Predicting Data Centre Locations Across European NUTS3 Regions

A Machine Learning Approach Using Socioeconomic, Infrastructure, and Environmental Indicators

Group 12 · Part 3 — Predictive Modelling Report · 2025

ABSTRACT

This report presents a supervised machine learning study predicting both the presence and concentration of data centers across 1,345 European NUTS3 regions spanning 37 countries. Two complementary tasks are addressed: binary classification (does a region host at least one data center?) and regression (how many data centers does it host?). Following data cleaning — including the removal of ten highly correlated redundant variables — interpretable and high-performance models are trained and cross-validated. For classification, a Logistic Regression model (statsmodels) serves as the interpretable baseline (test accuracy = 0.718, macro F1 = 0.645), while a Random Forest (scikit-learn) achieves 71.2% mean 5-fold accuracy and 70.6% mean macro F1. For regression, an OLS model reaches a test-set R^2 of 0.567, while a Random Forest Regressor achieves a cross-validated R^2 of 0.614. Across all models, regional GDP and Internet Exchange Point (IXP) infrastructure consistently emerge as the dominant predictors of data center activity.

Keywords: data centers, NUTS3, machine learning, classification, regression, Random Forest, Logistic Regression, OLS, European digital infrastructure

1. INTRODUCTION

Data centers are among the most capital-intensive and strategically significant assets of the digital economy. Siting decisions depend on a complex matrix of factors: access to fiber networks and internet exchanges, affordable and renewable energy, regulatory environments, climate suitability for cooling, and proximity to dense, high-income labor markets. For investors, planners, and policymakers, anticipating which sub-national regions are likely hosts — using only publicly available statistics — carries high practical value.

This study addresses that challenge using NUTS3-level data across 37 European countries. Our research question is: What socioeconomic, infrastructure, and environmental characteristics drive the presence and concentration of data centers across European NUTS3 regions? We hypothesize that economic mass (GDP, population) and digital connectivity (IXP presence, internet quality) are the dominant determinants, consistent with findings that digital infrastructure investment mirrors the geography of economic activity and agglomeration [1, 2]. A secondary role is expected for energy and environmental variables, given the industry trend toward renewable-powered, climate-efficient facilities [3].

2. DATASET

The dataset contains 1,345 NUTS3 regions with 45 raw variables spanning five domains: (i) economic indicators — GDP in million euros (gdp_eur_million), population, and electricity price; (ii) digital infrastructure — IXP count, IXP total participants, distance to nearest IXP, average download speed and latency, and broadband test count (total_tests); (iii) energy capacity — installed renewable and non-renewable MW and number of installations; (iv) environmental variables — wind speed statistics at 100 m height, Global Horizontal Irradiance, drought indices (SPEI, SPI), soil moisture, water abstraction, and pollutant releases; and (v) climate risk — heat and cold mortality rates, climate-related economic losses (1980–2024).

The target variable Data_Center_Count is heavily right-skewed: 55.9% of regions (752 of 1,345) host no data centers; 44.1% (593 regions) host at least one, with a maximum of 106 in a single region (mean = 2.0, std = 6.0). For classification, a binary indicator has_dc was derived.

3. DATA CLEANING & FEATURE SELECTION

Two cleaning steps were applied. First, seven non-predictive columns were removed unconditionally: the region identifier, both target variables, and four metadata year fields. Second, to reduce multicollinearity, one variable from each pair with Pearson $|r| > 0.85$ was dropped, retaining the more interpretable or higher target-correlated member. Ten variables were removed in this step (see Appendix A). After cleaning, 29 features were retained for modelling.

4. PREPROCESSING PIPELINE

All transformers were fitted exclusively on the training partition to prevent data leakage. The dataset was split 80/20 (1,076 train / 269 test) using stratified sampling on the binary target (random seed = 42), preserving the 44.1% positive class rate in both partitions. Missing values were imputed with column means (SimpleImputer) and all features were scaled to [0, 1] via Min-Max normalization, both fitted on training data only. For the interpretable models, the top 8 features by absolute Pearson correlation with the respective target were used as inputs, following the course recommendation to begin with parsimonious models.

5. CLASSIFICATION

5.1 Feature Relevance

The eight features most strongly correlated with has_dc were: gdp_eur_million ($r = 0.345$), total_tests (0.308), climate_losses_million_eur_1980_2024 (0.295), avg_download_mbps (0.284), population (0.267), non_renewable_number (0.226), avg_latency_ms (−0.226), and ixp_count (0.214). The negative sign on latency reflects data centers clustering in low-latency, well-connected urban hubs. Climate losses likely confound with economic wealth: richer regions both suffer larger monetary losses from climate events and host more data centers.

5.2 Logistic Regression — Interpretable Model

A statsmodels Logit model was fitted on the top 8 features most correlated with has_dc. On the held-out test set it achieved accuracy 0.718, macro-precision 0.720, macro-recall 0.703, and macro-F1 0.645. Five statistically significant drivers were identified at $p < 0.05$. gdp_eur_million is the strongest positive driver ($\beta = +46.86$, $p < 0.001$), confirming that economic output is the primary attractor. total_tests ($\beta = +10.69$, $p = 0.015$) reflects internet infrastructure maturity. non_renewable_number ($\beta = +1.95$, $p = 0.009$) captures energy availability. Conversely, avg_latency_ms ($\beta = -4.01$, $p = 0.007$) shows that high-latency, poorly connected regions are significantly less likely to host data centers. population ($\beta = -15.61$, $p = 0.011$) is negative conditional on GDP: raw headcount does not predict DC presence once economic productivity is controlled. climate_losses, avg_download_mbps, and ixp_count were not statistically significant ($p > 0.05$).

5.3 Random Forest — High-Performance Model

A Random Forest Classifier (300 trees, max_depth = 10, min_samples_leaf = 5, class_weight = 'balanced') was trained on all 29 features. Feature importance (mean impurity decrease) ranks gdp_eur_million first (0.167), followed by total_tests (0.104), population (0.069), Water Releases (0.065), and climate_losses_million_eur (0.058). The environmental pollutant variables likely reflect co-location with energy-intensive industrial regions rather than a direct causal effect. On the test set the model reached accuracy 0.699, macro-precision 0.695, macro-recall 0.695, macro-F1 0.661. Five-fold stratified cross-validation yielded mean accuracy 71.2% (±4.1%) and mean macro-F1 70.6% (±3.9%), confirming robust generalization. Table 1 summarizes all classification results.

Metric	Logit	RF (test)	RF CV
Accuracy	0.718	0.699	0.712 ± 0.041
Precision (macro)	0.720	0.695	—
Recall (macro)	0.703	0.695	—
F1 (macro)	0.645	0.661	0.706 ± 0.039

Table 1. Classification performance summary.

6. REGRESSION

6.1 Feature Relevance

Correlations with Data_Center_Count are substantially stronger than for the binary target, confirming that data centre concentration is highly predictable from regional characteristics. The top features are gdp_eur_million ($r = 0.758$), ixp_count (0.657), ixp_total_participants (0.627), total_tests (0.499), population (0.485), climate_losses_million_eur (0.478), non_renewable_number (0.248), and avg_download_mbps (0.238).

6.2 OLS Regression — Interpretable Model

An OLS model on the top 8 features identifies gdp_eur_million ($\beta = +37.0$) and ixp_total_participants ($\beta = +42.1$) as the dominant positive drivers of data center count. ixp_count adds a further positive effect ($\beta = +9.0$). Notably, population enters with a negative coefficient ($\beta = -9.2$) conditional on GDP and IXP infrastructure, indicating that raw headcount adds no explanatory power once economic output and connectivity are controlled: what drives data center density is the quality of the digital ecosystem, not population size per se. This conditional negative effect is consistent with co-location market research showing tier-1 IXP cities attract disproportionate facility concentration relative to their population [2].

6.3 Random Forest — High-Performance Model

A Random Forest Regressor (same architecture as classifier) concentrates 68.6% of feature importance on gdp_eur_million alone, with ixp_total_participants (8.6%) and ixp_count (5.2%) as distant second and third. The OLS model outperforms the RF on the held-out test set ($R^2 = 0.567$ vs 0.404), a result consistent with the dominant linear correlation of gdp_eur_million with the target ($r = 0.758$): a parsimonious linear model captures the primary structural signal effectively. Cross-validated performance is comparable across models (RF CV $R^2 = 0.614$ vs OLS test $R^2 = 0.567$), suggesting the test-set advantage of OLS partly reflects split variance. Table 2 summarizes all regression results.

Metric	OLS	RF (test)	RF CV
R^2	0.567	0.404	0.614 ± 0.138
MAE	1.585	1.607	1.537 ± 0.385
RMSE	5.126	6.012	—

Table 2. Regression performance summary.

7. DISCUSSION & CONCLUSION

Three findings are held consistently across all four models. First, economic mass is the primary driver of data center activity. GDP alone carries $r = 0.758$ with data center count and is the most important feature in every model. This aligns with the digital infrastructure agglomeration literature [1], confirming that data centers concentrate where enterprise IT demand — proxied by economic output — is greatest.

Second, IXP infrastructure provides an independent predictive signal beyond economic size. IXP count and total participants rank highly in both correlation analysis and regression feature importance, consistent with research showing proximity to internet exchanges reduces peering costs and latency, making IXP-dense cities especially attractive to hyperscale operators and content delivery networks [2].

Third, environmental and energy variables play a secondary role. Wind speed, renewable capacity, and pollutant release indicators appear in feature importance rankings, reflecting the growing industry emphasis on low-cost renewable energy and favorable cooling climates as siting criteria [3].

Several limitations warrant acknowledgement. The dataset is a cross-sectional snapshot and does not capture market dynamics. OLS and Logistic Regression assume spatial independence, which is unlikely given the geographic clustering of NUTS3 regions; future work should explore spatially aware models. Practically important variables absent from the dataset — submarine cable proximity, corporate tax regimes, planning environments — would likely improve predictive accuracy if incorporated.

In conclusion, this study demonstrates that publicly available NUTS3-level data supports reliable, interpretable predictions of data center location and concentration. The models provide a data-driven first screen for site selection and reveal the degree to which European digital infrastructure investment is geographically concentrated in already-wealthy, well-connected regions — a finding with direct implications for regional equity and the digital divide.

REFERENCES

[1] Grubesic, T. H., & Murray, A. T. (2004). Waiting for broadband: Local competition and the spatial distribution of advanced telecommunication services in the United States. Growth and Change, 35(2), 139–165.

[2] Malecki, E. J. (2002). The economic geography of the internet's infrastructure. Economic Geography, 78(4), 399–424.

[3] Masanet, E., et al. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984–986.

[4] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[5] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. JMLR, 12, 2825–2830.

[6] European Commission (2021). NUTS — Nomenclature of territorial units for statistics. Eurostat.

APPENDIX A — REMOVED FEATURES (|R| > 0.85)

Dropped Feature	Retained Feature	|r|
GHI_horizontal (kWh/m²/y)	GHI_optimal (kWh/m²/y)	0.995
SPEI_min	SPEI_mean	0.952
water_estimated_mixed	population	0.948
SPI_Max	SPI_Mean	0.951
SPI_Min	SPI_Mean	0.946
cold_mortality_3.0°C	cold_mortality_presentC	0.938
SPEI_max	SPEI_mean	0.914
climate_fatalities	climate_losses_eur	0.900
avg_upload_mbps	avg_download_mbps	0.890
heat_mortality_3.0°C	heat_mortality_presentC	0.854