Politecnico di Torino

## Group 8

1. Mohammad Mahdi Molla Jafari - S328723

2. Mohammadreza Goudarzi - S324343

3. Asma Zeynalpoor - S358540

4. Negin Alsadat Hosseini Gorji - S360179

5. Armin Akbari - S363457

2025/2026

# Predicting the Presence of Data Centers in European NUTS3 Regions:

## An Infrastructure and Demand-Driven Machine Learning Approach

## 1. Abstract / Introduction


Data centers are an essential part of today's digital economy, yet their distribution across Europe is far from uniform. While some NUTS3 regions host multiple facilities, many others do not have any. This project explores the factors that influence whether a data center is located in a particular region across 1,345 European NUTS3 regions.

The study examines whether data center locations are mainly driven by demand and infrastructure characteristics—such as GDP, population size, internet usage, and network quality—or by operational and environmental conditions, including electricity costs and climate-related risks. We expect that regions with stronger economies and better digital connectivity are more likely to attract data centers.

To investigate this, we model the problem as a binary classification task. A region is assigned a value of 1 if it hosts at least one data center (has_dc = 1) and 0 otherwise (has_dc = 0). We apply both Logistic Regression, using statsmodels to interpret the effects of individual variables, and Random Forest, using scikit-learn to maximize predictive performance.
The optimized Random Forest model achieves an accuracy of about 74% and a ROC–AUC score of 0.784. The findings indicate that GDP, internet usage, population, and network quality are the most influential factors in predicting data center presence.

Overall, the results suggest that the location of data centers in Europe is primarily shaped by economic demand and digital infrastructure rather than by operational or environmental considerations.


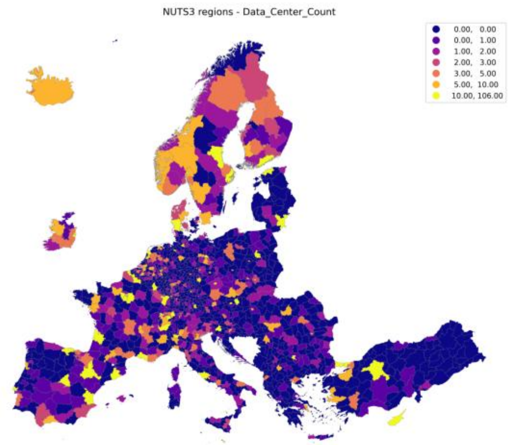


*Figure 1: Spatial distribution of data centers across European NUTS3 regions, illustrating the total count of facilities per region (Data_Center_Count) prior to binary transformation.*

## 2. Related Work

Previous studies on data center location show that different factors can influence where these facilities are built. Economic conditions, infrastructure quality, energy costs, and environmental constraints are usually considered important. In the U.S. county-level framework developed by Bagnasco, data center presence is studied using variables such as population, GDP, electricity prices, water availability, and infrastructure-related indicators. This framework is useful for our project because it connects data center location with both regional demand and operational conditions.

In this project, we use a similar logic but apply it to European NUTS3 regions instead of U.S. counties. The goal is not to directly compare Europe and the United States, but to use the same general idea to understand which variables help explain data center presence in Europe.

Based on this framework, we divide the predictors into three groups. The first group includes economic and market variables, such as GDP and population, which represent regional size and demand. The second group includes network variables, such as latency, download speed, upload speed, total internet tests, IXP count, and distance to the nearest IXP, which describe digital connectivity. The third group includes energy and environmental variables, such as electricity price, renewable and non-renewable capacity, and climate-related indicators. These groups help us test whether data center presence in Europe is mainly related to demand and connectivity, or whether energy and environmental factors also play an important role.

## 3. Methods

The methodology follows three main steps: data preparation, interpretable modeling, and predictive modeling.

First, the dataset was prepared for binary classification. The original variable `Data_Center_Count` was converted into the binary target `has_dc`. Regions with at least one data center were labeled as 1, while regions with zero data centers were labeled as 0. This choice focuses on the extensive margin of data center location: whether a region hosts a facility or not. Environmental and climate-related indicators were already included in the provided NUTS3-level dataset, but they did not emerge among the strongest predictors in the final models.

Before modeling, irrelevant year columns were removed because they represent metadata rather than useful predictors. Missing numerical values were handled using mean imputation, and numerical features were standardized. To avoid data leakage, the train-test split was performed before imputation and scaling. Therefore, the imputer and scaler were fitted only on the training data and then applied to the test data.

Second, an interpretable Logistic Regression model was trained using `statsmodels`. This model was used to identify statistically significant predictors through coefficients and p-values. It provides a clear interpretation of which variables are positively or negatively associated with the probability of hosting at least one data center.

Third, a Random Forest classifier was trained using `scikit-learn`. This model was selected because it can capture non-linear relationships and interactions between variables. Hyperparameters were tuned using 5-fold GridSearchCV, optimizing the F1-score. The best configuration was:

- `n_estimators = 100`
- `max_depth = 10`
- `min_samples_split = 5`

The selected explanatory variables include GDP, population, IXP count, distance to the nearest IXP, average download speed, average upload speed, average latency, total internet tests, electricity price, renewable capacity, and non-renewable capacity.

**Table 1: Feature Groups**

| Group | Key Variables |
|---|---|
| Economic | GDP (EUR M), population |
| Network | IXP count, latency, DL/UL speed, tests |
| Energy | renewable/non-renew. MW, price (EUR/kWh) |
| Climate | SPEI, SPI, temperature |

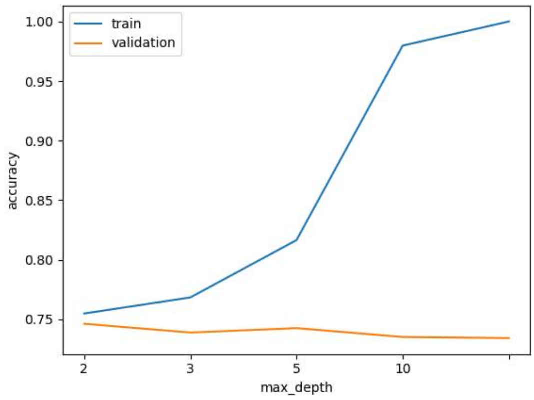


*Figure 2: Validation curve demonstrating model performance across varying tree depths, illustrating the divergence between training and validation accuracy and the onset of overfitting at higher depths.*

## 4. Results

The inferential Logistic Regression model confirmed that economic size is the most statistically significant driver of data center presence, with GDP showing the strongest effect ($p < 0.001$). Network-related variables also contributed to the explanation, particularly IXP count and latency, suggesting that connectivity quality plays an important role in data center location.

The optimized Random Forest model outperformed the Logistic Regression baseline, achieving an accuracy of 74.3% and a ROC-AUC score of 0.784. This indicates that the model can reasonably distinguish between NUTS3 regions with and without data centers.

**Table 2: Random Forest Test Performance**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.76 | 0.72 | 0.74 |
| Recall | 0.79 | 0.69 | 0.74 |
| F1 | 0.77 | 0.70 | 0.74 |
| AUC-ROC | — | — | 0.784 |

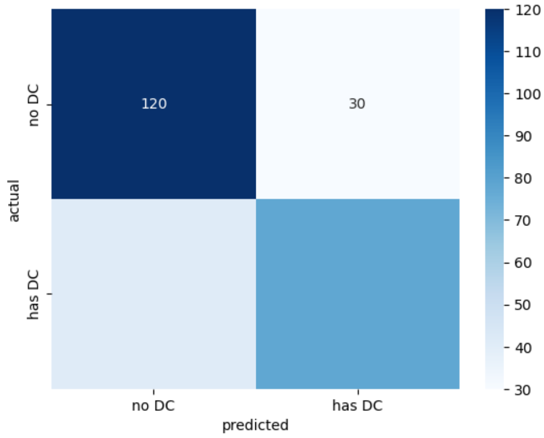


*Figure 3: Confusion matrix for the optimized Random Forest model on the test set*

Feature importance from the Random Forest revealed a clear hierarchy of drivers. GDP was the dominant predictor, followed by digital demand indicators such as total internet tests and population. Network quality variables, including latency, download speed, and distance to the nearest IXP, formed the second most important group. In contrast, electricity price, renewable capacity, and climate-related variables showed lower predictive importance.

Complementing Random Forest's importance rankings, Logistic Regression reveals directional effects: GDP (+4.01) and IXP count (+0.58) positively influence data center presence, whereas average latency (-0.32) has a negative impact.

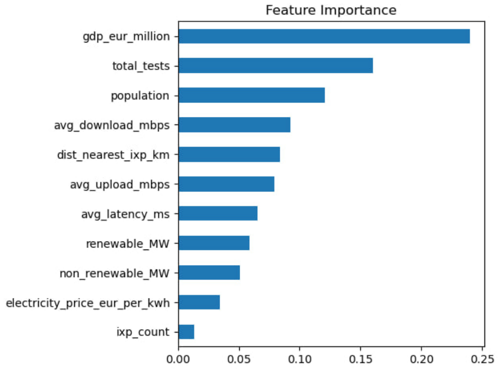


*Figure 4: Random Forest Feature Importance, illustrating the dominance of GDP and demand-side indicators, with network quality variables in the mid-range and energy and climate metrics ranked lowest.*

## 5. Discussion and Conclusion

The results largely support the main hypothesis that the presence of data centers across European NUTS3 regions is primarily driven by demand and infrastructure factors. Among all variables considered, GDP, internet usage, population size, and network quality emerge as the most important predictors. In contrast, electricity prices and climate-related variables have a much weaker influence on the model's predictions. This suggests that data centers in Europe are more likely to be located in regions with strong economic activity, large user bases, and reliable digital infrastructure than in areas that simply offer lower energy costs or lower environmental risks.

These findings are consistent with the U.S. county-level framework discussed in the literature review, which also highlights the importance of economic scale and infrastructure in shaping data center location decisions. However, this study does not attempt to make a direct comparison between Europe and the United States. Instead, the U.S. framework serves as a useful reference for interpreting the European results.

The findings can also be interpreted through the lens of four possible data center location logics: local, cheap, national, and foreign. The observed distribution does not strongly support a local data center logic, as facilities are concentrated in specific regions rather than being evenly distributed across Europe. Similarly, there is limited evidence for a cheap data center logic, since electricity prices have relatively low predictive importance. Instead, the results are more consistent with a market-driven or foreign data center logic, where large operators choose economically strong and highly connected regions in order to serve wider markets efficiently. A national logic may also play a role, but the available dataset does not contain information on ownership structures, government policies, or planning regulations that would allow this hypothesis to be tested directly.

From a policy perspective, these findings suggest that low energy prices or abundant land alone may not be enough to attract data center investment. Regions that currently host few or no facilities may need to strengthen their digital infrastructure, improve connectivity to Internet Exchange Points (IXPs), and invest in broader digital development strategies. At the same time, the results should be interpreted with caution. The analysis identifies statistical associations rather than causal relationships, and the binary target variable only captures whether a region hosts at least one data center, without accounting for differences in the number or size of facilities. Future research could address these limitations by modeling data center counts directly and by incorporating additional variables related to land availability, electricity grid capacity, water stress, fiber-optic networks, and local policy incentives.

## 6. References


[1] Bagnasco, M. (2026). Spatial Determinants of AI Data Center Location: A County-Level Econometric Analysis in the United States [Master's thesis, Politecnico di Torino].

[2] Eurostat. (2024). *GDP at current market prices by NUTS 3 regions*. Retrieved from https://ec.europa.eu/eurostat/data-browser/view/nama_10r_3gdp/default/table

[3] Eurostat. (2024). *Population on 1 January by NUTS 3 region*. Retrieved from https://ec.europa.eu/eurostat/data-browser/view/demo_r_pjanaggr3/default/table

[4] Packet Clearing House (PCH). (2024). *Internet Exchange Point Data*. Retrieved from https://www.pch.net/ixp/data

[5] Ookla. (2024). *Speedtest Global Performance Dataset*. AWS Open Data. Retrieved from https://registry.opendata.aws/speedtest-global-performance

[6] World Resources Institute (WRI). (2024). *Global Power Plant Database*. Retrieved from https://datasets.wri.org/datasets/global-power-plant-database

[7] Joint Research Centre (JRC). (2024). *PESETA V: Temperature-Related Mortality*. European Commission. Retrieved from https://joint-research-centre.ec.europa.eu/