on Blockchain, Information Technology and Smart Finance (ICBIS 2023)*, 2023, vol. 14, p. 59.
[26] M. Karimi, M. S. Mesgari, and R. S. Purves, "A comparative assessment of machine learning methods in extracting place functionality from textual content," *Trans. GIS*, vol. 26, no. 8, pp. 3225–3252, 2022.
[27] E. AI, "City Happiness Index 2024." 2024.