What Drives Data Center Location in Europe? A Multi-Factor Classification Analysis

ABSTRACT

The rapid growth of cloud computing and artificial intelligence has increased the demand for data centers worldwide, despite their significant environmental and social impacts. Understanding the factors that influence data center location decisions is therefore important. In tackling those factors, related work studied the variable predicting the presence of data centers in USA, however those results cannot be automatically assumed in europe. To address this, we studied the main factors associated with data center presence across European NUTS3 regions. We built a classification pipeline using regional economic, environmental, and infrastructure variables, applying data cleaning, feature engineering, correlation filtering, logistic regression interpretation, permutation importance, and machine learning model comparison. The findings suggest that data center presence in Europe is mainly associated with economic strength, digital connectivity, infrastructure availability, and access to key resources.

Related work:

Matlide's work studied the spatial determinants of data center location in the United States using a county-level econometric pipeline. Her dataset combined data center presence with population, GDP per capita, water supply, and electricity prices, after harmonizing the variables at county level and addressing geographic inconsistencies. The analysis used logistic regression for the probability of hosting at least one data center and additional robustness checks, including interaction effects, correlation analysis, and VIF diagnostics. The main results showed that U.S. data center presence is mainly driven by population size, GDP per capita, and water availability, while electricity prices were not significant for initial location decisions. This provides the foundation for our project, which extends the same research question to Europe using NUTS3 regions and a broader machine-learning classification pipeline.

Methodology:

Data Collection

Data center locations were identified across Europe and compiled into a single dataset. Explanatory variables were collected at the NUTS3 regional level from a range of public data sources. These included Eurostat datasets covering population, GDP, electricity prices, and national water abstraction, as well as climate-related indicators such as SPEI and SPI. All variables were harmonized using NUTS3 identifiers and merged into a unified dataset. The final dataset contained 1,344 NUTS3 regions and 39 explanatory variables.

Preprocessing

The preprocessing workflow began by removing non-predictive columns, mainly year-reference columns used only to indicate the source year of each variable. Missing values were then examined by calculating the missing-value percentage for each column (figure 1) and visualizing each variable's distribution. Skewed variables were imputed using the median, while more symmetric variables were imputed using the mean. To prevent data leakage, the dataset was first split into training (80%) and test (20%) sets, and all imputation values were learned from the training set only before being applied to the test set.

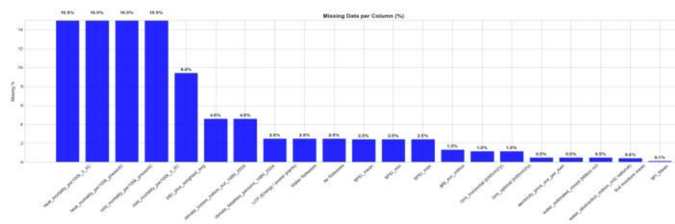


Figure 1 Missing data (%) per feature

After missing-value treatment, feature engineering was used to reduce redundancy and create more meaningful predictors. GDP per capita, computed from GDP and population, average network speed, computed from download and upload speeds, and renewable share, computed from renewable and non-renewable energy. The original variables used for the compution were then removed to reduce redundancy. Correlation-based filtering[2], built on the correlation of the features in the training set was applied, using a threshold of 0.8 to capture strong correlations while avoiding overly aggressive removal of potentially useful features. After each feature was selected for removal, the remaining correlations were rechecked to avoid dropping variables only because they were correlated with an already removed feature.

Finally, the selected features were standardized using StandardScaler rather than normalization to center variables around zero and scale them by standard deviation, and multicollinearity was checked using the Variance Inflation Factor. No VIF value exceeded 10, indicating the effectiveness of the correlation-based filtering, as no multicollinearity was severe after filtering. Following this preprocessing workflow, the feature set was reduced from 39 to 27 variables.

Feature Importance and Interpretation:

To identify the variables most strongly associated with data center presence, feature importance was evaluated using two complementary approaches.

First, logistic regression coefficients (figure 3) were examined to determine the direction and magnitude of each variable's effect on the probability of hosting a data center. Positive coefficients indicate factors associated with a higher likelihood of data center presence, while negative coefficients indicate the opposite. Statistical significance was assessed using p-values less than 0.05 to distinguish robust relationships from weaker associations.

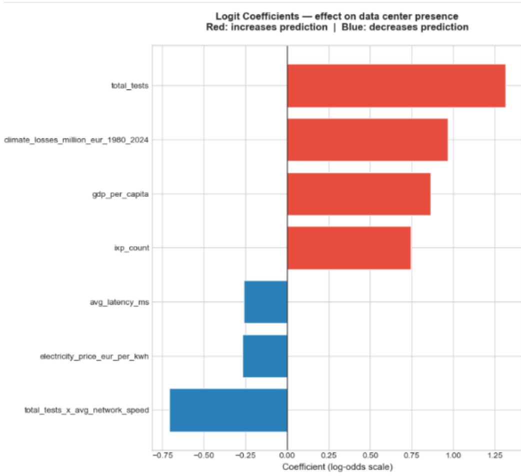


Figure 3: Logit Coefficients

Second, cross-validated permutation importance was applied using a logistic regression model on the training set, where each feature was randomly shuffled and the resulting decrease in Macro F1-score was measured (Figure 4).

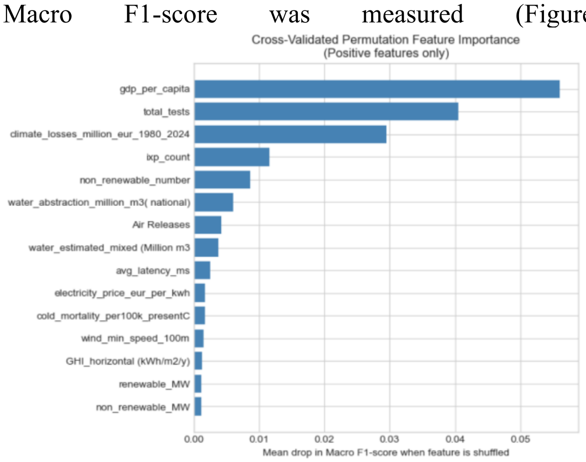


Figure 4 Drop in Macro F1 when the feature is shuffled

While logistic regression provides interpretable coefficients, it assumes primarily linear relationships between the predictors and the outcome. As some factors influencing data center location may exhibit non-linear effects, feature importance was also evaluated using an SVM-based approach to verify whether the same variables remained important when non-linear relationships were taken into account:

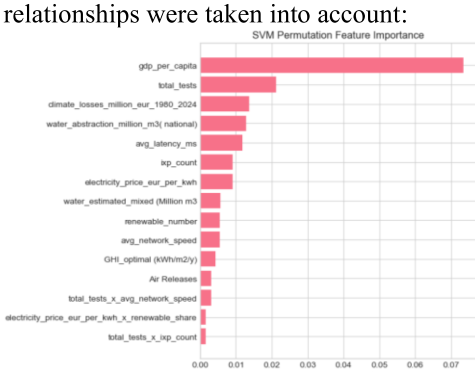
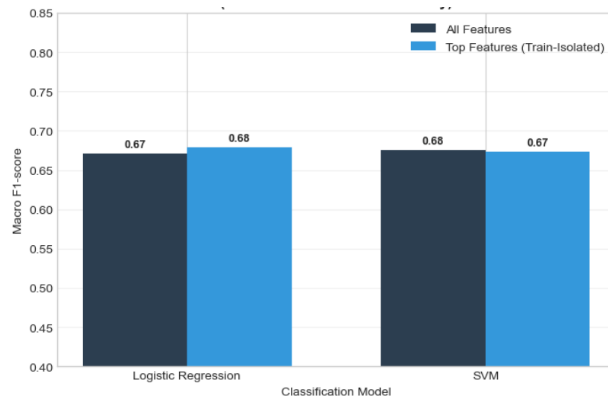


Figure 1

Macro F1-scores were nearly identical across models and feature sets, indicating stable classification performance. The small differences suggest that reducing the feature set had little impact on the models' ability to balance predictions across both classes.



Both methods consistently highlighted GDP per capita, total tests done on the network, climate-related losses, water availability indicators such as water abstraction, internet variables like exchange point (IXP) and latency, as well as electricity price among the most influential predictors of data center presence in Europe, this suggests the importance of these variables is robust and does not depend strongly on the modeling approach.

Model developement:

Three classification algorithms were evaluated: Logistic Regression, Support Vector Machines (SVM), and Random Forest. To investigate the effect of feature representation, each model was trained using three datasets: the full set of selected features after the correlation filtering, the subset of important features identified during the features interpretation stage (, and principal component analysis (PCA) features. Hyperparameter tuning was performed using GridSearchCV with five-fold stratified cross-validation on the training set. Model selection was based on the ROC AUC metric, which was chosen because it evaluates the trade-off between true positive and true negative rates across all classification thresholds and is less sensitive to class imbalance than accuracy alone. The test set was reserved exclusively for the final evaluation of the best-performing models.

PCA was evaluated as an alternative feature representation to reduce dimensionality while preserving most of the information contained in the original variables (fig 5), and to have a model we can compare to. The products of some features, like of total tests with network speed, and total tests with ixp count were introduced to get correlation-free coefficients. Based on the cumulative explained variance curve, 17 principal components were selected, preserving approximately 90% of the variance while reducing the dimensionality from 27 original features to

17 components.

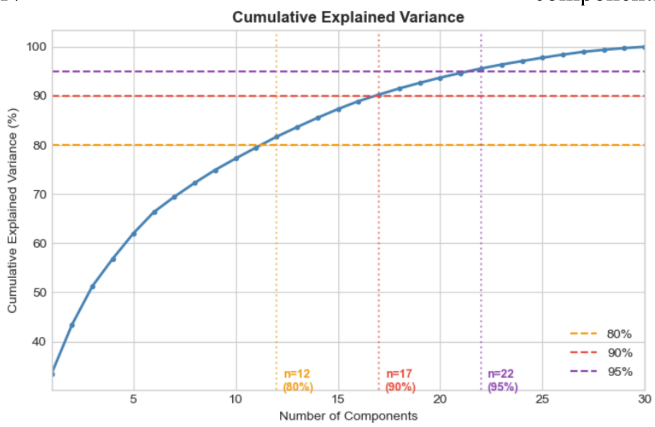


Figure 5 PCA Components' Explained Variance

Logistic Regression was optimized by testing regularization strengths C of 0.001, 0.01, 0.1, 1, 10, and 100. Both L1 and L2 regularization penalties were evaluated, while the Liblinear solver was used for all experiments.

Feature Set	Best Parameters
PCA Features	C = 100; penalty = L1; solver = liblinear
All Selected Features	C = 0.1; penalty = L1; solver = liblinear
Top Features	C = 100; penalty = L2; solver = liblinea

SVM models were optimized by testing both linear and radial basis function (RBF) kernels. The regularization parameter C was varied between 0.01 and 100, while gamma values of scale, auto, 0.001, 0.01, 0.1, and 1 were evaluated to control the influence of individual training samples.

Feature Set	Best Parameters
PCA Features	C = 100; gamma = 0.001; kernel = RBF
All Selected Features	C = 0.1; gamma = scale; kernel = linear
Top Features	C = 10; gamma = 0.01; kernel = RBF

Random Forest models were optimized by varying the number of trees (100, 200, and 500), maximum tree depth (5, 10, 20, and unlimited depth), minimum samples required for node splitting (2, 5, and 10), minimum samples per leaf node (1, 2, and 4), and the feature sampling strategy (square root, log2, and all features).

Feature Set	Best Parameters
PCA Features	max_depth = 5; max_features = sqrt; min_samples_leaf = 1; n_estimators = 300
All Selected Features	max_depth = 5; max_features = sqrt; min_samples_leaf = 4; n_estimators = 300
Top Features	max_depth = 5; max_features = sqrt; min_samples_leaf = 1; n_estimators = 300

Figure 6 compares the best cross-validated ROC AUC obtained after hyperparameter tuning for each model-feature set combination. Random Forest using all selected features achieved the highest overall performance (ROC AUC = 0.8234). However, the top-feature models also performed very strongly, with SVM and Logistic Regression reaching ROC AUC values of 0.8217 and 0.8206, respectively. PCA-based models showed lower performance across all algorithms, suggesting that dimensionality reduction removed some useful predictive information. Overall, the small differences between the full and top-feature sets indicate that a limited number of important variables captured most of the predictive information in the dataset.

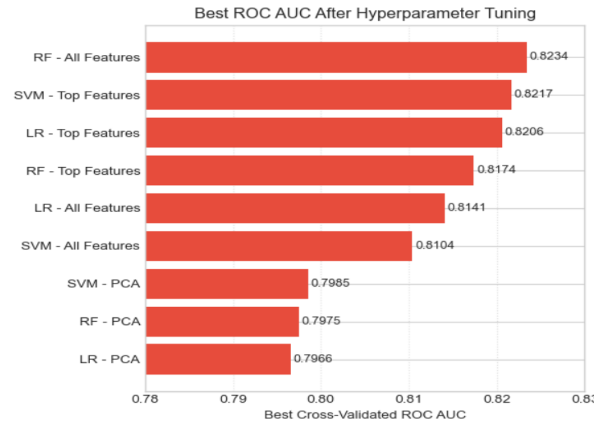


Figure 6 ROC AUC Scores

Validation and Final Evaluation:

After hyperparameter tuning, the final validation focused on the interpretable feature representations only: the complete selected feature set and the reduced top-feature subset. PCA-based models were excluded from this step because principal components are not directly interpretable in terms of the original variables, while the objective of this study is to identify the factors associated with data center presence.

For the final validation, a single tuned configuration was selected for each model based on its best overall cross-validation performance and applied consistently across both feature sets to enable a fair comparison between the complete and reduced feature representations. Performance was assessed using ROC AUC as the main metric, supported by accuracy, recall, F1-score, and confusion matrices. The following table shows the performance of each model, sorted by Roc AUC.

Model	Feature Set	Accuracy	Recall	F1	ROC AUC
Random Forest	Top Features	0.729	0.706	0.697	0.777
SVM	Top Features	0.729	0.605	0.664	0.773
Logistic Regression	Top Features	0.736	0.672	0.693	0.770
Random Forest	All Selected Features	0.688	0.672	0.656	0.768
Logistic Regression	All Selected Features	0.721	0.655	0.675	0.766
SVM	All Selected Features	0.718	0.655	0.672	0.763

Overall, all models achieved similar results, with ROC AUC values ranging from 0.763 to 0.777. The best performance was obtained by Random Forest trained on the top-feature subset (ROC AUC = 0.777), followed closely by SVM and Logistic Regression using the same reduced feature set. In general, the top-feature models performed slightly better than the full-feature models, suggesting that the selected variables retained most of the predictive information while reducing noise from less useful features. The small differences between models indicate that the classification results are stable across different algorithms. Overall, the models demonstrated a reasonable ability to distinguish between regions with and without data centers on unseen data.

Discussion:

All three models achieved very similar performance, indicating that the relationship between the predictors and data center presence is robust and can be captured effectively by both linear and non-linear algorithms. Interestingly, the best results were obtained using only the top-ranked features, suggesting that a small subset of variables contains most of the information needed for prediction.

GDP per capita emerged as one of the strongest predictors, likely because wealthier regions provide stronger economic ecosystems, better infrastructure, and higher demand for digital services. Network activity, represented by total tests, was also highly influential, suggesting that data centers tend to be located in regions with greater internet capacity. Similarly, IXP count reflects the availability of internet exchange infrastructure, which improves network efficiency and reduces communication costs. Average latency showed a negative relationship with data center presence, indicating that regions with better connectivity and lower network delays are more attractive for decision makers. Water abstraction was consistently important, which makes sense given the substantial cooling requirements of modern data centers. Climate-related losses reflect broader economic activity and infrastructure investment, but they also indicate that data centers are concentrated in highly developed regions where both economic output and exposure to climate-related damages are greater. Finally, electricity price appeared among the important predictors, although its effect was weaker than that of the economic and infrastructure variables. This suggests that while energy costs may influence operational expenses, the initial location decision is driven more strongly by the other features discussed.

The results are broadly consistent with those reported by Matilda in her county-level analysis of data center locations in the United States. In both studies, indicators of economic development and infrastructure emerged as the strongest determinants of data center presence. Bagnasco found that GDP per capita, population, and water availability significantly increased the probability of hosting a data center, while electricity prices were not significant for the initial location decision. Similarly, the European analysis identified GDP per capita, network activity, IXP infrastructure, and water-related variables among the most influential predictors, whereas electricity price played a comparatively minor role. Although the specific variables differ due to differences in data availability and geographic scale, both studies suggest that data center locations are primarily driven by structural factors such as economic strength, connectivity, and access to critical resources rather than by energy costs alone. These similarities provide evidence that the determinants of data center location may be consistent across different regions and modeling frameworks.

The results also relate to spatial organization of data centers, ranging from highly decentralized local systems to concentration in a few dominant regions. The importance of GDP per capita, network activity, IXP infrastructure, and water availability suggests that data centers are more likely to concentrate in economically developed and well-connected regions rather than being distributed evenly across all areas. This supports scenarios in which digital infrastructure becomes increasingly clustered around regions with strong economic and technical advantages, such as cases of national and foreign data centers. At the same time, the importance of water and energy-related variables highlights potential environmental and resource constraints that may influence future location decisions as demand for data centers continues to grow.

Several limitations should be acknowledged when interpreting these results. First, the analysis is conducted at the NUTS3 regional level, which may obscure finer-grained variation in data center location drivers, within a single region, infrastructure quality and economic conditions can vary considerably. Second, the binary outcome variable (presence vs. absence of a data center) does not capture the density or scale of digital infrastructure, meaning that regions hosting one small facility are treated identically to major hubs. Third, correlation-based filtering reduced the feature set from 39 to 27 variables, but excluded features should not be considered unimportant: variables such as population were removed because they are strongly correlated with retained predictors like GDP per capita, and their explanatory power is likely still indirectly represented in the final feature set rather than absent from the model. Finally, the dataset captures a single snapshot in time, so while the analysis can identify which regional characteristics are associated with data center presence, it cannot confirm that those characteristics actually caused data centers to locate there. For example, GDP per capita and data center presence may be correlated simply because both tend to develop together in advanced economies, rather than because high GDP directly attracts data centers.

Conclusion:

This study examined the main factors linked to data center presence across European NUTS3 regions. The results show that data centers are more likely to be found in economically strong, well-connected regions with good infrastructure and access to key resources. The most important predictors were GDP per capita, total network tests, climate-related losses, IXP count, water abstraction, average latency, and electricity price. The models achieved similar results, with the best ROC AUC reaching 0.777 on unseen data. This shows that the models could reasonably distinguish between regions with and without data centers. The top-feature models also performed very close to the full-feature models, suggesting that a small group of variables captured most of the useful information. Overall, the findings are consistent with Matilde’s U.S. study and suggest that data center location is mainly shaped by regional strength and infrastructure rather than by electricity price alone.

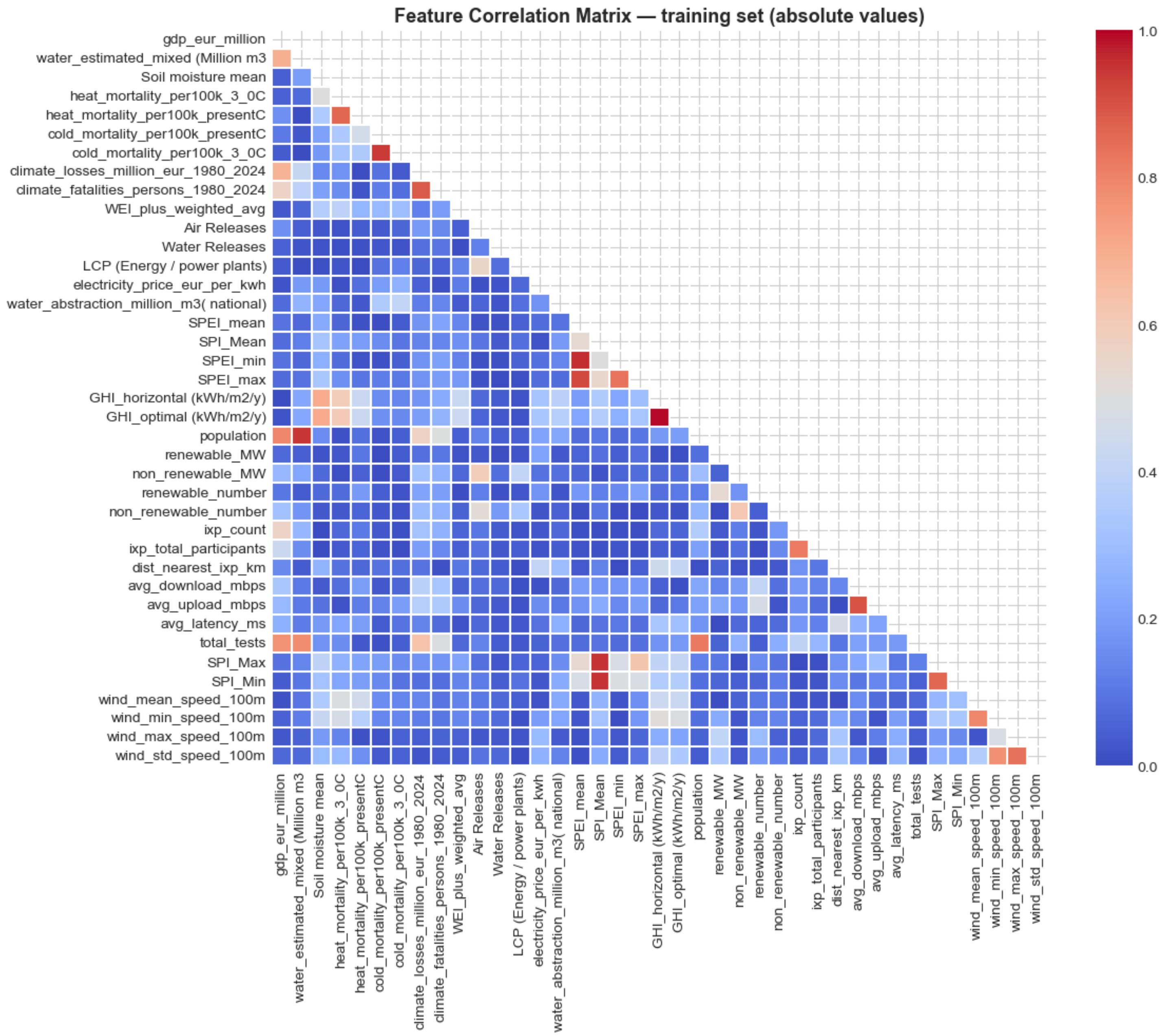
Appendix:

- Features:

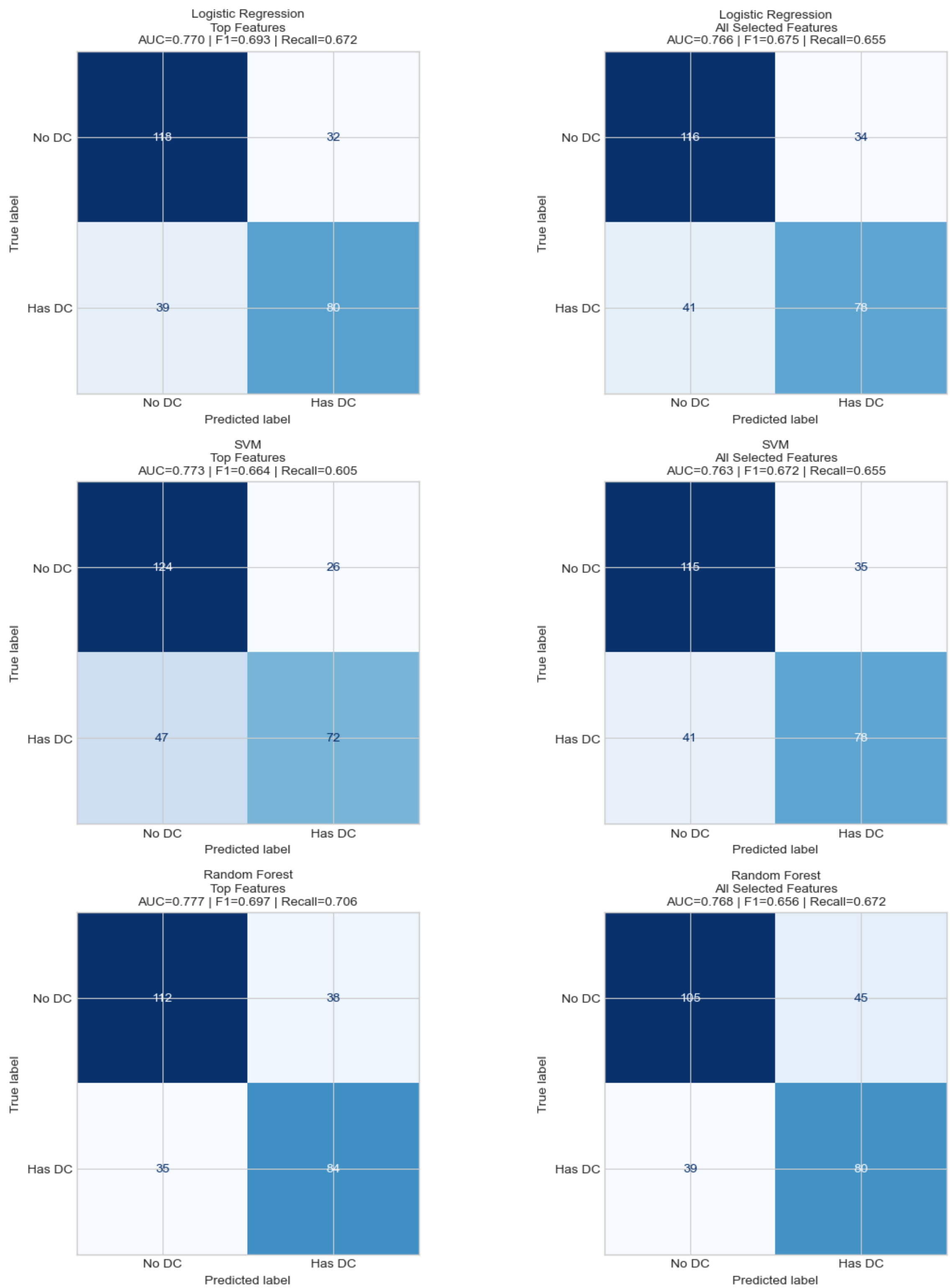
Feature	Description	Year / Reference Period	Source
NUTS_ID	NUTS region identifier	2024	Eurostat GISCO 2024
Data_Center_Count	Number of data centers in the region (average from multiple sources)	Current dataset	Aggregated from multiple sources
population	Population of the NUTS3 region	Latest available per region	Eurostat demo_r_pjanaggr3
population_year	Reference year for population value	Varies by region	Eurostat demo_r_pjanaggr3
gdp_eur_million	Gross Domestic Product (GDP)	Latest available per region	Eurostat nama_10r_3gdp
gdp_year	Reference year for GDP value	Varies by region	Eurostat nama_10r_3gdp
electricity_price_eur_per_kwh	National electricity price replicated to NUTS3 regions	Reference year in dataset	Eurostat Energy Database
electricity_price_year	Reference year for electricity price	Varies	Eurostat Energy Database
water_abstraction_million_m3	National water abstraction volume	Reference year in dataset	Eurostat Water Statistics
water_abstraction_year	Reference year for water abstraction	Varies	Eurostat Water Statistics
water_estimated_mixed	Estimated NUTS3 water abstraction using GDP and population weighting	Derived dataset	Derived from Eurostat data
renewable_MW	Installed renewable energy capacity	Current dataset	WRI Global Power Plant Database
non_renewable_MW	Installed non-renewable energy capacity	Current dataset	WRI Global Power Plant Database
renewable_number	Number of renewable power plants	Current dataset	WRI Global Power Plant Database
non_renewable_number	Number of non-renewable power plants	Current dataset	WRI Global Power Plant Database
ixp_count	Number of Internet Exchange Points (IXPs)	Current dataset	PCH IXP Data
ixp_total_participants	Total participants connected to IXPs	Current dataset	PCH IXP Data
dist_nearest_ixp_km	Distance to nearest IXP	Current dataset	PCH IXP Data
avg_download_mbps	Average internet download speed	Current dataset	Ookla Speedtest Global Performance
avg_upload_mbps	Average internet upload speed	Current dataset	Ookla Speedtest Global Performance
avg_latency_ms	Average network latency	Current dataset	Ookla Speedtest Global Performance
total_tests	Number of speed tests used in calculations	Current dataset	Ookla Speedtest Global Performance
SPEI_mean	Mean Standardized Precipitation-Evapotranspiration Index	2023	Copernicus EDO
SPEI_min	Minimum SPEI value	2023	Copernicus EDO
SPEI_max	Maximum SPEI value	2023	Copernicus EDO
SPI_Mean	Mean Standardized Precipitation Index	2023	Copernicus EDO
SPI_Min	Minimum SPI value	2023	Copernicus EDO
SPI_Max	Maximum SPI value	2023	Copernicus EDO
Soil_moisture_mean	Mean soil moisture index	2023	Copernicus EDO
Avg_summer_temp	Average summer temperature (June–August)	2005–2024 average	Copernicus Climate Data Store (ERA5)
Avg_summer_tmax	Average summer daily maximum temperature	2005–2024 average	Copernicus Climate Data Store (ERA5)
Standard_summer_tmax	Standard deviation of summer daily maximum temperature	2005–2024 average	Copernicus Climate Data Store (ERA5)

No_of_days_tmax_gt35	Average annual days with Tmax > 35°C	2005–2024 average	Copernicus Climate Data Store (ERA5)
No_hot_nights	Average annual hot nights	2005–2024 average	Copernicus Climate Data Store (ERA5)
heat_mortality_per100k_presentC	Heat-related mortality under current climate	Aug 2024 dataset	JRC PESETA V
heat_mortality_per100k_3_0C	Heat-related mortality under +3°C warming	Aug 2024 dataset	JRC PESETA V
cold_mortality_per100k_presentC	Cold-related mortality under current climate	Aug 2024 dataset	JRC PESETA V
cold_mortality_per100k_3_0C	Cold-related mortality under +3°C warming	Aug 2024 dataset	JRC PESETA V
climate_losses_million_eur_1980_2024	Climate-related economic losses	1980–2024	EEA Climate Losses Dataset
climate_fatalities_persons_1980_2024	Climate-related fatalities	1980–2024	EEA Climate Losses Dataset
WEI_plus_weighted_avg	Average Water Exploitation Index (WEI+)	2000–2023 average	EEA WEI+
Air_Releases	Total air pollutant releases	Current dataset	EEA E-PRTR
Water_Releases	Total water pollutant releases	Current dataset	EEA E-PRTR
LCP_Energy_power_plants	Emissions from Large Combustion Plants (SO2, NOx, Dust)	Current dataset	EEA LCP Register
GHI_horizontal_kWh_m2_y	Annual global horizontal solar irradiance	Current dataset	PVGIS SARAH-2
GHI_optimal_kWh_m2_y	Annual solar irradiance on optimally inclined surface	Current dataset	PVGIS SARAH-2
wind_mean_speed_100m	Mean wind speed at 100 m height	Approx. 2008–2017	Global Wind Atlas v3.0
wind_min_speed_100m	Minimum wind speed at 100 m height	Approx. 2008–2017	Global Wind Atlas v3.0
wind_max_speed_100m	Maximum wind speed at 100 m height	Approx. 2008–2017	Global Wind Atlas v3.0
wind_std_speed_100m	Standard deviation of wind speed at 100 m height	Approx. 2008–2017	Global Wind Atlas v3.0

- Correlation matrix :



- Validation Results



- Authors :
  - Karim Samarji s350602
  - Walid Alkhouly s357493
  - Joy Aramouny s335163
  - Liliane Sadek s339673
  - Ghazal Jamaliparsa s357968

- References :

  - Hao, K. (2021). The carbon footprint of US data centers. MIT Technology Review.
  - Shehabi, A. et al. (2016). United States Data Center Energy Usage Report. LBNL.
  - Avgerinou, M. et al. (2017). Trends in Data Centre Energy Consumption Under the European Code of Conduct. Energies, 10(10), 1470.
  - Eurostat (2024). NUTS — Nomenclature of Territorial Units for Statistics. European Commission.
  - Garriga, A. et al. (2023). Data centre location factors: A systematic literature review. Journal of Cleaner Production, 384, 135552.
  - M. Bagnasco, “Spatial Determinants of AI Data Center Location: A County-Level Econometric Analysis in the United States,” Master’s Thesis, Department of Management and Production Engineering, Politecnico di Torino, Turin, Italy, Academic Year 2025–2026