Predicting Data Centers Presence in Europe

Team JAM (ID 11)

1. ABSTRACT

Studying the variables that predict the presence of data centers in Europe is increasingly important, because the number of these facilities is rapidly growing and where they are located matters for host communities, energy, water systems and regional economies. Similar work focused on predicting the presence of data centers and studying their driving factors in the United States of America has been done, but comparable studies for Europe are scarce. Therefore, to tackle the problem of studying the driving factors and predicting data center locations in the European NUTS3 regions, we made the following contributions:

- collection of data about the data centers already present in Europe;
- collection of data about socio-economical, infrastructure and environmental/climate variables for each NUTS3 region present in Europe;
- data pre-processing to remove duplicates and handle missing values;
- model training to predict the presence of a data center in a NUTS3 region and study of variables that drive the prediction.

2. RELATED WORK

The rapid growth of artificial intelligence and cloud computing has increased interest in understanding the factors that drive data center location. Existing studies mainly focus on the environmental and operational challenges of data centers, including energy consumption, water use, and sustainability, while industry reports identify energy costs, water availability, climate conditions, network connectivity, human capital, land availability, and regulatory frameworks as the main site-selection criteria (Bashir & Donti, 2024; Kamiya & Bertoldi, 2024; Omdia, 2024; Pengfei et al., 2025).

Empirical evidence, however, remains limited. A notable contribution is (Bagnasco, 2026), which developed a county-level model for the United States to predict the presence of data center using electricity prices, water availability, GDP and population. The study suggests that economic development, population size, and infrastructure availability are important drivers of data center localization, while energy and environmental factors also play a relevant role in location decisions.

To sum up, existing studies provide useful insights into the factors associated with data center location, but their findings are based on the United States and cannot be readily extended to European NUTS3 regions.

3. METHODOLOGY

3.1 Data collection

First, we created a main dataset focusing on data centers, including their NUTS3 level, coordinates, and location for existing facilities across European countries. For the second part, we collected three main types of data: population, GDP, and additional variables like solar potential for both flat and tilted surfaces. The data were organized to fit the NUTS3 level. For population and GDP, we ensured they aligned perfectly with the NUTS3 map, dropping any records that did not fit. Sunlight data came as a continuous grid across the map. To address this, we applied a zonal statistics method to find the average sunlight within each NUTS3 boundary. Afterward, we cleaned up the data by eliminating areas with missing identifiers. Incomplete country records were supplemented with data from adjacent years. After completing this part, the consolidated dataset was compiled by combining the information collected by all the student teams using the NUTS3 code, resulting in a single complete table of 1,345 regions, each described by 45 variables spanning three families. We considered socio-economic (e.g. population, GDP), infrastructure (e.g. internet-exchange-point count, network speeds) and environmental/climate (e.g. solar radiation, wind speed, drought indices, heat- and cold-related mortality) variables to build the model for predicting the data center locations.

3.2 Data pre-processing

To predict data center presence, we derived a binary target variable called "Presence." We marked this as "Yes" if the region had at least one data center and "No" if it did not. To prevent target leakage, we then dropped the actual data center count column, along with a few other columns like region IDs and specific reporting years that did not hold any actual predictive value. Next, we converted our "Yes" and "No" targets into 1 and 0 and split our data, reserving 20% for a testing set while making sure the class ratio remained balanced. We then worked on our missing data by looking at the distributions of our features to figure out the best way to fill in the gaps. For the imputation step, we used a median strategy to fill in missing values for highly skewed features, like soil moisture mean (Figure 1). However, for variables with a higher percentage of missing data, such as climate fatalities and heat/cold-related mortality, we used an iterative imputation method to estimate the missing numbers by seeing how they related to the other features rather than from a single variable.

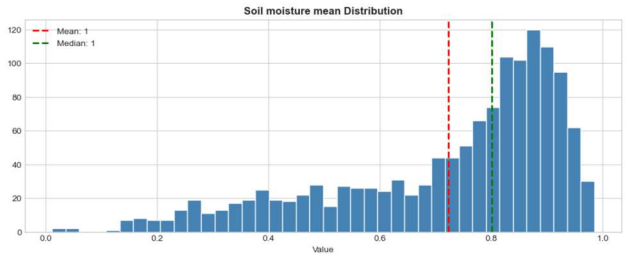

Figure 1 - Skewed distribution for Soil Moisture mean.

3.3 Feature Analysis & Selection

With the dataset split, we computed the correlation matrix on the training set to find out where redundancy existed. We used a correlation coefficient of 0.7 as the threshold to handle the challenge of multicollinearity that could bias the models. About 22 pairs of correlated features showed up, including duplicated solar energy measures and overlapping wind statistics.

Based on those results, we rebuilt the feature set, having both redundancy and real-world application in mind. We dropped one variable from each correlated pair; for the wind measures, for instance, we kept only maximum speed, since that is what reflects both renewable energy potential and the structural safety limits a facility must withstand. We collapsed the SPEI (Standardized Precipitation Evapotranspiration Index) and SPI (Standardized Precipitation Index) min/max columns into ranges to capture volatility instead of absolute levels, and replaced raw future mortality projections with the gap relative to present-day levels, since that change is what represents a region's added climate-risk exposure under warming. We also folded economic losses and climate fatalities into a single z-scored climate-risk index. For absolute totals like estimated water abstraction, number of tests used for network metrics and GDP, we normalized them to per-capita terms to capture intensity and make the regions comparable regardless of population, but since all three then shared the same denominator and became artificially correlated, we kept only GDP per capita. For the heavily skewed variables, we simply log-transformed them to stabilize their variance.

Finally, we ran a VIF (Variance Inflation Factor) analysis to check for any remaining multicollinearity, and every feature came back below the threshold of 5, confirming it had been resolved.

3.4 Modelling design

We set up two distinct tracks for our modeling – one to understand data relationships, the other to predict new locations. First, we used statsmodels to fit a logistic regression model, which provided easy-to-read coefficients, p-values, and odds ratios. To maximize interpretability, we experimented with different feature combinations but only kept the simplest version that cleared a pseudo-R^2 of 0.2 while retaining solely the variables significant at a p-value below 0.05. We selected this value of pseudo-R^2 to increase the interpretability of the coefficients and obtain a meaningful model.

The second track was about boosting prediction performance. We first standardized all the features with a z-score standardization, to prevent the model from learning from differently scaled variables and used PCA (Principal Component Analysis) to lower the dimensionality while keeping at least 90% of the original cumulative explained variance. With dimensionality reduction, we were able to make predictive operations easier and faster.

We then tested and compared four classifiers: Logistic Regression, Decision Tree, Random Forest, and SVM and fine-tuned each with grid search under stratified 5-fold cross-validation. Since missing a true host region costs more than raising a false alarm, the whole evaluation leaned toward catching the positive class. This shaped two choices: we improved the positive class's F1-score rather than accuracy alone, and applied class-balanced weights to keep the model fair and unbiased given the uneven split between regions with and without data centers. To assess our best classifier, we examined its confusion matrix and ROC curve on the held-out test set. To strengthen recall so that as few real host regions as possible slipped through undetected, we carried out decision threshold tuning, testing a range of probability thresholds from 0.20 to 0.50 and tracking how precision, recall, and F1 shifted for the positive class. From these, we selected the operating point that gave the best balance between recovering true host regions and keeping false alarms in check.

4. RESULTS AND EVALUATION

4.1 The drivers of presence

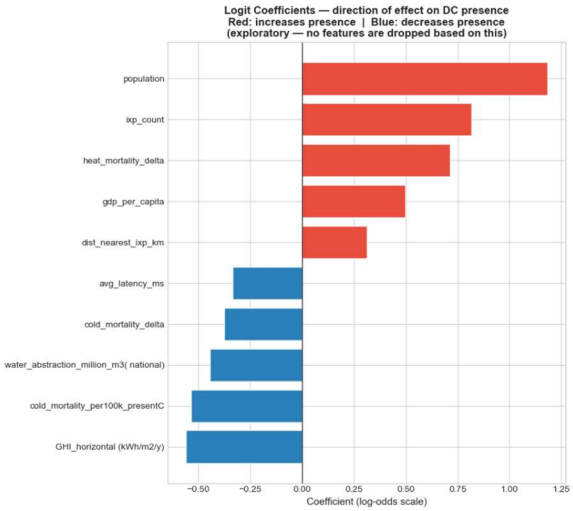

Figure 2: Logit coefficients chart.

Figure 2 gives an exploratory view of how strongly each feature influences data center location. The variables that contribute most positively to the prediction are:

- population: most recent available value per NUTS3 region. More populous regions generate higher demand for digital services and provide stronger economic

ecosystems, making them attractive locations for data centers;

- ixp_count: number of IXP (Internet Exchange Point) in the region. A higher number of IXPs indicates an advanced network infrastructure and connectivity, which are important factors for the data centers;
- heat_mortality_delta: it is a difficult feature to interpret. It was created during the analysis of the feature correlation from 'heat_mortality_per100k_3_0C' and 'heat_mortality_per100k_presentC'. A possible interpretation is that the positive effect suggests data centers are more likely to be located in densely populated and economically developed regions that also happen to be more vulnerable to future heat impacts, meaning the feature likely acts as a proxy for development rather than a real preference for heat-vulnerable areas;
- gdp_per_capita: wealthier regions typically have greater investments in technology, higher business activity, and better infrastructure, all of which support data center development;
- dist_nearest_ixp_km: distance to the nearest IXP. This is a difficult feature to interpret, since we would normally expect data centers to be located closer to IXPs to benefit from lower latency and better connectivity. The positive sign may therefore reflect noise or a confounding effect with large, sparsely connected regions, rather than a genuine preference for being far from connectivity.

The features with the negative contributions are:

- avg_latency_ms: average network latency. Higher latency reflects lower connectivity quality, reducing a region's attractiveness for data centers;
- cold_mortality_delta: this variable is also difficult to interpret. It comes from the analysis of the feature correlation from 'cold_mortality_per100k_3_0C' and 'cold_mortality_per100k_presentC', the negative effect suggests that regions more affected by climate-related changes in cold mortality are less likely to host data centers, potentially reflecting colder and less urbanized areas;
- water_abstraction_million_m3: water abstraction at the national level. Higher water abstraction levels may characterize regions with high-water stress, therefore not suitable for the water demands of data centers;
- cold_mortality_per100k_presentC: cold-related mortality under present climate conditions. This suggests that regions experiencing harsher climatic conditions tend to less likely host data centers;
- GHI_horizontal: annual global horizontal irradiance averaged over the region. Higher solar irradiance is associated with warmer regions, while data centers are often located in cooler regions where cooling costs are lower.

After the exploratory analysis, we then searched for the most parsimonious specification that kept only statistically significant predictors. The simplest interpretable model we obtained from the statsmodels logistic regression is:

Presence ~ log(population) + water_abstraction + ixp_count + cold_mortality.

The model uses four main predictors: population, water abstraction (that refers to the removal or diversion of water from natural sources), IXP count and current cold mortality. By analyzing the odds ratios, we have observed that a one-unit increase in log-population led to about 3.4 times higher chance of having a data center in that NUTS3 region. Additionally, adding another IXP increased these odds by a factor of 3.2, making these effects the biggest in the model. Meanwhile, higher cold mortality showed a small negative link, and although water abstraction is statistically significant, it is small in impact. Comparing our results with Bagnasco (2026), who found population, GDP and water availability to be the main drivers in the US, we noticed that in our Europe model, it is really economic size, digital connectivity and infrastructures that are the main driving forces, and climate and environmental factors play a marginal role. Although water was a major driver in the US, in our European model water abstraction hardly matters at all, which suggests it plays a different role on each side of the Atlantic. As for cold mortality having a negative effect, it could be explained by the fact that when the climate is too cold, freezing could interfere with cooling-water systems and the electricity supply, making such a location less appealing to site a data center.

4.2 Predictive performance

We looked at four classifiers on a test set, shown in Table 1. While Logistic Regression achieved the highest ROC-AUC (0.74), Random Forest yielded the best overall performance – it led in accuracy at 0.70 and matched that with a strong positive-class F1 score of 0.67. The decision tree increased the recall to 0.71 but lost on precision, dropping its F1 score. To prioritize the identification of true host regions while limiting false positives, we selected Random Forest as the final model.

Model	Accuracy	F1 (cls1)	Recall	Precision	ROC-AUC
Random Forest	0.70	0.67	0.68	0.65	0.71
Logistic Regression	0.67	0.60	0.56	0.65	0.74
Decision Tree	0.67	0.66	0.71	0.61	0.70
SVM	0.66	0.60	0.56	0.63	0.72

Table 1: Test-set performance of the four tuned classifiers

Figure 3 shows the ROC curve of the Random Forest classifier. The model achieves an AUC of 0.71, indicating an acceptable ability to discriminate between NUTS3 regions hosting at least one data center and regions without data centers. The ROC curve consistently lies above the random-classification baseline, showing that the selected socio-economic, infrastructure, and

climatic variables contain useful information for predicting data center presence. However, the moderate AUC value suggests that additional factors could influence the presence of data centers across European regions.

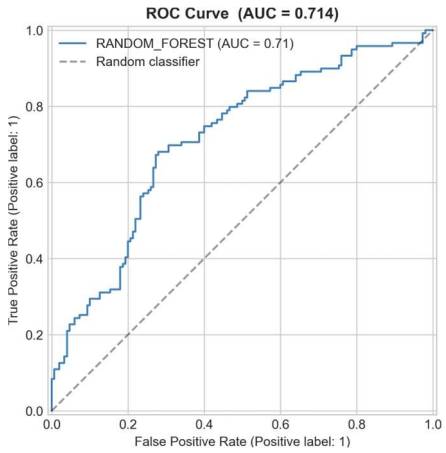

Figure 3 - ROC Curve for Random Forest model on the Test Set

4.3 Threshold tuning

The initial performance under the default 0.50 threshold is illustrated in Figure 4. At this setting, the model follows a more conservative decision rule, resulting in a lower number of false positives (43) and correctly identifying 81 true host regions. However, this comes at the cost of 38 false negatives, which in a screening context represent a critical limitation, as potentially relevant host regions stay undetected.

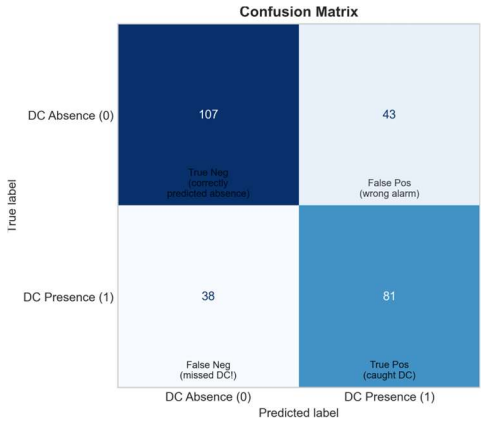

Figure 4 - Confusion Matrix before the threshold tuning

When the decision threshold is reduced to 0.35, as shown in Figure 5Figure 4, the model becomes more sensitive, leading to a noticeable shift in the classification outcomes. In particular, the number of false negatives decreases substantially from 38 to 20, meaning that 18 previously missed host regions are now correctly identified. This improvement is reflected in an increase in true positives from 81 to 99, raising the recall to 83.2%.

This gain in sensitivity is accompanied by the expected trade-off in precision, as false positives increased from 43 to 77, while true negatives decrease from 107 to 73. Overall, the comparison between the two thresholds provides a clear empirical justification for selecting a lower cutoff of 0.35, since the increase in false alarms is operationally acceptable considering the substantial reduction in missed host regions and the corresponding improvement in screening reliability.

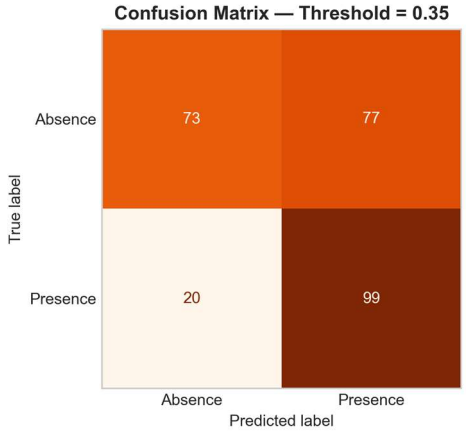

Figure 5 - Confusion Matrix after the threshold tuning

5. DISCUSSION

Our findings align closely with what Bagnasco (2026) identified as the main drivers of data centers in the US, showing that similar patterns hold in both cases. What stands out is that IXP count comes through almost as strongly as population, making the vague connectivity claim from public reports concrete and measurable.

An interesting perspective is how these drivers impact policy, and that depends greatly on who controls where data centers get built. National policymakers could steer investment toward areas with solid infrastructure to draw both domestic and foreign interests, though it could risk friction with host communities. If the decision is left to regional or local authorities, communities get to decide on trade-offs, but could become too weak to attract investments.

The driving variables we found can also be divided into two groups, depending on whether policy can actually move them. Locally actionable variables include digital connectivity (avg_latency_ms, dist_nearest_ixp_km, ixp_count) and water management (water_abstraction_million_m3), which can be shaped by targeted regional investments. In contrast, globally driven variables act as fixed environmental and structural baselines, encompassing socioeconomic indicators (population, gdp_per_capita) and climate metrics. Climate is the interesting case. It looks fixed at the regional level, but whether it stays that way depends on who governs siting. Since the Paris Agreement works through national commitments, the EU is well placed to enforce sustainability standards on data center siting, as it already does through the Code of Conduct for data centers under the Corporate Sustainability Reporting Directive.

TEAM JAM (ID: 11):

Matteo Decandia s353983

Jarno Valfrè s358249

Samin Yasar Rahman s353394

Abel Inalegwu Onuh s363268

Maisha Nusrat s363274

REFERENCES

Bagnasco, M. (2026). *Spatial Determinants of AI Data Center Location: A County-Level Econometric Analysis in the United States*. Politecnico di Torino.

Bashir, N., & Donti, P. (2024). *The Climate and Sustainability Implications of Generative AI.*

Kamiya, G., & Bertoldi, P. (2024). *Energy consumption in data centres and broadband communication networks in the EU*. Publications Office of the European Union. https://doi.org/doi/10.2760/706491

Omdia. (2024). *Best Data Center Locations 2024*.

Pengfei, L., Jianyi, Y., Mohammad A., I., & Shaolei, R. (2025). *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*.