Politecnico di Torino

Data Science & Machine Learning for Engineering Applications

Prof. Daniele Quercia

What Drives the Location of Data Centers in Europe?

Date: 05/06/2026

What Drives the Location of Data Centers in Europe?

A Machine Learning Analysis at NUTS3 Regional Level

Hamieh, Ali
Georesources and Geoenergy Engineering
Politecnico di Torino
s353164@studenti.polito.it

Nizam, Alaa
Georesources and Geoenergy Engineering
Politecnico di Torino
s354195@studenti.polito.it

Guidetti, Giovanni
Georesources and Geoenergy Engineering
Politecnico di Torino
s358720@studenti.polito.it

Tanios, Joe
Georesources and Geoenergy Engineering
Politecnico di Torino
s3455372@studenti.polito.it

Abstract

This study investigates the socioeconomic, infrastructural, and environmental determinants of data center placement across 1,345 European NUTS3 regions. The classification task predicts whether a region hosts at least one data center (44% of regions) or none (56%). We compare an interpretable logistic regression model (M2: economy + digital infrastructure) against a Random Forest classifier. Both models are validated with 5-fold cross-validation on the training set. On the held-out test set, Logit M2 achieves ROC-AUC ≈ 0.776 and the Random Forest ≈ 0.781. Permutation importance identifies GDP as the strongest single driver, followed by renewable energy capacity, network latency, download speed, and water-related variables. Geographic maps confirm that data centers cluster in economically dense Western European corridors, with underrepresentation in Eastern and peripheral Southern regions. KNN imputation (fitted on training data only) was used to handle missing values in a leakage-free preprocessing pipeline.

1 Research Question, Hypothesis & Related Work

Research Question: What structural characteristics of a European NUTS3 region — economic scale, digital infrastructure, energy availability, and environmental conditions — help explain whether at least one data center is located there?

Hypothesis: Regions with stronger economic activity, better digital connectivity, lower network latency, and greater proximity to internet infrastructure are more likely to host data centers. Energy and environmental factors are expected to contribute additional predictive information, especially in nonlinear models, but are likely to play a secondary role compared with economic and digital infrastructure variables.

Related Work

Digital infrastructure clusters where demand, fibre connectivity, and peering are already strong [1,2]. IXPs, broadband quality, and latency indicate regional readiness; electricity availability, price, and reliability are critical operational constraints; and recent literature adds renewable energy, water stress, and climate vulnerability as siting factors [3].

Three sources specifically guided our hypotheses: a multi-criteria site-selection framework for data centers [3]; CBRE's Global Data Center Trends 2024 [1], highlighting power availability and market demand; and a report on European data center expansion into emerging markets [2], confirming that infrastructure readiness shapes where growth spreads beyond Western hubs.

2 Methodology

2.1 Dataset & Target Variable

The dataset covers 1,345 European NUTS3 regions and combines four categories of features: (i) economic indicators — GDP and population; (ii) digital infrastructure — IXP count, distance to nearest IXP, average download speed, and latency; (iii) energy indicators — renewable and non-renewable installed capacity (MW) and electricity price (EUR/kWh); and (iv) environmental indicators — drought index (SPEI), heat mortality, climate-related losses, water-availability variables, and the Water Energy Index (WEI+).

The binary target variable is defined as:

- 1 = the region hosts at least one data center (593 regions, ≈ 44 %)
- 0 = the region does not host any data center (752 regions, ≈ 56 %)

ROC-AUC is used as the primary metric; class_weight='balanced' is applied in tree-based models.

2.2 Preprocessing Pipeline

All preprocessing steps were designed to prevent data leakage from the test set:

- Removal of identifiers, target-leakage variables (Data_Center_Count), year metadata, and theoretically redundant variables (e.g., plant count, which is dominated by installed MW capacity).
- Stratified 80/20 train/test split performed before any imputation or scaling, so the test set never influences preprocessing.
- Correlation-based feature reduction removed highly redundant predictors (threshold |r| > 0.95), applied after the train/test split.

◆ KNN Imputation — Handling Missing Values

Features were first scaled to the [0,1] range using MinMaxScaler before KNN imputation and then inverse-transformed back to their original scale for interpretability.

2.3 Models

Interpretable model (M2) — A logistic regression estimated with statsmodels. Three specifications were compared — economy only (M1), economy + digital infrastructure (M2), and a full model including energy and environmental variables (M3). M2 was selected as the best balance between parsimony, interpretability, and predictive performance. Its features are: population, GDP, electricity price, IXP count, distance to nearest IXP, average download speed, and latency.

Predictive model — Several scikit-learn classifiers were compared: Logistic Regression, Random Forest (500 trees, min_samples_leaf = 3, class_weight = 'balanced'), and Gradient Boosting. The Random Forest was selected as the best predictive model based on ROC-AUC on the held-out test set. It was trained on all 31 cleaned features to maximise predictive performance rather than coefficient interpretability.

2.4 Validation Strategy

Both models are evaluated on the held-out test set using ROC-AUC (primary metric, robust to class imbalance), accuracy, precision, recall, F1, and confusion matrices.

To verify robustness, 5-fold cross-validation was applied on the training set only. For the logistic model, the scaler was embedded in a Pipeline so that scaling was fitted within each fold, eliminating any within-fold leakage. Permutation importance (30 repeats, ROC-AUC scoring) was applied to the Random Forest to quantify each feature's contribution to predictive discrimination.

3 Results & Interpretation

3.1 Target Distribution & Geographic Pattern

The target distribution shows a moderate class imbalance: 593 of 1,345 NUTS3 regions host at least one data center (≈ 44%), while 752 do not (≈ 56%). The spatial distribution is markedly uneven: data centers cluster in economically dense Western European corridors (Greater London, Benelux, Rhine-Ruhr, Ile-de-France, Po Valley), while Eastern and peripheral Southern regions are largely without data centers. Model performance is evaluated on the held-out test set and by cross-validation.

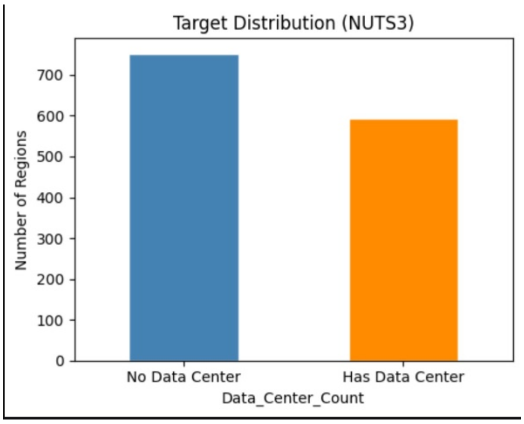

Figure 1: Target distribution across NUTS3 regions.

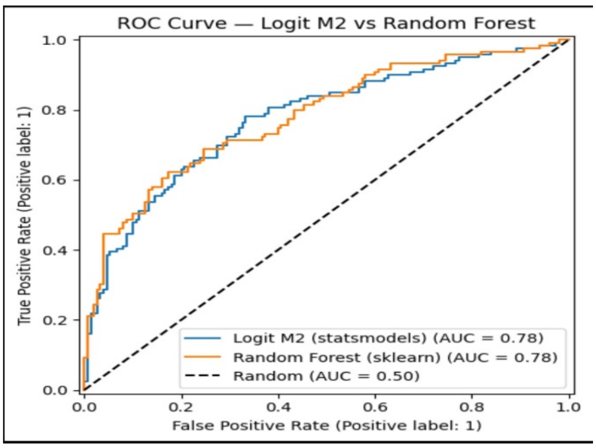

Figure 2: ROC Curve — Logit M2 vs Random Forest (AUC = 0.78 each).

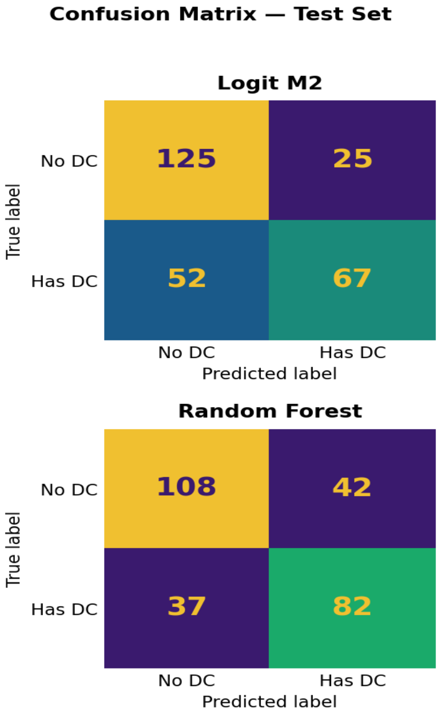

Figure 3: Confusion matrices on the held-out test set.

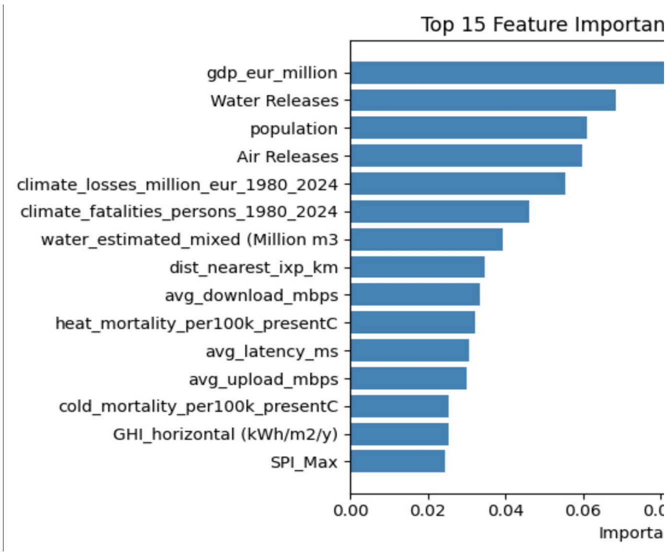

Figure 4: Top-15 permutation feature importance — Random Forest.

3.2 Cross-Validation Results

5-fold CV on the training set confirms that neither model overfits; the logistic model's scaler was embedded in a Pipeline to prevent within-fold leakage.

Model / Metric	Value
Logit M2 (Interpretable)	
CV ROC-AUC (mean)	0.815
± Std Dev	0.022
Random Forest (Predictive)	
CV ROC-AUC (mean)	0.816
± Std Dev	0.031

Table 1: 5-fold cross-validated ROC-AUC (training set only, scaler fitted within each fold). Low standard deviations indicate stable, generalisable performance.

3.3 Test-Set Performance Comparison

Model / Metric	Value
Logit M2 (Interpretable)	
Accuracy	~0.714
ROC-AUC	~0.776
F1 (DC = 1)	~0.635
Random Forest (Predictive)	
Accuracy	~0.706
ROC-AUC	~0.781
F1 (DC = 1)	~0.675

Table 2: Performance on the held-out test set. ROC-AUC values are close, but the Random Forest improves recall for the positive class (DC = 1).

Both models achieve nearly identical ROC-AUC (≈0.776 vs ≈0.781). The Random Forest improves recall for the positive class at the cost of more false positives; the Logit is more conservative. The Logit is preferred for coefficient interpretability; the Random Forest for detecting regions likely to host data centers.

3.4 Permutation Importance & Key Drivers

Permutation importance shuffles each feature's values 30 times and records the mean drop in ROC-AUC. Features with high importance are those the Random Forest genuinely relies on. The analysis reveals the following hierarchy of drivers:

- **GDP (gdp_eur_million) —** The strongest and most stable predictor. Regions with higher economic output generate greater demand for digital infrastructure and are more likely to possess the complementary assets (skilled labour, reliable power, developed logistics) that data centers require. This result is consistent across both the interpretable and predictive models.
- **Renewable energy capacity (renewable_MW) —** The second most important feature in the Random Forest. This reflects corporate sustainability commitments by major cloud operators and the growing requirement for green energy certification (e.g., Power Purchase Agreements). Regions with substantial renewable capacity are increasingly preferred siting locations.
- **Heat mortality & climate losses —** Environmental stress indicators that the nonlinear Random Forest uses to distinguish vulnerable regions from stable ones. High heat mortality and climate-related losses signal regions with

unfavourable conditions for sensitive computing infrastructure.

- **Average download speed & latency —** Broadband quality and network responsiveness proxy for fibre infrastructure density and backbone connectivity. High-speed, low-latency regions already have the duct capacity and routing equipment that data centers require for last-mile connectivity.
- **Water-related variables (water_estimated_mixed, WEI+) —** Water availability is a non-trivial location constraint: large data centers require significant cooling water. Regions under high water stress (high WEI+) are penalised.
- **Population —** Correlated with market size and demand for latency-sensitive services (CDN, cloud gaming, financial APIs), though its effect is secondary once GDP and connectivity are controlled for.

GDP is the clearest and most stable driver; infrastructure and environmental variables add further predictive information. The Random Forest captures nonlinear interactions not visible in the logit.

3.5 Interpretable Model: Coefficients & Odds Ratios (Logit M2)

The Logit M2 model provides statistically grounded evidence on the direction and significance of key associations after standardisation of all features:

- **GDP (+, significant) —** Regions with higher economic output are significantly more likely to host data centers. The positive coefficient is statistically significant ($p < 0.05$) and consistent with the permutation importance results.
- **Latency (−, significant) —** Regions with higher average latency are significantly less likely to host data centers, confirming the importance of high-quality digital infrastructure for low-latency-sensitive applications.
- **Population (−, significant after controls) —** After controlling for GDP and connectivity, population has a negative coefficient. This counterintuitive result should be interpreted carefully: conditional on economic output, higher population may capture regions with lower economic concentration or a different regional structure rather than indicating any inherent disadvantage.
- **IXP count (+, marginal) —** A positive but only marginally significant effect. The linear model partially captures what permutation importance confirms in the nonlinear model: IXP proximity matters, but its effect is partially absorbed by the correlated variables GDP and download speed.
- **Electricity price — not significant in M2.** Its role should be interpreted cautiously: it may still matter operationally, but the linear model does not provide strong statistical evidence once the other variables are included.

The confusion matrix for Logit M2 reveals that the main error mode is false negatives — regions that host a data center but are predicted not to. These are concentrated among medium-GDP Eastern European cities that are beginning to attract investment but still lack the connectivity density typical of Western hubs.

4 Conclusion

This analysis demonstrates that data center location in Europe follows a structured spatial logic rather than a random pattern. Economic scale is the most robust predictor: regional GDP is strongly and consistently associated with the probability of hosting at least one data center, across both the interpretable logistic model and the Random Forest's permutation importance ranking. Digital infrastructure — particularly latency and download speed — is the second dimension, while renewable energy capacity, water availability, and climate-related variables add meaningful predictive information in the nonlinear model.

The two models serve complementary purposes: the Logit M2 provides statistically interpretable coefficients and odds ratios useful for policy communication, while the Random Forest achieves higher recall for the positive class and captures nonlinear interactions. Cross-validation confirms that both models generalise well beyond the training set, and KNN imputation ensured that missing-value handling did not introduce any leakage.

From a policy perspective, regions seeking to attract data center investment should focus on digital connectivity (latency, fibre), renewable energy access, and water/climate constraints — not only fiscal incentives. The persistent East-West digital divide reflects structural inequalities that fiscal policy alone cannot bridge.

The main limitation is the absence of land prices, tax incentives, grid reliability, permitting constraints, proximity to cloud availability zones, and operator strategies. Future work should add longitudinal data to study how regions transition from zero to at least one data center over time.

References

1. CBRE. (2024). Global Data Center Trends 2024. CBRE Research.
2. Hittinger, E., & Jaramillo, P. (2019). Internet of Things: Energy boon or bane? Science, 364(6438), 326–328. https://doi.org/10.1126/science.aau8825
3. Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984–986. https://doi.org/10.1126/science.aba3758
4. Siting & selection for data centres: A multi-criteria decision-making model -Tugrul U. Daim, Ashok Bhatla & Mohammad Mansour- https://doi.org/10.1080/19397038.2012.719554. (Project literature paper cited in the analysis.)
5. Report on European data center growth into emerging markets - Owen Thomson & Coen Hutters. (Project literature paper cited in the analysis.)
6. Van Dijk, J. (2019). The Network Society (3rd ed.). SAGE Publications.