Predicting Data Center Presence in European NUTS3 Regions: A Machine Learning Approach

Team 2 / GENYZ

"Every time that we blink, there is a new data center popping up," said Aurora Gómez of a Spanish campaign group [8]. Europe is in the midst of a data center construction boom. The geographical determinants of data center location are studied for the USA, but no equivalent NUTS3-level analysis exists for Europe.This study addresses this gap by investigating the factors that influence the presence of data centers across European NUTS3 regions. The research question is: What are the main factors that contribute to the presence of data centers in a European region? To answer this question, we develop an end-to-end data science pipeline encompassing data collection, preprocessing, feature engineering, model development, evaluation, and interpretation.Our hypothesis is that data center presence is driven by a favorable combination of economic, infrastructural, and environmental conditions: reliable power, proximity to internet hubs, water availability, cheap electricity, and renewable energy supply.

1 RELATED WORK

Bagnasco [1] shows that data centers concentrate where reliable electricity, suitable climate, and strong connectivity converge. AI workloads further amplify demands for cooling capacity and computational resources.

Charles [2] highlights proximity to business centers, transportation infrastructure, and time-zone coverage as secondary locational drivers.

The IEA [3] reports hyperscale operators lead renewable energy procurement through power purchase agreements (PPAs) rather than site selection near renewables. Together these studies underscore the complex interplay motivating our multi-dimensional feature set.

2 METHODOLOGY

2.1 Data Collection

Data were collected from three sources: the Data Center Map (facility-level presence), Eurostat (socio-economic and energy variables at NUTS3), and Google searches for supplementary indicators. Variables span six categories: socio-demographic/economic, energy/environmental, digital infrastructure, climate risk, and water security. Variables were collected across different reference years (2023–2025). The target — Data_Center_Presence — is binary (0: absence, 1: presence), with 44.1 % positive and 55.9 % negative regions (Figure 1).

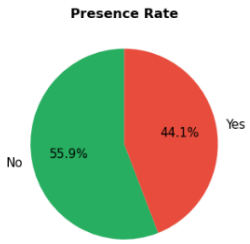


Figure 1: Target variable distribution — 44.1 % presence (Yes), 55.9 % absence (No)

2.2 Preprocessing Pipeline

The pipeline follows these sequential steps:

1. Feature removal: 13 uninformative or redundant columns dropped. The variables containing the year are not relevant for the prediction, also for the min and max variables their mean was computed. GHI horizontal was redundant with the optimal GHI. For the cold mortality and heat mortality per 100k_3.0 °C, the data centers are present locally so we disregarded the global warming scenario.
2. Imputation: NaN values replaced by mean (symmetric distributions) or median (skewed/outlier-prone) — strategy chosen per-variable.
3. Train/test split (80/20, stratified, random_state=42) applied before any scaling to prevent data leakage.
4. Derived features: Average_speed = (download + upload)/2 and GDP_per_capita = GDP/population computed before scaling to avoid distorted non-linear relationships.
5. Collinearity: correlation matrix with threshold 0.8 [11]; highly correlated raw features dropped. Residual multicollinearity verified with VIF.
6. PCA: dimensionality reduction and visualization of feature contributions per principal component.

2.3 Interpretable Model (statsmodels)

A logistic regression (statsmodels Logit, LBFGS solver) was fitted to the scaled, collinearity-filtered features. Useful for sanity checking, a logit coefficient helps us to track the direction of features related to the presence of data center. Only features with $p < 0.05$ are shown in Figure 2. Population and GDP per capita carry the largest positive log-odds coefficients, indicating economically denser regions are significantly more likely to host data centers. Average latency and water abstraction show small negative coefficients. As the latency increases, connectivity quality worsens so the probability of the data center decreases. On the other hand the water abstraction may represent the resource stress where it indicates water scarcity and sustainability concerns. It can be also

a proxy for other activities as agricultural regions and heavy industrial areas.

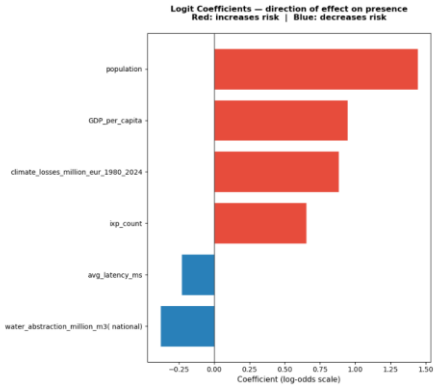


Figure 2: Logit coefficients for significant features (p < 0.05). Red: increases presence risk; Blue: decreases risk.

2.4 Predictive Models (sklearn)

Six classifiers were trained on PCA-transformed features: Logistic Regression, Random Forest (balanced weights), SVM-RBF (balanced weights), Decision Tree, kNN, and Gaussian Naïve Bayes. Evaluation used 5-fold StratifiedKFold cross-validation with metrics: accuracy, precision, recall, F1, and ROC-AUC. Hyperparameters optimized via GridSearchCV, targeting F1 — appropriate because the task requires balancing missed detections (false negatives) against false alarms.

2.5 Clustering (KMeans)

KMeans clustering was conducted to determine the concentration of clusters where the datacenter could be or not by the nearest point. The optimal number of clusters was identified as k=2 using the elbow curve and the silhouette score.

The elbow curve (Figure 3) plummeted in the inertia as the number of clusters increased from 2 to 10. The inertia measures how complex the clusters are. The inertia decreases because adding more clusters allows data points to be grouped more tightly. Initially the decrease is a steep towards 2, the curve changes to a more gradual decline. The figure below represents a good trade-off between having few clusters (simple model) to having compact clusters (good fit).

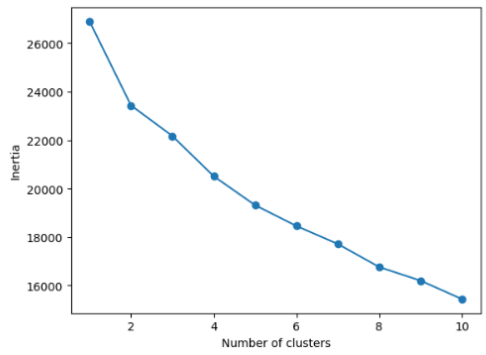


Figure 3: Elbow cure

The silhouette score curve (Figure 4) was analyzed to evaluate how well-separated and compact the clusters are. The highest silhouette score is around 0.2, it indicates that the structure in the data is weak. It indicates that the points are slightly closer to their own cluster than to neighboring clusters. The clusters overlap considerably. To visualize the representation of data points , the Principal Components Analysis (PCA) (Appendix 1) was used to reduce dimensionality, the plot shows one large cloud without separation. For this classification, this clustering algorithm is not appropriate.

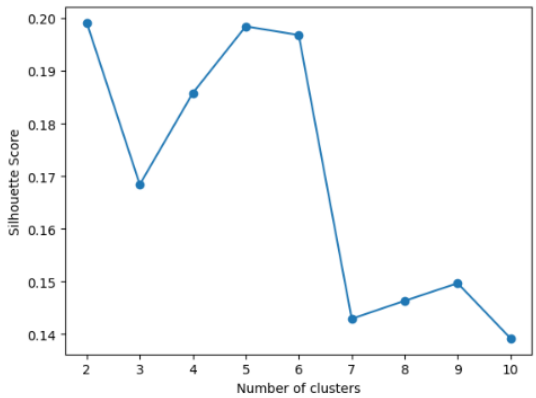


Figure 4: Silhouette score curve

3 RESULTS AND INTERPRETATION

3.1 Model Comparison

Table 1 summarizes classifier performance. SVM-RBF achieves more balanced values than other models, indicating acceptable discrimination [4]. After threshold tuning (0.5), the confusion matrix shows 113 true negatives, 77 true positives, 37 false positives, and 42 false negatives (Figure 5). The tuned model yields accuracy of 69.5%, recall of 66.4%, precision of 65.3%, and F1 of 65.9 %.

Model	Acc.	Prec.	Rec.	F1	Roc_auc
Log. Reg.	0.74	0.72	0.65	0.69	0.80
Rand. Forest	0.71	0.69	0.60	0.64	0.78
SVM-RBF ★	**0.72**	**0.70**	**0.65**	**0.67**	**0.79**
Dec. Tree	0.62	0.57	0.55	0.56	0.61
kNN	0.65	0.72	0.35	0.47	0.70
Gauss. NB	0.66	0.68	0.42	0.52	0.70

★ Best model selected for further analysis

Table 1. Model Performance after Tuning (Test Set)

The baseline model (left) shows 108 true negatives, 79 true positives, 42 false positives, and 40 false negatives. After threshold tuning to 0.5 (right), true negatives increase to 113, true positives to 77, false positives decrease to 37, while false negatives increase

slightly to 42. The tuning improves specificity while maintaining recall balance.

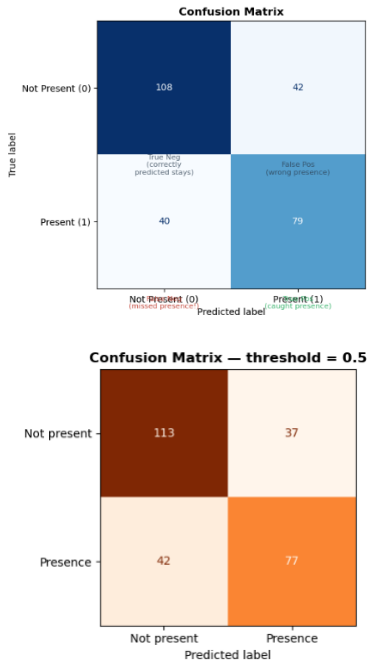


Figure 5: Confusion matrix vs tuned confusion matrix

3.2 Hyperparameter tuning and final model

To improve the model, different hyperparameters of each model were tested through GridSearchCV which combines all the parameters and picks the best one using 5-fold cross-validation on the training set. Which key metric should be optimized to be aligned with the strategic goal? The accuracy is 69.5%, the recall 66.4%, the precision 65.3% and the F1-score is 65.9%. For this study, awareness is focus on both missing opportunities and avoids bad recommendations. F1-score is more suitable. The best parameters show us:

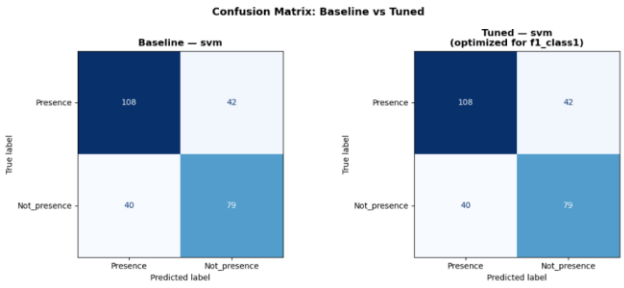


Figure 6: Confusion matrix : baseline vs tuned

The fact that the F1 model achieves 79 true positives while keeping false positives at 42 suggests it provides the most practical balance for identifying candidate datacenter regions without generating an excessive number of false leads.

3.3 Feature Importance

Permutation importance on the tuned SVM (Figure 7) ranks GDP per capita highest, followed by climate-related economic losses (1980–2024), average connection speed, GHI optimal, population, and electricity price. Renewable energy capacity (renewable_number) ranks last, contradicting the initial hypothesis. This confirms operators prioritize economic and connectivity conditions.

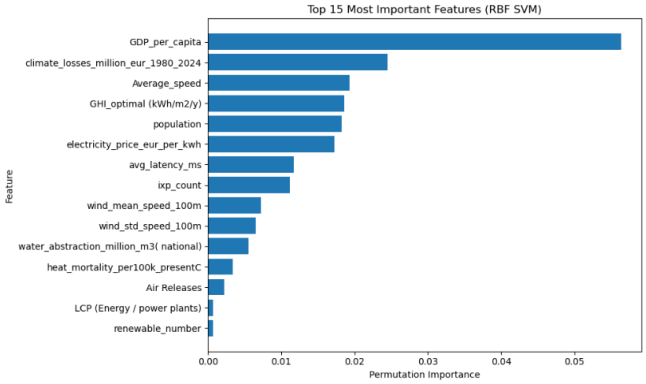


Figure 7: Top 15 most important features (RBF SVM permutation importance)

3.4 Partial Dependence Analysis

Partial dependence plots (Figure 8) reveal directional relationships. GDP per capita, climate losses, average speed, and population all show a positive association with data center presence. GHI optimal shows a negative trend — higher solar irradiance (associated with southern, rural regions) discourages presence. Electricity price displays a non-linear U-shape: both very low and very high prices are associated with higher presence probability, suggesting different operator segments respond differently to energy costs.

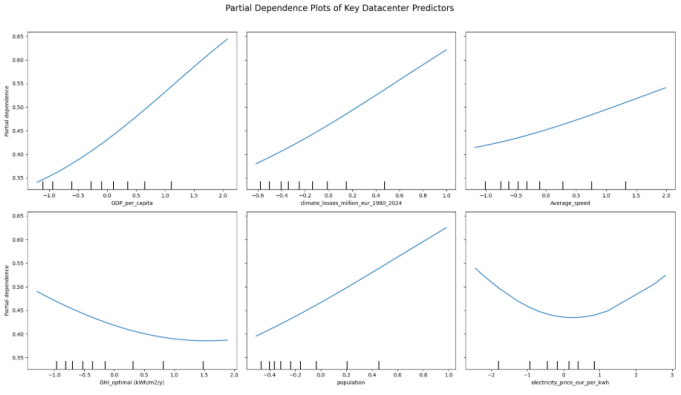


Figure 8: Partial dependence plots for top 6 predictors

4 DISCUSSION

The work by Bagnasco [1] for the USA is partially aligned with our European results. In both contexts, agglomeration effects, market size, and infrastructure availability are key drivers of data center spatial distribution. In contrast, electricity price does not show a statistically significant effect in either case. The USA study faced limitations in obtaining local climate conditions — variables that, if available, might have altered the cross-regional comparison.

The key features driving the presence of data enter are not controlled by a single authority. They emerge from communications between national governments, local authorities and private actors.

Based on the results of the importance of the features, at national level, governments can affect the digital infrastructure through a regulatory framework but also the electricity price by infrastructure planning. In contrast, at local level, authorities can have influence on local environment regulations (as water abstraction). Factors such as GDP per capita and population are largely determined by broader socioeconomic dynamics and evolve over timescale. From this point of view, while governments can create favorable environments for digital infrastructure investments, the attractiveness of a location depends also on the demographic and economic variables that are less directly controllable by public authorities.

5 CONCLUSIONS

Europe is experiencing an unprecedented data center construction boom, yet the geographic drivers of location at sub-national scale remain understudied. This paper builds a machine learning pipeline to classify data center presence across NUTS3 regions using socio-economic, energy, digital infrastructure, climate risk, and water security features. Six classifiers are trained; a Support Vector Machine (SVM-RBF) achieves the best discrimination (AUC = 0.79, F1 = 0.67). GDP per capita and climate-related economic losses are the dominant predictors. Renewable energy capacity contributes minimally, suggesting operators rely on power purchase agreements rather than colocation with generation assets. Partial dependence analysis reveals a non-linear U-shaped relationship between electricity price and data center presence. These findings align with international evidence on agglomeration effects and market-size dynamics.

ACKNOWLEDGMENTS

• Variables were collected at different temporal scales (2023–2025), introducing potential inconsistency across feature years.

• The Global Power Plant Database may omit small-scale or recently commissioned plants, leading to underestimation of renewable capacity in some regions.

• Moderate model performance (AUC 0.744, F1 0.659) suggests relevant unmeasured factors: land costs, permitting regimes, fiber topology, and tax incentives.

REFERENCES

[1] Bagnasco, M. (2026). Spatial Determinants of AI Data Center Location. Department of Management and Production Engineering, Torino.
[2] Charles, N. (2024). Guide to data center site selection: Key considerations and Best Practices.
[3] (IEA), I. E. (2023, 07 11). 1. Data Centres and Data Transmission Networks: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks#programmes
[4] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.
[5] Eurostat (2024). NUTS — Nomenclature of Territorial Units for Statistics.
[6] Data Center Map (2024). Global Data Center Directory. datacentermap.com
[7] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. JMLR, 12, 2825–2830.
[8] Vailshery, L. S. (2025). Data center boom in Europe. Statista Research Department.
[9] Ian Editor (Ed.). 2007. The title of book one (1st. ed.). The name of the series one, Vol.9. University of Chicago Press, Chicago. DOI: https://doi.org/10.1007/3-540-09237-4
[10] David Kosiur. 2001. Understanding Policy-Based Networking (2nd. ed.). Wiley, New York, NY..
[11] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning.
[12] Sten Andler. 1979. Predicate path expressions. In Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79). ACM Press, New York, NY, 226-236. DOI: https://doi.org/10.1145/567752.567774

AUTHORS

Elie Diab (DIATI, Politecnico di Torino, S335519)

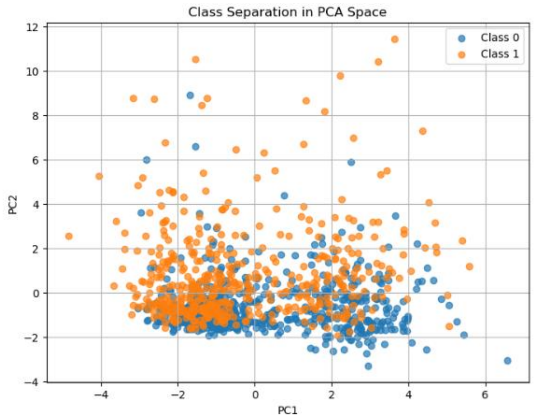
Farah Garayeva (DIATI, Politecnico di Torino, S324723)

Aref Alizadeh (DIATI, Politecnico di Torino, S359967)

Francis Obe (DIATI, Politecnico di Torino, S351802)

Karim Shamilov (DIATI, Politecnico di Torino, S360046).

APPENDIX



Appendix 1: The Principal Components Analysis (PCA)