

Nov, 2025

GRINS DISCUSSION PAPER SERIES DP N° 70/2025

ISSN 3035-5567



A multi-stage data quality management system for Energy Performance Certificates using statistical and machine learning methods

DP N° 70/2025

Authors:

Marco Sandri, Alessio Pirola, Paola Zuccolotto

A multi-stage data quality management system for Energy Performance Certificates using statistical and machine learning methods

Marco Sandri, Alessio Pirola, Paola Zuccolotto

KEYWORDS

Machine learning

energy performance

ACKNOWLEDGEMENTS

This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS - Growing Resilient, INclusive and Sustainable project (GRINS PE00000018). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

CITE THIS WORK

Author(s): Marco Sandri, Alessio Pirola, Paola Zuccolotto. Title: A multi-stage data quality management system for Energy Performance Certificates using statistical and machine learning methods. Publication Date: 2025.

Data quality is a critical prerequisite for the effective use of Energy Performance Certificates (EPCs), called Attestati di Prestazione Energetica (APE) in Italian regulations, for research purposes, monitoring activities, and policy evaluation. Large administrative EPC datasets often exhibit heterogeneous quality, with potential inaccuracies or inconsistencies that may compromise subsequent analyses.

To address this issue, we introduce a data quality control framework designed to evaluate the reliability of each certificate through a comprehensive set of checks applied at the individual-record level. The outcome of this process is a global penalty index that quantifies the extent of detected issues: higher values denote poorer data quality. This approach offers a systematic and reproducible way to summarize data reliability and to support the identification of certificates that warrant correction or additional review.

Within the proposed framework, we distinguish between two broad families of data quality checks. The first group consists of deterministic, rule-based validation checks, which rely on explicit logical constraints derived from the regulatory definition of EPCs and from basic physical relationships. These checks verify that required fields are present, that reported values lie within admissible domains, and that predefined relationships between fields are exactly satisfied. The second group comprises statistical, data-driven diagnostics, which assess the plausibility of the reported information in relation to the empirical distributions and multivariate patterns observed in the dataset. These diagnostics are designed to flag records that are formally admissible according to the deterministic rules but appear highly atypical or improbable, and are therefore at increased risk of measurement, coding, or transcription errors. Taken together, deterministic and statistical checks provide complementary perspectives on data quality, enabling a more comprehensive characterization of potential issues at the level of each individual certificate.

Nov, 2025

GRINS DISCUSSION PAPER SERIES DP N° 70/2025

ISSN 3035-5567

Within both the deterministic and statistical components, we distinguish between univariable and multivariable checks. Univariable checks focus on individual fields, such as detecting values outside admissible ranges or identifying unusually extreme observations relative to the empirical distribution. Multivariable checks evaluate the joint coherence of two or more variables, for example verifying mutually exclusive categories that should admit only one valid option, or detecting atypical combinations that deviate from expected multivariate patterns.

The underlying rationale of the proposed system is to examine each certificate through multiple complementary perspectives, enabling the detection of a wide range of data quality issues. Simple rule-based checks are designed to capture explicit inconsistencies—such as violated logical constraints or values falling outside admissible domains—that typically signal coding errors or clearly implausible entries. In parallel, both univariable and multivariable statistical diagnostics are used to identify more subtle anomalies that cannot be revealed through logical rules alone. These include values that are individually plausible but highly atypical in the context of the empirical distribution, as well as combinations of fields that deviate from expected multivariate patterns. By integrating these different layers of validation, the framework is capable of identifying issues that span from straightforward incongruities to complex, hard-to-detect anomalies arising from the joint behavior of multiple fields.

It is important to emphasize that the resulting penalty score does not represent a deterministic measure of data quality. Because the framework combines rule-based validations with statistical diagnostics, the score should be interpreted as an indicator of the likelihood that a certificate may contain inaccuracies, rather than a definitive classification of records as valid or invalid. Deterministic checks contribute directly to the score by flagging explicit inconsistencies, whereas statistical checks reflect deviations from empirical patterns that may arise from genuine heterogeneity as well as from potential errors. As a consequence, higher scores identify certificates that warrant closer examination, but they do not imply that the underlying data are necessarily incorrect. The penalty score therefore functions as a screening tool that prioritizes potentially problematic records and supports more informed decisions on data cleaning, verification, or subsequent exclusion.

**A multi-stage data quality management system
for Energy Performance Certificates
using statistical and machine learning methods**

Marco Sandri¹, Alessio Pirola², Paola Zuccolotto²

¹ Big & Open Data Innovation Laboratory, University of Brescia, C.da S. Chiara 50, 25122 Brescia, Italy

² Department of Economics and Management, University of Brescia, C.da S. Chiara 50, 25122 Brescia, Italy

The research is funded by the European Union - Next Generation EU, in the framework of the GRINS - Growing Resilient, Inclusive and Sustainable Project (GRINS PE00000018 CUP D73C24000400001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union and the European Union cannot be held responsible for them.

1. Introduction

Data quality is a critical prerequisite for the effective use of Energy Performance Certificates (EPCs), called *Attestati di Prestazione Energetica* (APE) in Italian regulations, for research purposes, monitoring activities, and policy evaluation. Large administrative EPC datasets often exhibit heterogeneous quality, with potential inaccuracies or inconsistencies that may compromise subsequent analyses.

To address this issue, we introduce a data quality control framework designed to evaluate the reliability of each certificate through a comprehensive set of checks applied at the individual-record level. The outcome of this process is a global penalty index that quantifies the extent of detected issues: higher values denote poorer data quality. This approach offers a systematic and reproducible way to summarize data reliability and to support the identification of certificates that warrant correction or additional review.

Within the proposed framework, we distinguish between two broad families of data quality checks. The first group consists of deterministic, rule-based validation checks, which rely on explicit logical constraints derived from the regulatory definition of EPCs and from basic physical relationships. These checks verify that required fields are present, that reported values lie within admissible domains, and that predefined relationships between fields are exactly satisfied. The second group comprises statistical, data-driven diagnostics, which assess the plausibility of the reported information in relation to the empirical distributions and multivariate patterns observed in the dataset. These diagnostics are designed to flag records that are formally admissible according to the deterministic rules but appear highly atypical or improbable, and are therefore at increased risk of measurement, coding, or transcription errors. Taken together, deterministic and statistical checks provide complementary perspectives on data quality, enabling a more comprehensive characterization of potential issues at the level of each individual certificate.

Within both the deterministic and statistical components, we distinguish between univariable and multivariable checks. Univariable checks focus on individual fields, such as detecting values outside admissible ranges or identifying unusually extreme observations relative to the empirical distribution. Multivariable checks evaluate the joint coherence of two or more variables, for example verifying mutually exclusive categories that should admit only one valid option, or detecting atypical combinations that deviate from expected multivariate patterns.

The underlying rationale of the proposed system is to examine each certificate through multiple complementary perspectives, enabling the detection of a wide range of data quality issues. Simple rule-based checks are designed to capture explicit inconsistencies—such as violated logical constraints or values falling outside admissible domains—that typically signal coding errors or clearly implausible entries. In parallel, both univariable and multivariable statistical diagnostics are used to identify more subtle anomalies that cannot be revealed through logical rules alone. These include values that are individually plausible but highly atypical in the context of the empirical distribution, as well as combinations of fields that deviate from expected multivariate patterns. By integrating these different layers of validation, the

framework is capable of identifying issues that span from straightforward incongruities to complex, hard-to-detect anomalies arising from the joint behavior of multiple fields.

It is important to emphasize that the resulting penalty score does not represent a deterministic measure of data quality. Because the framework combines rule-based validations with statistical diagnostics, the score should be interpreted as an indicator of the *likelihood* that a certificate may contain inaccuracies, rather than a definitive classification of records as valid or invalid. Deterministic checks contribute directly to the score by flagging explicit inconsistencies, whereas statistical checks reflect deviations from empirical patterns that may arise from genuine heterogeneity as well as from potential errors. As a consequence, higher scores identify certificates that warrant closer examination, but they do not imply that the underlying data are necessarily incorrect. The penalty score therefore functions as a screening tool that prioritizes potentially problematic records and supports more informed decisions on data cleaning, verification, or subsequent exclusion.

2. EPC Data Cleaning and Processing

The proposed set of R scripts was developed to implement a complete and reproducible data-cleaning workflow for the EPC dataset. Its purpose is to prepare a consistent and well-structured database suitable for subsequent statistical analysis and modeling.

The code is organized into a sequence of clearly defined steps that include data import, transformation, filtering, derivation of key indicators, and final export.

Data import

The workflow begins by loading the required R packages for data manipulation and file handling, including `fst`, `data.table`, `dplyr`, and `stringr`. The working directory is automatically set to the location of the script, ensuring that relative file paths remain valid across different execution environments. The script then checks whether a pre-processed version of the dataset in `.fst` format is available. If so, it loads this binary file, which enables very fast data access. Otherwise, it imports the original raw data from the `.csv` file, saves it as `.fst`, and proceeds with the analysis. At this stage, the dataset contains 1,780,577 observations.

Data transformation

The next phase involves a series of variable transformations designed to harmonize formats and correct inconsistencies in the raw data. Boolean variables originally stored as text values (“true”/“false”) are recoded into binary factors labeled “Yes” and “No.” Dates of EPC submission are converted into proper date objects, and the corresponding year of submission is extracted. The variable describing the year of construction is also carefully cleaned: malformed entries are corrected through a predefined mapping, and all buildings are subsequently grouped into seven historical construction periods ranging from “Before 1930” to “After 2006.”

Data filtering

After the variables have been harmonized and pre-processed, the dataset is progressively filtered to retain only relevant EPC records.

The first filter selects residential buildings classified as “E.1(1)” under national regulations, excluding other types of structures (variable `classificazione_dpr`). At this stage, the dataset contains 1,506,724 observations.

A second filter removes dwellings with a heated surface smaller than 25 square meters, as such units are considered atypical or unreliable for energy analysis. After this step, the dataset includes 1,477,114 observations.

Finally, in cases where multiple certificates exist for the same dwelling, only the most recent one is retained to ensure that each record corresponds to a unique and up-to-date observation. The resulting dataset comprises 1,114,184 observations.

Derivation of key indicators

Additional transformations are applied to derive analytical variables. The composite field `y` (classification and non-renewable global energy performance index $EP_{gl,nren}$ of comparable newly constructed buildings) is parsed to extract both the official energy class (`y_classe_energetica`, with categories A4–G) and the corresponding non-renewable global performance index (`y_ep_gl_nren`).

Binary indicators are created to capture the presence of up to seven recommended energy-efficiency interventions, and a new variable classifies each EPC according to whether it refers to an individual unit, an entire building, or a group of units (`oggetto_attestato`). Other textual variables, such as the type of heating system or the climatic zone, are recoded into categorical form to facilitate subsequent analyses.

Derived indicators of energy performance and building geometry are also computed. In particular, the ratio between dispersing surface area and gross heated volume is discretized into six ordered categories ranging from “Critical low” to “Critical high”, providing a compact measure of the building’s thermal compactness (`rapporto_sv_cat`). Similarly, the thermal transmittance index is categorized according to the climatic zone, assigning qualitative levels such as “Excellent,” “Good,” or “Poor” based on predefined numerical thresholds (`y_ie_cat`). The variable describing the energy carrier of exported energy is also harmonized, consolidating all variants referring to electricity under a single label.

Data export

Once all cleaning, filtering, and transformation steps have been completed, the final dataset is saved in three formats:

- .fst for internal high-speed processing,
- .csv for interoperability and external sharing, and

.dta for compatibility with Stata.

The script concludes by reporting the total number of cleaned observations, confirming the successful execution of the entire workflow.

Overall, this pipeline ensures that the EPC data are internally consistent, free of redundant or unreliable entries, and organized in a form that supports robust statistical and econometric analysis. The structure of the code allows for transparent replication and easy adaptation to future datasets collected under similar schemes.

3. Univariate diagnostics

Number of building units (w1_S1)

The plausibility of the reported number of building units is assessed through three elementary checks combined into a single warning score.

First, missing values are penalized as their absence indicates incomplete reporting of essential building data that should always be provided. These cases receive a penalty of 1. Second, logically invalid values (i.e., null or negative counts) are treated as severe errors and are assigned a penalty of 2. Third, atypical values are detected using Tukey's fences applied to a variance-stabilized scale, specifically $(x + 1)$. Inner and outer fences are defined as $Q_1 \pm 1.5 IQR$ and $Q_3 \pm 3 IQR$ respectively, where Q_1 , Q_3 , and IQR are computed on the transformed series. Observations beyond the inner fences are marked as moderate outliers (penalty 1), while those beyond the outer fences are marked as extreme outliers (penalty 2).

The three components (missingness, invalidity, and outlyingness) are summed to form the composite indicator w1_S1. By construction, higher scores indicate greater concern, with null or negative counts and extreme outliers contributing more heavily than simple missing values. The logarithmic transformation controls right-skew and ensures that the outlier detection is not dominated by a few very large buildings while preserving the relative ordering of typical cases.

Energy consumption (w2_S1)

This diagnostic applies to all energy consumption variables (those with the cons_ prefix). Each variable's score combines three checks: missing values (penalty 1), negative readings (penalty 2), and outliers identified using Tukey's fences on (x) . Moderate and extreme outliers are assigned penalties of 1 and 2. Scores are summed across all variables to obtain the composite indicator w2_S1. Zero values are treated as valid, reflecting legitimate measurement scenarios, and are excluded from the logarithmic transformation. The transformation mitigates right skew and prevents extreme consumptions from dominating the analysis.

Surfaces and volumes (w3_S1)

This diagnostic covers all variables describing building surfaces and volumes, including derived ratios such as the surface-to-volume ratio. Missing and negative values receive penalties of 1 and 2, respectively, while zeros are accepted as valid. Outliers are detected via Tukey's fences on (x) , with penalties of 1 and 2 for moderate and extreme cases. The resulting per-variable scores are aggregated to form $w3_S1$. Higher values indicate poorer data reliability, particularly when negative or extreme values are present.

Energy performance indices and exported energy ($w4_S1$)

This diagnostic includes all variables representing non-renewable global performance indices achievable after recommended interventions, renewable energy performance indices by service, and non-renewable energy performance indices by service. Each variable is checked for negative values (penalty 2) and outliers on (x) , with moderate and extreme cases penalized by 1 and 2. The total indicator $w4_S1$ is obtained by summing all per-variable penalties. The logarithmic transformation stabilizes variance and prevents very large values from distorting the results.

Seasonal average system efficiency ($w5_S1$)

All efficiency variables (those with the `eff_` prefix) are tested for negative values (penalty 2) and outliers on (x) , which are penalized by 1 or 2 according to their severity. Zero values are considered valid and excluded from the transformation. Aggregating per-variable penalties yields $w5_S1$, where higher scores indicate the presence of implausible or extreme efficiency values.

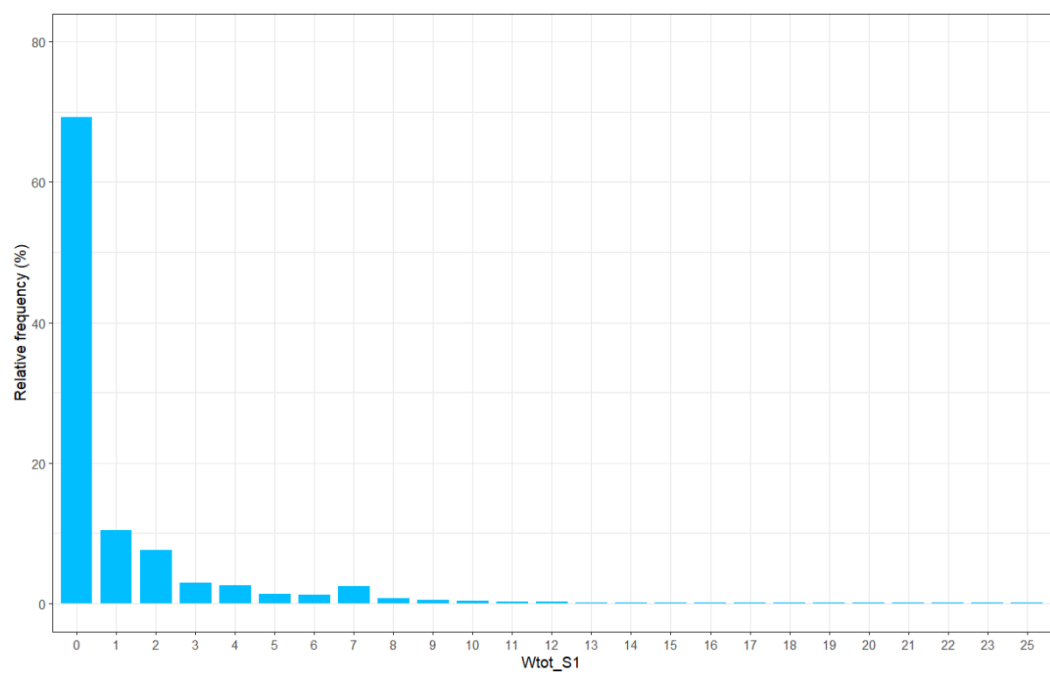
Installation year vs. construction year ($w6_S1$)

This diagnostic checks the temporal consistency between each system's installation year (heating, cooling, domestic hot water, etc.) and the building's construction year. For each system, a lower admissible bound is derived from the construction-year class, and installations preceding it are flagged (penalty 1). Missing values are not penalized. Summing these penalties across all installation-year variables yields $w6_S1$, with higher scores reflecting more temporal inconsistencies.

Construction year ($w7_S1$)

The plausibility of the raw construction-year variable (`anno_costruzione`) is verified independently of its cleaned categorical counterpart (`AnnoCostruzione`). Values earlier than 1400 or later than 2025 are considered implausible and receive a penalty of 1, while all other numeric entries are left unpenalized. The resulting indicator, $w7_S1$, captures unrealistic or temporally inconsistent construction years within the dataset.

Figure 1. Distribution of the total number of univariate inconsistencies (Wtot_S1) obtained by summing the seven diagnostic indicators (W1_S1–W7_S1). Bars represent the percentage of Energy Performance Certificates exhibiting zero, one, or multiple anomalies across the univariate checks.



4. Bivariate diagnostics

Destination: residential vs non-residential (W1_S2)

A logical consistency check was applied to verify the mutual exclusivity of the two destination variables *residenziale* (residential use) and *non_residenziale* (non-residential use). According to the study design, exactly one of the two variables must be marked as “Yes”, while the other must be “No”. This rule is equivalent to an exclusive OR (XOR) condition between the two variables.

A control variable (*W1_S2*) was then defined as follows: $W1_S2=0$ if exactly one of the two indicators was equal to 1 (valid condition), and $W1_S2=1$ if both were equal to 0 or both were equal to 1 (inconsistent condition).

Certificate scope: whole building vs single unit vs group of units (W2_S2)

A logical consistency check was applied to verify the mutual exclusivity among the three certificate scope variables *gruppo_unita_immobiliari* (group of units), *intero_edificio* (whole building), and *unita_immobiliare* (single unit). According to the study design, exactly one of these variables must be marked as “Yes”.

A control variable (*W2_S2*) was defined as follows: $W2_S2=0$ if exactly one of the three indicators was equal to 1 (valid condition), and $W2_S2=1$ if none or more than one indicator was equal to 1 (inconsistent condition). This rule is equivalent to a three-variable XOR (exclusive OR) condition.

Number of units vs whole building (W3_S2)

A logical consistency check was applied to verify the relationship between *numero_unita_immobiliari* (number of units) and *intero_edificio* (whole building). According to the study design, when *intero_edificio* is marked as “Yes”, the number of units must be equal to 1.

A control variable (*W3_S2*) was defined as follows: $W3_S2=1$ when *intero_edificio*=“Yes” and *numero_unita_immobiliari* > 1, and $W3_S2 = 0$ otherwise.

Reason for certificate (W4_S2)

A logical consistency check was applied to verify the mutual exclusivity among the six variables describing the reason for certificate: *nuova_costruzione* (new construction), *passaggio_proprieta* (change of ownership), *locazione* (rental), *ristrutturazione_importante* (major renovation), *oggetto_altro* (other), and *riqualificazione_energetica* (energy improvement). According to the study design, exactly one of these variables must be marked as “Yes”. This rule is equivalent to a six-variable exclusive OR (XOR) condition.

A control variable (*W4_S2*) was defined as follows: $W4_S2=0$ if exactly one of the six indicators was equal to 1 (valid condition), and $W4_S2=1$ if none or more than one indicator was equal to 1 (inconsistent condition).

Global energy class vs envelope and performance indicators (W5_S2)

A logical consistency check was applied to verify the coherence between the global energy class (*classe_energetica*) and the envelope (*pe*) and performance (*pi*) indicators. High energy classes (A4 or A3) are expected to correspond to good envelope and performance indicators, while low energy classes (F or G) should not be associated with high performance levels. A control variable (*W5_S2*) was defined to flag violations of the following logical conditions:

- *classe_energetica* ∈ {A4, A3} AND *pi* = “Red”;
- *classe_energetica* ∈ {F, G} AND *pi* = “Green”;
- *classe_energetica* ∈ {A4, A3} AND *pe* = “Red”.

Each violation increased *W5_S2* by 1.

Energy class vs new construction (W6_S2)

A logical consistency check was applied to verify the coherence between the global energy class (*classe_energetica*) and the *nuova_costruzione* (new construction) indicator reported in the EPC. Newly constructed buildings are expected to achieve high energy efficiency ratings and should not fall into low energy classes (E, F, or G).

A control variable (*W6_S2*) was defined based on the following logical condition:

nuova_costruzione = “Yes” AND *classe_energetica* ∈ {E, F, G}.

Records satisfying this condition were flagged (*W6_S2*=1), whereas all others were considered consistent.

Energy carrier versus consumption consistency (W7_S2)

This diagnostic verifies the internal consistency between the declared energy carriers and the corresponding energy consumption values reported in the certificate.

Each energy system (space heating, cooling, domestic hot water, etc.) is associated with one or more energy carrier variables, for example, natural gas, fuel oil, electricity, biomass, or district heating. The variables considered include all those describing the carrier used by each subsystem (such as *ci_vettore_energetico_1*, *ce_vettore_energetico_2*, *pa_vettore_energetico*, and others).

To ensure comparability across systems, carrier labels were first harmonized (e.g., *gasolio e olio combustibile* → *gasolio*, *energia elettrica* → *solare_fotovoltaico* for renewable production). Each declared carrier was then matched to its corresponding consumption variable (e.g., *gas_naturale* → *consumi_gas_naturale*, *energia elettrica* → *consumi_energia_elettrica*).

For each record, the algorithm checks whether the consumption values associated with the declared carriers are logically consistent, that is, strictly positive and non-missing. Whenever one or more of the matched consumption variables are zero, negative, or missing, a warning is raised (*warn_ve_cons*), proportional to the number of inconsistencies detected.

This procedure allows the identification of incoherent cases, such as systems declaring the use of a given energy source but reporting no corresponding consumption, or implausible negative readings, thus ensuring the internal validity of the energy-balance information provided in the dataset.

Consumption versus energy carrier consistency (W8_S2)

This diagnostic mirrors the previous check (W7_S2) but applies the verification in the opposite direction. Instead of starting from the declared energy carriers, the algorithm begins with the reported energy consumption values and checks whether the corresponding carriers are consistent with them.

For each record, all consumption variables with strictly positive values are identified and mapped to their expected carrier labels (e.g., *consumi_gas_naturale* → *gas_naturale*, *consumi_gpl* → *gpl*). The set of declared carriers is then retrieved from the fields describing each subsystem (such as *ci_vettore_energetico_1*, *ce_vettore_energetico_2*, *pa_vettore_energetico*, etc.), after harmonizing the carrier nomenclature (e.g., *gasolio e olio combustibile* → *gasolio*, *biomasse_liquide_e_gassose* → *biomasse_liquide* + *biomasse_gassose*).

The diagnostic then compares the two sets. Whenever a positive consumption is reported for an energy source that does not appear among the declared carriers for that building, a warning is issued. The indicator counts the number of such mismatches per record, and its aggregated version (W8_S2) provides a measure of the overall inconsistency between the reported consumptions and the declared energy vectors.

This control complements W7_S2 by ensuring bidirectional coherence between the declared energy systems and their corresponding consumption data, thereby strengthening the internal reliability of the energy certificate database.

Consistency between Energy Class and EP_{gl,nren} (W9_S2)

Verifying the consistency between the energy class and the non-renewable global energy performance index (EP_{gl,nren}) is crucial for EPCs. The energy class represents the final and most visible outcome of the certification process. It is the element used by citizens, technicians, and the real estate market to evaluate a building's energy performance.

The energy class is directly derived from the EP_{gl,nren} value through predefined regulatory thresholds. Therefore, any inconsistency between the two compromises the reliability of the entire energy certification system. Such discrepancies may result from data-entry errors, computational mistakes, or intentional data manipulation.

In Lombardy, these aspects are governed by the Regional Decree DDUO 6480/2015, which regulates building energy efficiency and the EPCs. According to this decree, the energy class must be determined based on the ratio:

$$r_{EP} = \frac{EP_{gl,nren}}{EP_{gl,nren,rif}}$$

where $EP_{gl,nren,rif}$ is the non-renewable global energy performance index of a reference building. This reference building is a “virtual” construction with the same geometry, orientation, and location as the actual building, but with standardized technical characteristics defined by regulation (e.g., thermal transmittance, system efficiency, etc.).

Both $EP_{gl,nren,rif}$ and the energy class of the reference building are retrieved from column y in the dataset, described as “*Classification and non-renewable global energy performance index ($EP_{gl,nren}$) of comparable new buildings*”.

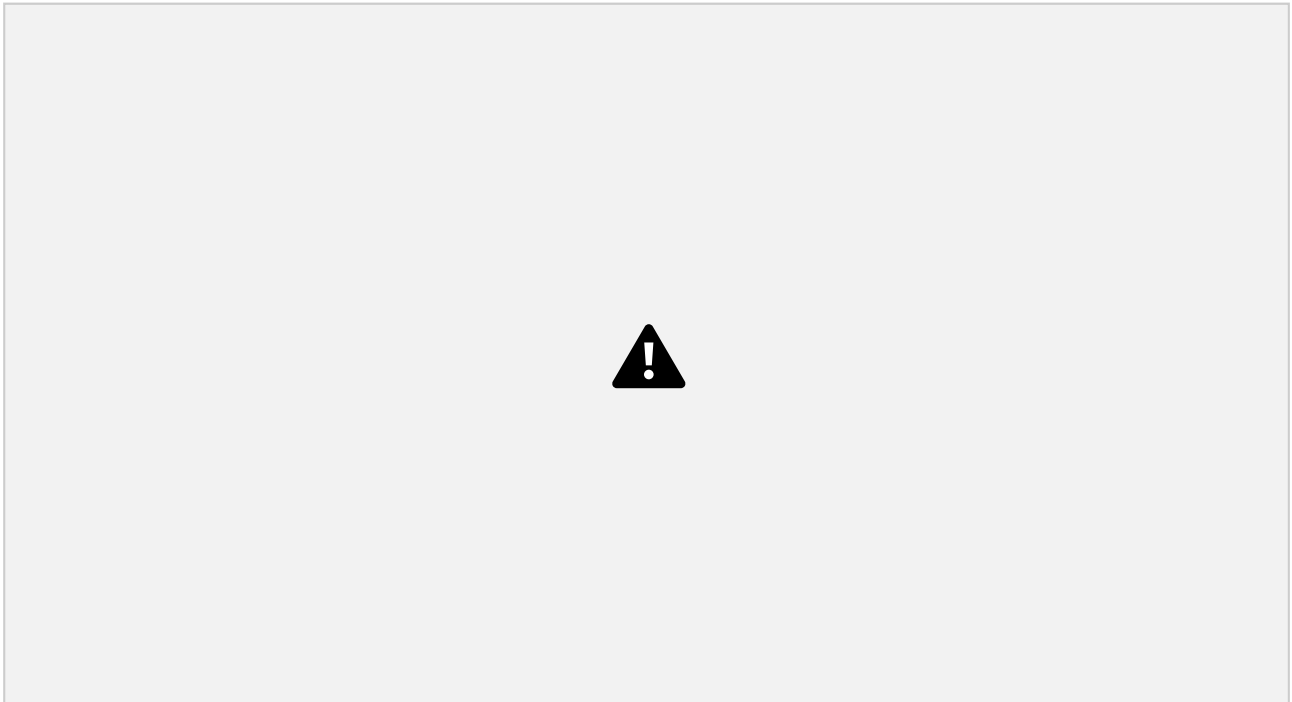
For each energy class, the threshold values of $EP_{gl,nren}$ (defined as the product of a class-specific constant and $EP_{gl,nren,rif}$) are reported in Table 5 of the DDUO 6480/2015 decree. These class-specific constants therefore represent the threshold values for the r_{EP} ratio.

Table 1. Classification scale of buildings based on the non-renewable global energy performance index, reproduced from Table 5 (p. 24) of Decree DDUO 6480/2015.

	Classe A4	$\leq 0,40 EP_{gl,nren,rif}$
$0,40 EP_{gl,nren,rif} <$	Classe A3	$\leq 0,60 EP_{gl,nren,rif}$
$0,60 EP_{gl,nren,rif} <$	Classe A2	$\leq 0,80 EP_{gl,nren,rif}$
$0,80 EP_{gl,nren,rif} <$	Classe A1	$\leq 1,00 EP_{gl,nren,rif}$
$1,00 EP_{gl,nren,rif} <$	Classe B	$\leq 1,20 EP_{gl,nren,rif}$
$1,20 EP_{gl,nren,rif} <$	Classe C	$\leq 1,50 EP_{gl,nren,rif}$
$1,50 EP_{gl,nren,rif} <$	Classe D	$\leq 2,00 EP_{gl,nren,rif}$
$2,00 EP_{gl,nren,rif} <$	Classe E	$\leq 2,60 EP_{gl,nren,rif}$
$2,60 EP_{gl,nren,rif} <$	Classe F	$\leq 3,50 EP_{gl,nren,rif}$
	Classe G	$> 3,50 EP_{gl,nren,rif}$

Figure 2 shows the distribution of the ratio r_{EP} across the different declared energy classes. As can be observed, for classes C, D, E, F, and G, the thresholds defined in Table 5 are generally well respected, with only minor tails extending beyond the regulatory ranges. In contrast, for classes B, A1, A2, A3, and A4, the distribution appears less concentrated, and a considerable number of EPCs display values outside the prescribed limits.

Figure 2. Density distribution of the ratio r_{EP} across the declared energy classes. The x-axis is shown on a logarithmic scale. Classes C–G display distributions largely consistent with the regulatory thresholds defined in DDUO 6480/2015, whereas classes B and A1–A4 exhibit wider spreads with several EPCs outside the expected ranges.



Consistency between periodic thermal transmittance and year of construction (W10_S2)

The periodic thermal transmittance (Y_{IE}) quantifies the ability of a building envelope to attenuate and delay heat transfer under periodic thermal loads, such as daily temperature fluctuations. It is a key parameter for assessing the dynamic thermal behavior of opaque components and their contribution to summer comfort and overall energy performance. Within EPCs, an accurate estimation of this index is essential, as unrealistic or inconsistent values can lead to significant errors in the evaluation of cooling demand and in the assignment of the energy class. For this reason, a systematic quality control of the periodic thermal transmittance data is necessary to ensure the physical consistency and reliability of EPC records.

A series of consistency checks was applied to the periodic thermal transmittance (Y_{IE}) to identify physically implausible or noncompliant values:

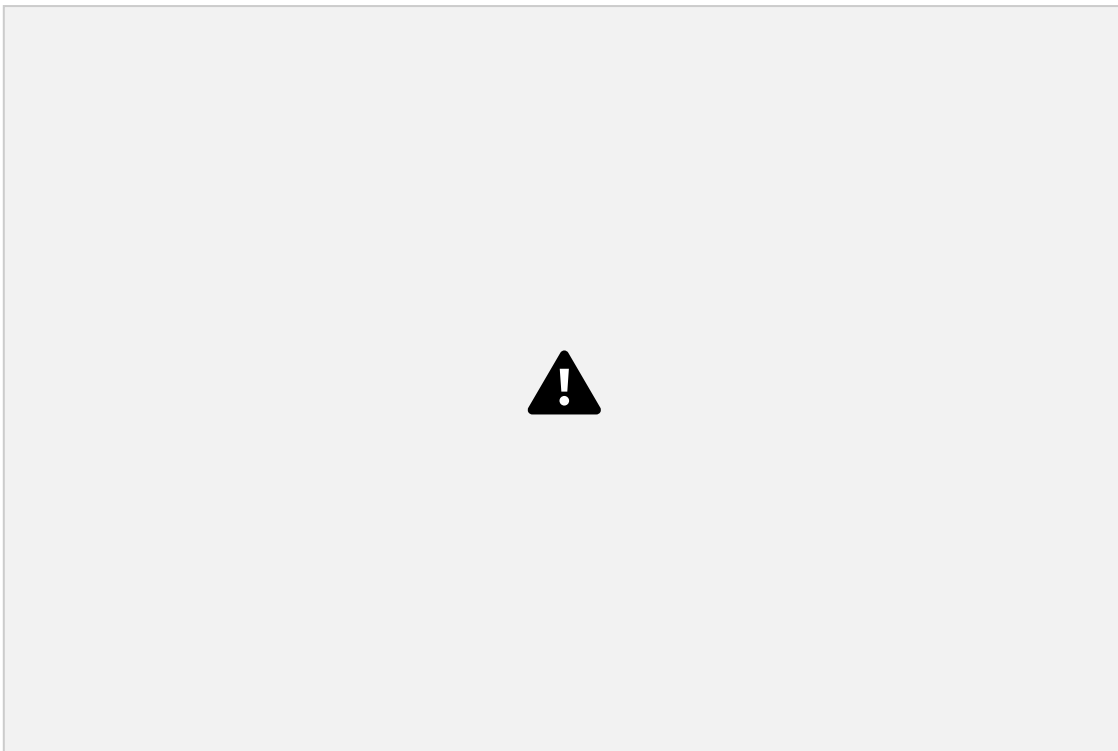
- records with Y_{IE} exceeding 1 were automatically flagged as inconsistent;
- for buildings constructed after 2006 or 2016, Y_{IE} values above 0.30 and 0.18 respectively were considered incompatible with regulatory standards;
- EPCs classified as new constructions or belonging to high-efficiency energy classes (A2–A4) were required to exhibit $Y_{IE} \leq 0.20$.

These criteria collectively ensure that the reported thermal transmittance values are coherent with the building's age, declared energy performance, and expected envelope quality.

Consistency of EP_{gl,nren} values for new constructions (W11_S2)

A consistency check was applied to the non-renewable global energy performance index for buildings declared as new constructions. According to current regulatory standards, newly built or nearly zero-energy buildings are expected to exhibit very low non-renewable energy consumption values. Therefore, EPC records indicating EP_{gl,nren} greater than 100 kWh/m²-year were flagged as inconsistent, as such values are unlikely to correspond to compliant new buildings. This check helps detect potential data entry or calculation errors and ensures the reliability of EPC information for high-efficiency constructions.

Figure 3. Distribution of the total number of cross-variable inconsistencies (W_{tot_S2}) in the dataset. Bars represent the percentage of Energy Performance Certificates showing zero, one, or multiple inconsistencies across the eleven logical checks.



5. Regression-based diagnostic

The third stage of the quality assessment focuses on regression-based diagnostics aimed at identifying records that substantially deviate from expected physical or statistical relationships. In this framework, multivariate regression models are estimated on large, homogeneous subsets of EPCs sharing the same configuration of energy vectors. These subsets ensure that model coefficients represent comparable technological and system contexts.

For each record, residuals and model-derived metrics are used to detect anomalous or implausible patterns. Observations showing significant deviations from the regression-predicted behaviour are classified as outliers and penalized in the overall EPC quality score.

In the regression-based diagnostics stage, the chosen models act as implicit representations of the “normal” multivariate relationships among key EPC variables. In a sense, the regression serves to learn the typical conditional structure of the data: once a model is fit on a majority of certificates—within reasonably homogeneous subgroups delineated by energy-vector configuration—outlier detection is conducted by examining deviations (i.e. residuals or prediction errors) that diverge from the learned norm.

This approach is analogous to unsupervised methods such as isolation forests or autoencoders, which learn a model of the “normal” data manifold and flag observations that lie far from it. For instance, Isolation Forest isolates instances by repeated random partitioning: anomalies are those that require fewer splits to isolate, because they tend to differ more from the bulk of the data. Autoencoders, similarly, learn a compressed representation of typical data; points that reconstruct poorly (i.e. yield large reconstruction error) are considered anomalous.

Here, in our hybrid regression-based scheme, the model residuals play the role of a divergence score from the learned norm. Records whose residuals exceed empirical thresholds can therefore be classified as outliers and penalized in the EPC quality scoring. Because our models are estimated within energy-vector subgroups, this method respects heterogeneity in system configurations and avoids forcing one global relationship across incompatible types.

To ensure that regression models are estimated on homogeneous populations, records were partitioned into subgroups defined by the configuration of energy vectors reported in the EPC. Specifically, six categorical variables were used to characterize the vector mix across systems (*ci_vettore_energetico_1*, *ci_vettore_energetico_2*, *ce_vettore_energetico_1*, *pa_vettore_energetico*, *pfr_vettore_energetico_1*, *pfr_vettore_energetico_2*). For each record, these variables were concatenated into a composite label uniquely identifying the energy-vector configuration; records sharing the same configuration were assigned the same subgroup identifier (*group_label*).

Subgroup sizes were then tabulated, and only “large” subgroups (more than 7,500 records) were retained for regression fitting to guarantee stable parameter estimates and outlier detection thresholds. The resulting table reports, for each retained subgroup, the corresponding energy-vector composition.

Table 2. Energy-vector subgroups of Energy Performance Certificates defined by the combination of energy vectors used for space heating (*ci_vettore_energetico_1*, *ci_vettore_energetico_2*), cooling (*ce_vettore_energetico_1*), domestic hot water (*pa_vettore_energetico*), and renewable energy generation (*pfr_vettore_energetico_1*, *pfr_vettore_energetico_2*). The number N of certificates in each subgroup is also reported. Only subgroups with more than 7,500 certificates are shown.

Group label	N	ci_vettore_energetico_1	ci_vettore_energetico_2	ce_vettore_energetico_1	pa_vettore_energetico	pfr_vettore_energetico_1	pfr_vettore_energetico_2
GAS_N_N_GAS_N_N	521865	Natural gas			Natural gas		
GAS_N_ELE_GAS_N_N	203447	Natural gas		Electricity	Natural gas		
GAS_N_N_ELE_N_N	48908	Natural gas			Electricity		
GAS_N_N_N_N_N	28623	Natural gas					
N_N_N_N_N_N	18871						
ELE_N_ELE_ELE_ELE_ELE	17839	Electricity		Electricity	Electricity	Electricity	Electricity
GAS_N_ELE_GAS_ELE_N	16028	Natural gas		Electricity	Natural gas	Electricity	
GAS_N_ELE_ELE_N_N	15310	Natural gas		Electricity	Electricity		
GAS_ELE_ELE_GAS_ELE_ELE	13850	Natural gas	Electricity	Electricity	Natural gas	Electricity	Electricity
GAS_ELE_N_GAS_ELE_ELE	12358	Natural gas	Electricity		Natural gas	Electricity	Electricity
GAS_GAS_N_GAS_N_N	11924	Natural gas	Natural gas		Natural gas		
TEL_N_N_TEL_N_N	11598	District heating			District heating		
TEL_N_N_GAS_N_N	10292	District heating			Natural gas		
GAS_N_N_GAS_SOL_N	9591	Natural gas			Natural gas	Solar thermal	
ELE_N_N_ELE_ELE_ELE	8657	Electricity			Electricity	Electricity	Electricity
TEL_N_N_ELE_N_N	8408	District heating			Electricity		
TEL_N_ELE_TEL_N_N	7780	District heating		Electricity	District heating		
ELE_N_ELE_ELE_ELE_N	7223	Energia elettrica		Energia elettrica	Energia elettrica	Energia elettrica	

GAS_N_N_GAS_N_N + GAS_N_N_N_N_N + GAS_GAS_N_GAS_N_N subgroups (w1_S3)

A regression-based diagnostic was developed to identify records whose reported natural gas consumption deviates from the expected relationship with building characteristics and energy performance. The analysis focused on the union of three subgroups representing buildings whose energy systems rely almost exclusively on natural gas: (1) buildings using gas for both space heating and domestic hot water, with no other energy vector; (2) buildings using gas solely for space heating; and (3) buildings with multiple gas-fueled systems, such as primary and secondary heating circuits and domestic hot water.

The following linear regression model was fitted:

$$(cons_GasNatur) = \beta_0 + \beta_1 \cdot (1 + SupUtileRisc) + \beta_2 \cdot (1 + ep_gl_nren)$$

where *cons_GasNatur* denotes the consumption of natural gas, *SupUtileRisc* the heated surface area, and *ep_gl_nren* as the non-renewable global performance index. Records with missing values in any of these variables or with zero natural gas consumption were excluded from model fitting.

For each record included in the regression, the residual was computed and expressed as a relative percentage error with respect to the log-transformed consumption. This error served as a measure of deviation from the model's predicted "normal" behavior. The function `step4_warn_resid_errpct()` then discretized the residual percentage error into a three-level penalty according to two thresholds (2.5% and 5%), corresponding to increasing levels of inconsistency (0 = none; 1 = moderate; 2 = severe). The resulting penalties were merged into the main dataset to form the regression-based indicator *w1_S3*. Records not

included in the modeled subgroups, or excluded due to missing or zero values, received a neutral score ($w1_S3=0$).

The fitted regression model yielded an R^2 of 0.996, indicating an excellent fit. Overall, 2,147 of 562,380 certificates (3.81%) were assigned a penalty of $w1_S3=1$, and 1,347 certificates (2.48%) a penalty of $w1_S3=2$.

ELE_N_ELE_ELE_ELE_ELE, ELE_N_ELE_ELE_ELE_N, ELE_N_N_ELE_ELE_ELE subgroups ($w2_S3$)

A regression-based diagnostic was developed to identify records whose reported electricity consumption deviates from the expected relationship with building characteristics and energy performance. The analysis focused on the union of three homogeneous subgroups representing all-electric buildings, in which electricity is used as the sole energy vector for space heating, cooling, domestic hot water production, and renewable-based systems. A linear regression model was fitted on the logarithmic scale, with (*cons_EnergElettr*) as the dependent variable and ($1 + SupUtileRisc$) (heated surface area) and ($1 + ep_gl_nren$) (non-renewable global performance index) as predictors. Records with zero electricity consumption or missing values were excluded from model fitting. Residuals were expressed as relative percentage errors with respect to the observed values, and penalties were assigned according to two predefined thresholds: a value of 1 for moderate deviations ($2.5\% < |error| < 5\%$) and 2 for large deviations ($|error| > 5\%$).

The fitted regression model yielded an R^2 of 0.977, indicating an excellent fit. Overall, 240 of 33,621 certificates (0.71%) were assigned a penalty of $w1_S3=1$, and 531 certificates (1.58%) a penalty of $w1_S3=2$.

GAS_N_ELE_GAS_N_N subgroup ($w3_S3$)

A regression-based diagnostic was applied to the subgroup characterized by natural gas used for space heating and domestic hot water production, and electricity as the energy vector for cooling. The model was fitted on the log-transformed scale, with natural gas consumption as the dependent variable and the following predictors: heated surface area (*SupUtileRisc*), non-renewable primary energy performance index for space heating (*ci_epnren*), and for domestic hot water production (*pa_epnren*).

$$(cons_GasNatur) = \beta_0 + \beta_1 \cdot (1 + SupUtileRisc) + \beta_2 \cdot (1 + ci_epnren) + \beta_3 \cdot (1 + pa_epnren)$$

Records showing residual-based deviations exceeding predefined thresholds were penalized proportionally to the magnitude of the discrepancy.

The fitted regression model yielded an R^2 of 0.976, indicating an excellent fit. Overall, 2382 of 203,442 certificates (1.17%) were assigned a penalty of $w1_S3=1$, and 1805 certificates (0.89%) a penalty of $w1_S3=2$.

GAS_N_N_ELE_N_N subgroup ($w4_S3$)

A regression-based diagnostic was applied to the subgroup characterized by natural gas as the energy vector for space heating and electricity for domestic hot water, with no other vectors present. The model was fitted on the logarithmic scale using a +1 offset to accommodate zeros, with natural gas consumption as the dependent variable and two predictors: heated surface area (*SupUtileRisc*) and the non-renewable primary energy performance index for space heating (*ci_epnren*). Records with zero natural gas consumption or missing values in the modeled variables were excluded prior to fitting.

$$(cons_GasNatur) = \beta_0 + \beta_1 \cdot (1 + SupUtileRisc) + \beta_2 \cdot (1 + ci_epnren)$$

Residuals were converted into relative percentage errors with respect to the log-transformed outcome and mapped to a discrete warning score via two predefined thresholds: a penalty of 1 for moderate deviations ($2.5\% < |error| \leq 5\%$) and 2 for large deviations ($|error| > 5\%$). The resulting penalties were merged back into the full dataset to form the indicator *w4_S3*; non-modeled records received *w4_S3*=0.

The fitted regression model yielded an R^2 of 0.99, indicating an excellent fit. Overall, 250 of 48,849 certificates (0.51%) were assigned a penalty of *w1_S3*=1, and 380 certificates (0.78%) a penalty of *w1_S3*=2.

GAS_N_N_GAS_SOL_N subgroup (*w5_S3*)

A regression-based diagnostic was applied to the subgroup characterized by natural gas used as the energy vector for space heating and domestic hot water production, and solar energy as the vector associated with renewable production systems, with no other energy sources present.

The model was fitted on the logarithmic scale using a +1 offset to accommodate possible zero values. The dependent variable was the natural gas consumption (*cons_GasNatur*), while the predictors were the heated surface area (*SupUtileRisc*) and the non-renewable global performance index (*ep_gl_nren*).

$$(cons_GasNatur) = \beta_0 + \beta_1 \cdot (1 + SupUtileRisc) + \beta_2 \cdot (1 + ep_gl_nren)$$

Residuals were converted into relative percentage errors with respect to the log-transformed outcome and mapped to a discrete warning score via two predefined thresholds: a penalty of 1 for moderate deviations ($2.5\% < |error| \leq 5\%$) and 2 for large deviations ($|error| > 5\%$). The resulting penalties were merged back into the full dataset to form the indicator *w5_S3*; non-modeled records received *w5_S3* =0.

The fitted regression model yielded an R^2 of 0.982, indicating an excellent fit. Overall, 173 of 9,590 certificates (1.80%) were assigned a penalty of *w1_S3*=1, and 148 certificates (1.54%) a penalty of *w1_S3*=2.

Regression-based diagnostics were applied to the largest and most homogeneous energy-vector subgroups, covering 858,077 certificates (77% of the entire dataset). In total, five regression models were estimated across nine subgroups defined by the combination of energy vectors used for space heating, cooling, domestic hot water production, and renewable systems. This represents a substantial portion of the dataset, ensuring broad applicability of the regression-based consistency checks.

Overall, 5,192 certificates (0.6%) were assigned a penalty of 1, indicating moderate deviations from model-predicted values, and 4,256 certificates (0.5%) received a penalty of 2, corresponding to large deviations.

6. Diagnostics Using Multivariable Machine Learning Methods

The fourth stage of the EPC quality assessment introduces a complementary diagnostic approach based on machine learning. As in the regression-based controls described in the previous stage, the objective is to learn/approximate the relationship between a dependent variable (in this case, the non-renewable global energy performance index, $EP_{gl,nren}$) and a comprehensive set of building, system, and climatic variables.

The learning algorithm adopted here is XGBoost, an ensemble method designed to capture complex, non-linear interactions among predictors (Chen and Guestrin, 2016). XGBoost is widely recognized as one of the most effective algorithms for structured datasets and frequently achieves state-of-the-art results in competitive machine-learning tasks (Shmuel, Glickman, and Lazebnik, 2025).

When the algorithm succeeds in identifying strong and consistent relationships, i.e., when the model accurately predicts $EP_{gl,nren}$ for a large portion of certificates, the resulting pattern can be interpreted as representing the “normal” behavior of the population. Certificates that substantially deviate from this well-defined normality are flagged as anomalies and assigned a penalty according to the magnitude of their deviation. More specifically, for each certificate, the absolute percentage deviation is computed between the declared value of $\log_{10}(EP_{gl,nren})$ and the corresponding value predicted by the XGBoost model. Records showing prediction errors above predefined thresholds receive progressively higher penalty levels within the warning system.

Two XGBoost predictive algorithms were trained:

- *XGBoost1*: the first using an extended set of predictors, and
- *XGBoost2*: the second using the same set excluding the two most influential variables identified in the first algorithm.

6.1 XGBoost1 ($W1_{S4}$)

The predictive algorithm XGBoost1 was trained using the following set of potentially predictive variables: year of construction (*AnnoCostruzione*), heated floor area (*SupUtileRisc*), gross heated volume (*VolLordoRisc*), dispersing surface (*SupDisperd*), surface-to-volume ratio (*rapporto_sv*), heating energy performance index (*ep_h_nd*), climate zone (*zona_climatica*), certificate scope (*oggetto_attestato*), and CO₂ emissions (*emissioni_co2*), along with the characteristics of the main heating, auxiliary, and renewable systems (including system type, installation year, energy vector, and mean efficiency). This set of explanatory variables was partly defined following the criteria proposed by Buratti et al. (2014).

The predictive performance of the algorithm is summarized by the in-sample and out-of-sample coefficients of determination (R^2), together with the error metrics reported in in Table 3.

Table 3. Error metrics (XGBoost1)

Metric	Train	Test
RMSE	0.0119	0.0175
MAE	0.0049	0.0057
R ²	0.998	0.996

These results indicate that the algorithm successfully captures the behavior of a large portion of EPCs, which can therefore be considered representative of the “normal” population. The absolute percentage error between observed and predicted values is then computed for each record. EPCs exhibiting prediction errors beyond predefined thresholds are classified as anomalous and assigned a penalty level within the warning system:

- errors between 2.5% and 5% receive penalty level 1,
- errors between 5% and 10% receive penalty level 2, and
- errors exceeding 10% receive penalty level 3.

This procedure enables the automatic attribution of a quality score to each certificate, highlighting cases that deviate substantially from the patterns learned from the broader, internally consistent dataset, and thereby improving the overall reliability assessment process.

Figure 4 presents the estimated variable importance for all explanatory variables used in XGBoost1. It is immediately evident that the variables most strongly associated with the response variable EPgl,nren are CO₂ emissions and the heating energy performance index. Consequently, a second algorithm, XGBoost2, was developed by retraining the model after removing these two predictors from the input set.

The distribution of the penalty levels assigned by the algorithm is shown below, providing an overview of how the certificates are classified according to their prediction errors.

The distribution of the penalty levels assigned by the algorithm is shown below, providing an overview of how the certificates are classified according to their prediction errors. The penalties generated by XGBoost1 were stored in the variable w1_S4 (Table 4).

Table 4. Count of penalties generated by XGBoost1

0	1	2	3
1103256 (99.0%)	7732 (6.9‰)	2656 (2.4‰)	540 (0.5‰)

6.2 XGBoost2 (w2_S4)

The XGBoost2 algorithm was trained using the same set of explanatory variables adopted for XGBoost1, except for CO₂ emissions (emissioni_co2) and heating energy performance index (ep_h_nd), which were excluded from the predictor set.

Its predictive performance is summarized in Table 5, and the importance of the remaining variables is shown in Figure 5.

Table 5. Error metrics (XGBoost2)

Metric	Train	Test
RMSE	0.1290	0.1362
MAE	0.0977	0.1026
R ²	0.789	0.764

As shown, the predictive performance of XGBoost2 is good but clearly lower than that of XGBoost1. This indicates that the algorithm captures only a portion of the “normal” population in this case. For this reason, a different threshold system was adopted, designed to detect more pronounced deviations than those identified by XGBoost1, namely:

- errors between 25% and 35% receive penalty level 1,
- errors between 35% and 45% receive penalty level 2, and
- errors exceeding 45% receive penalty level 3.

The penalties assigned by XGBoost2 were saved in the variable w2_S4. Their frequency distribution highlights how the certificates were classified according to the magnitude of their prediction errors (Table 6).

Table 6. Count of penalties generated by XGBoost2

0	1	2	3
1110077 (99.6%)	3521 (3.2‰)	454 (0.4‰)	132 (0.12‰)

The low Spearman correlation observed between w1_S4 and w2_S4 ($\rho = 0.034$, 95% CI 0.032–0.036) suggests that XGBoost1 and XGBoost2 capture distinct types of anomalies, each associated with different patterns among the explanatory variables. This indicates that considering both algorithms was worthwhile, as they provide complementary and non-overlapping information.

The contingency table of w1_S4 and w2_S4 (Table 7) highlights the degree of agreement between the penalty levels assigned to EPCs by XGBoost1 and XGBoost2.

Table 7. Cross-table of w1_S4 and w2_S4

w1_S4 \ w2_S4	0	1	2	3
0	1099416	3346	408	86
1	7596	101	19	16
2	2563	54	22	17
3	502	20	5	13

Figure 4. Variable importance for all explanatory variables used in the XGBoost1 predictive algorithm, computed as mean decrease in accuracy. The figure ranks the predictors according to their relative contribution to the model, highlighting those most strongly associated with the non-renewable global energy performance index $\log_{10}(\text{EP}_{\text{gl,nren}})$.

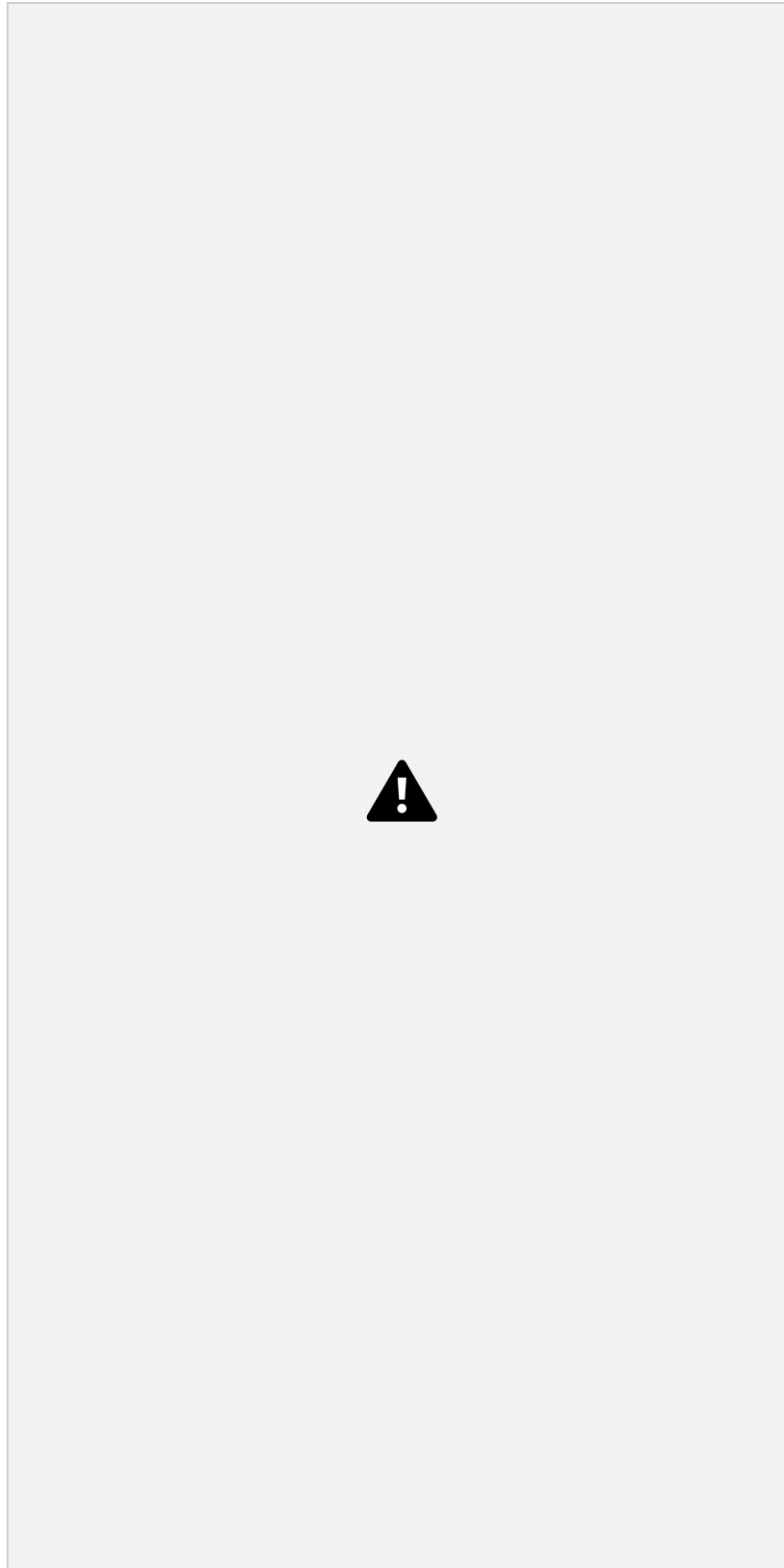
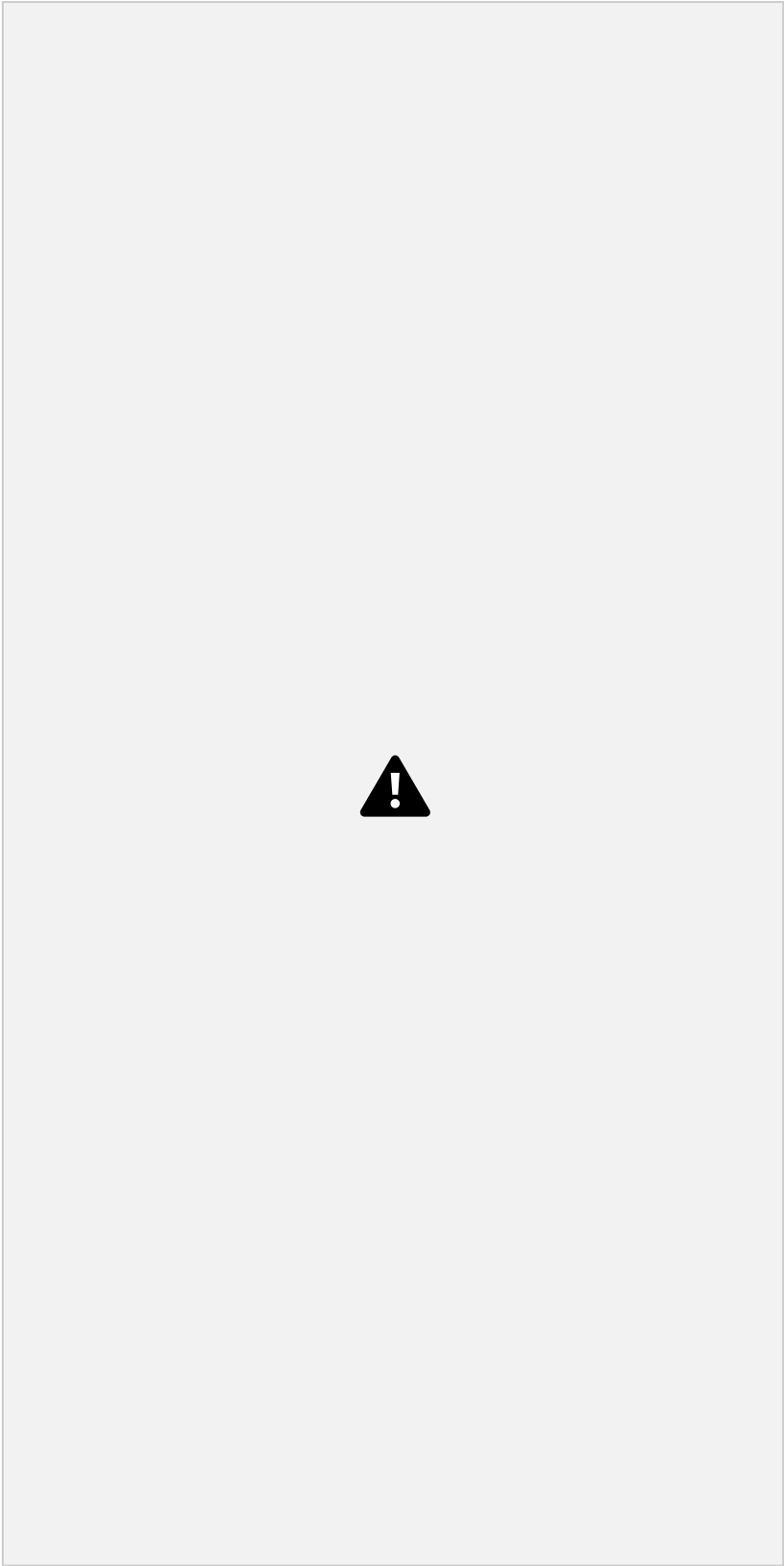


Figure 5. Variable importance for all explanatory variables used in the XGBoost2 predictive algorithm, computed as mean decrease in accuracy. The figure ranks the predictors according to their relative contribution to the model, highlighting those most strongly associated with the non-renewable global energy performance index $\log_{10}(\text{EPgl,nren})$.



6.3 Isolation Forests (W3_S4)

Anomaly detection aims to identify observations that deviate markedly from the general structure of the data, and that may therefore reflect measurement errors, rare events, or atypical behavior. In the context of EPCs, detecting such atypical records is essential to ensure data quality and to avoid biased analyses.

The Isolation Forest algorithm is an unsupervised ensemble algorithm specifically designed for anomaly detection. Its rationale differs from distance- or density-based methods. Isolation Forests operate by recursively partitioning the data space using random splits: observations that are easier to isolate (i.e., requiring fewer splits) are more likely to be anomalies. The method constructs an ensemble of randomized binary trees, and the average path length required to isolate each observation provides a quantitative measure of its degree of abnormality. Data points with shorter average path lengths are assigned higher anomaly scores. This approach is computationally efficient and well-suited for high-dimensional datasets.

In the present analysis, we employed the Extended Isolation Forest (EIF) algorithm, which generalizes the original Isolation Forest by using random hyperplanes rather than axis-aligned splits. This modification reduces the bias induced by feature axis orientation and improves the algorithm's ability to detect anomalies in datasets with correlated or non-axis-aligned structures.

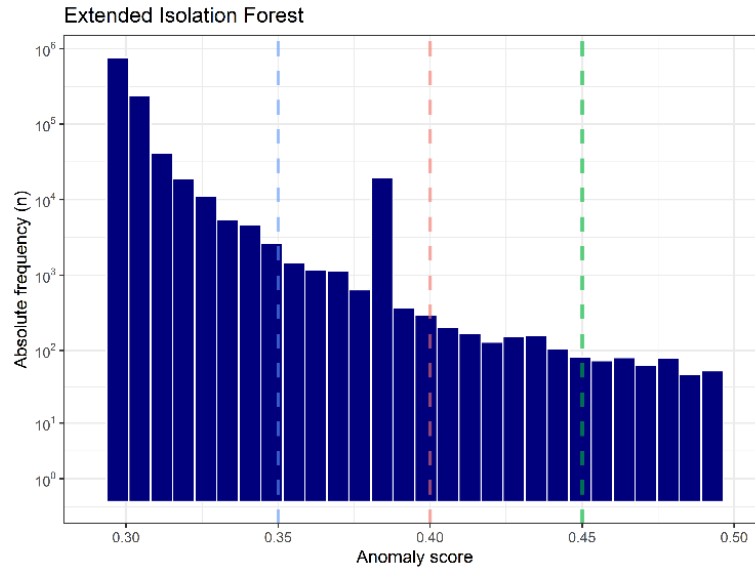
In our analysis, anomaly detection was conducted using the Extended Isolation Forest (EIF) algorithm on a feature space consisting of both continuous and categorical variables. Continuous variables included geometric characteristics, thermal performance indices, and energy carrier consumption measures, while categorical variables (such as climatic zone, certificate type, construction period, and certificate usage) were supplied to the model in their original form and automatically transformed into numerical encodings by the H2O framework for R during the conversion of the dataset using `as.h2o()`. By applying EIF to this mixed-variable representation, we identified records exhibiting atypical multivariate profiles, which were subsequently flagged for targeted review prior to downstream statistical analyses.

The Extended Isolation Forest model was fitted using the H2O implementation with the following settings:

- the number of trees was set to 200 (`ntrees = 200`),
- each tree was trained on a subsample of 65,536 observations (`sample_size = 65536`),
- the extension level was set to the maximum allowable value (`extension_level = p - 1`, where p is the number of input variables), enabling the use of fully general random hyperplanes rather than axis-aligned splits.

These parameter choices balance computational efficiency and stability of the anomaly score distribution while maximizing the algorithm's ability to detect atypical multivariate patterns in mixed-type feature space. EIF does not require complete cases. Observations with missing values are still evaluated: numeric NAs are randomly assigned to one of the split branches during tree construction, and categorical NAs are treated as an additional category. As a consequence, every record receives an anomaly score, even if it contains missing values.

Figure 6. Frequency distribution of the anomaly score computed using the Extended Isolation Forest algorithm for the 1,114,184 EPCs. The y-axis is shown on a logarithmic scale to enhance the visibility of the right tail. The vertical dashed lines indicate the threshold values used to define the penalty levels.



The distribution of the anomaly score generated by the Extended Isolation Forest is shown in Figure 6. The score was discretized into four intervals: < 0.35 , $[0.35, 0.40)$, $[0.40, 0.45)$, and ≥ 0.45 , which were assigned penalty levels 0, 1, 2, and 3, respectively. Higher penalty levels correspond to increasingly atypical observations and therefore lower data reliability. The absolute and relative frequencies of these categories were 1,087,066 (97.6%), 24,853 (2.2%), 1,081 (0.1%), and 1,184 (0.1%).

6.4 Autoencoder (W4_S4)

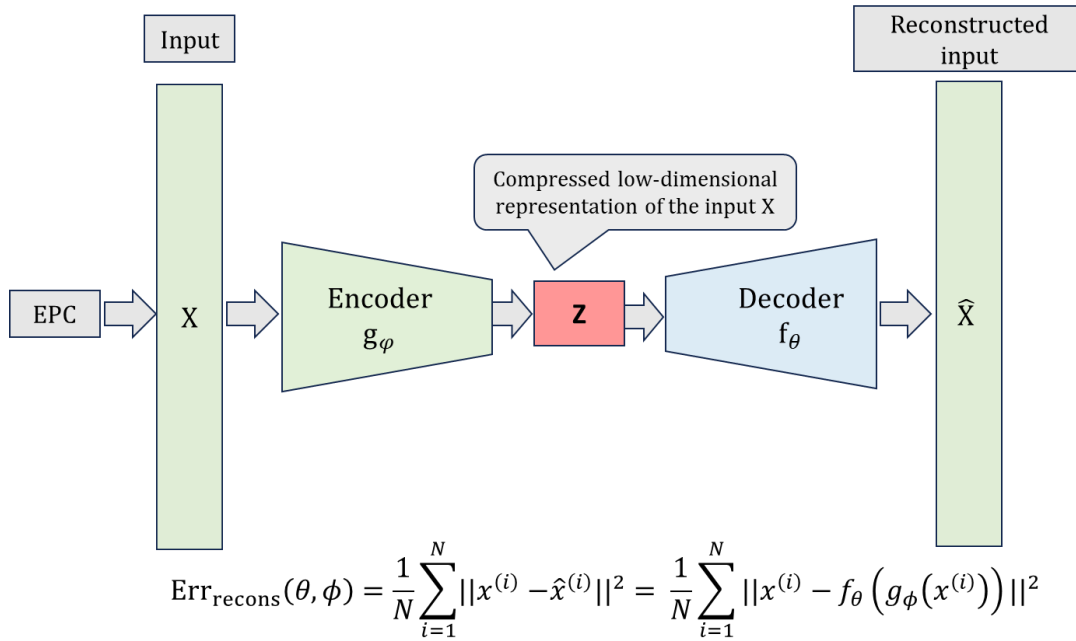
Autoencoders are neural network models designed to learn a compressed representation of multivariate data and to reconstruct the original input from this lower-dimensional encoding (Figure 7). When trained on a dataset that predominantly reflects regular or expected patterns, the autoencoder learns to reproduce such patterns efficiently, yielding low reconstruction error. Observations that diverge substantially from the learned structure (due to anomalous combinations of feature values or atypical relationships) result in higher reconstruction error, thereby distinguishing them from the majority of the data.

The use of autoencoders for outlier detection is particularly advantageous in high-dimensional settings, where traditional distance-based or threshold-based methods may suffer from reduced discriminative capacity. Autoencoders do not require pre-specified distributional assumptions and can capture complex, non-linear dependencies among variables. This makes them well-suited for identifying subtle deviations that would not be apparent using simpler diagnostic criteria.

In the context of the present analysis, the autoencoder was trained exclusively on the available data under the assumption that the majority of observations represent structurally consistent cases. The derived reconstruction error (or corresponding anomaly score) provides a continuous measure of deviation, which is subsequently stratified into ordinal anomaly categories to facilitate operational interpretation in subsequent

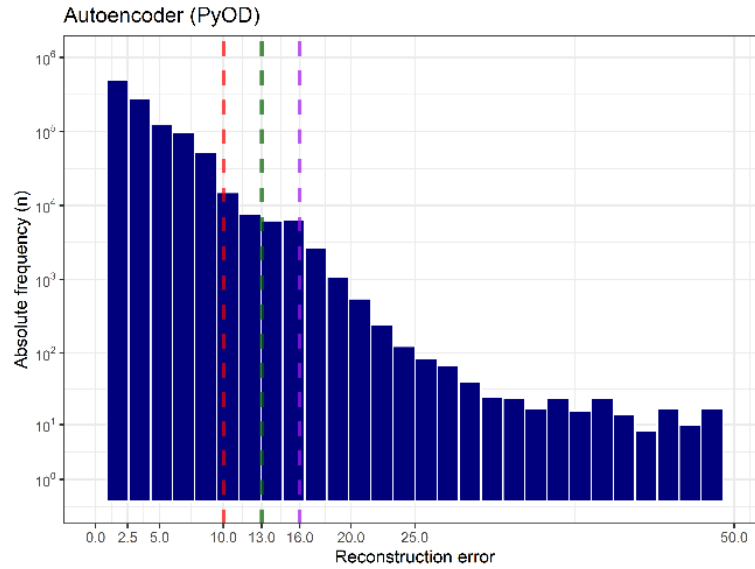
analyses. This methodology enables systematic identification of observations that warrant further review or validation due to their atypical multivariate profiles.

Figure 7. Block diagram of an autoencoder. The encoder g_ϕ maps the input data X into a latent representation Z , which is then used by the decoder f_θ to reconstruct the original input as $\hat{X}$. The model is trained by minimizing the reconstruction error $Err_{recons}(\theta, \phi)$ between X and $\hat{X}$.



The autoencoder was trained on standardized input features (internal preprocessing enabled), ensuring that all variables contributed on a comparable scale. The network architecture consisted of a symmetric feed-forward structure with two encoding layers (64 and 32 units) and their mirrored decoding counterparts (32 and 64), using $ReLU(x) = (0, x)$ activation functions to capture non-linear relationships in the data. Regularization was provided through dropout (0.20) and batch normalization, which together help prevent overfitting and stabilize training dynamics. Model optimization was performed using the Adaptive Moment Estimation (Adam) algorithm (learning rate = 0.001) for 30 epochs with a batch size of 32. The resulting reconstruction error provided a continuous anomaly score, which was subsequently stratified into ordered anomaly categories for interpretability and further analytical use.

Figure 8. Frequency distribution of the reconstruction error computed using the Autoencoder algorithm for the 1,114,184 EPCs. The y-axis is shown on a logarithmic scale to enhance the visibility of the right tail. The dashed lines indicate the thresholds applied for score assignment.



The distribution of the anomaly score produced by the autoencoder is reported in Figure 8. To support interpretation, the continuous score was partitioned into four ranges: < 10 , $[10, 13)$, $[13, 16)$, and ≥ 16 , which were mapped to penalty levels 0, 1, 2, and 3. Increasing penalty levels indicate increasingly atypical multivariate profiles and, consequently, reduced reliability of the corresponding observations. The absolute (and relative) frequencies associated with these levels were: 1,060,615 (96.9%), 16,274 (1.5%), 11,369 (1.04%), and 6,792 (0.62%), respectively.

6.5 Distribution of the Overall Data Quality Score (W_{tot})

The overall data quality score ($W_{tot} = W_{tot_S1} + W_{tot_S2} + W_{tot_S3} + W_{tot_S4}$), which summarizes the quality level of the energy performance certificates within the data quality management system, was examined through its frequency distribution (Figure 9). In Figure 10 the y-axis is displayed on a logarithmic scale to enhance the visibility of less frequent values and to illustrate the full range of the distribution.

Figure 9. Histogram of the overall score ($W_{tot} = W_{tot_S1} + W_{tot_S2} + W_{tot_S3} + W_{tot_S4}$), showing its frequency distribution in the study sample.

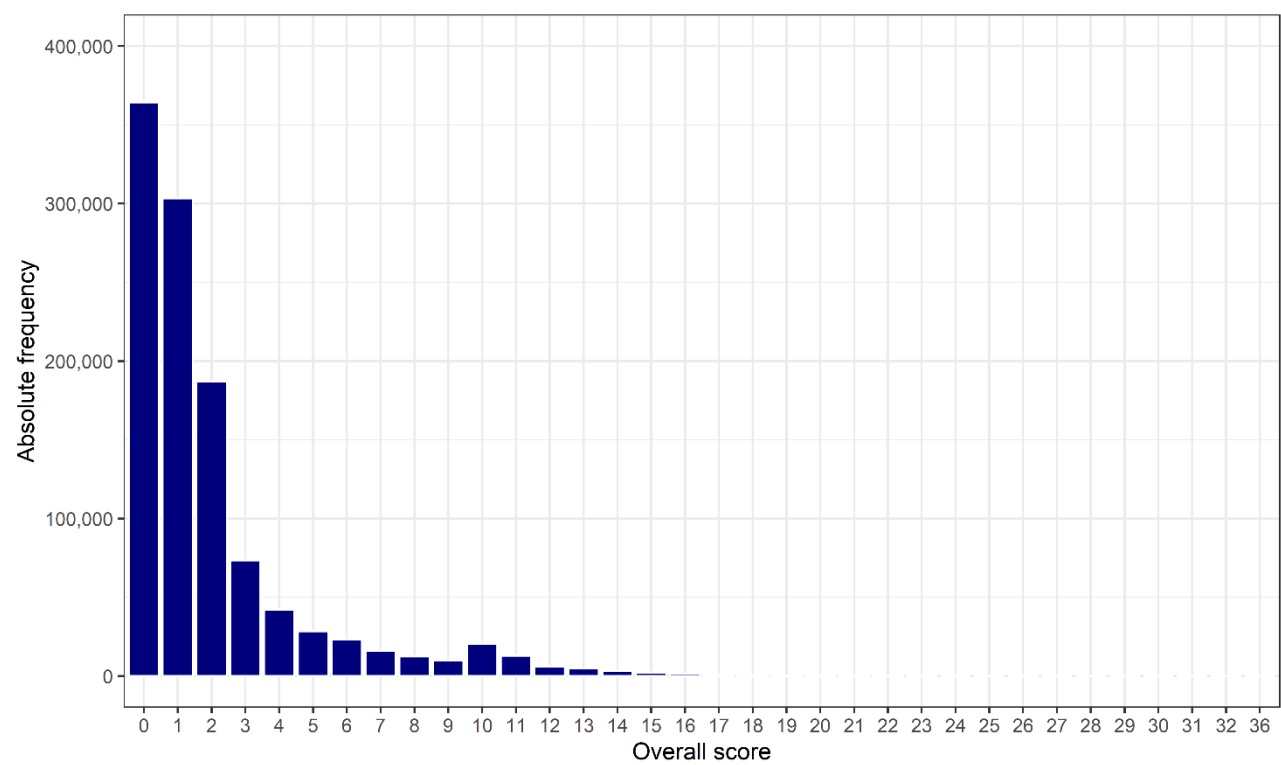
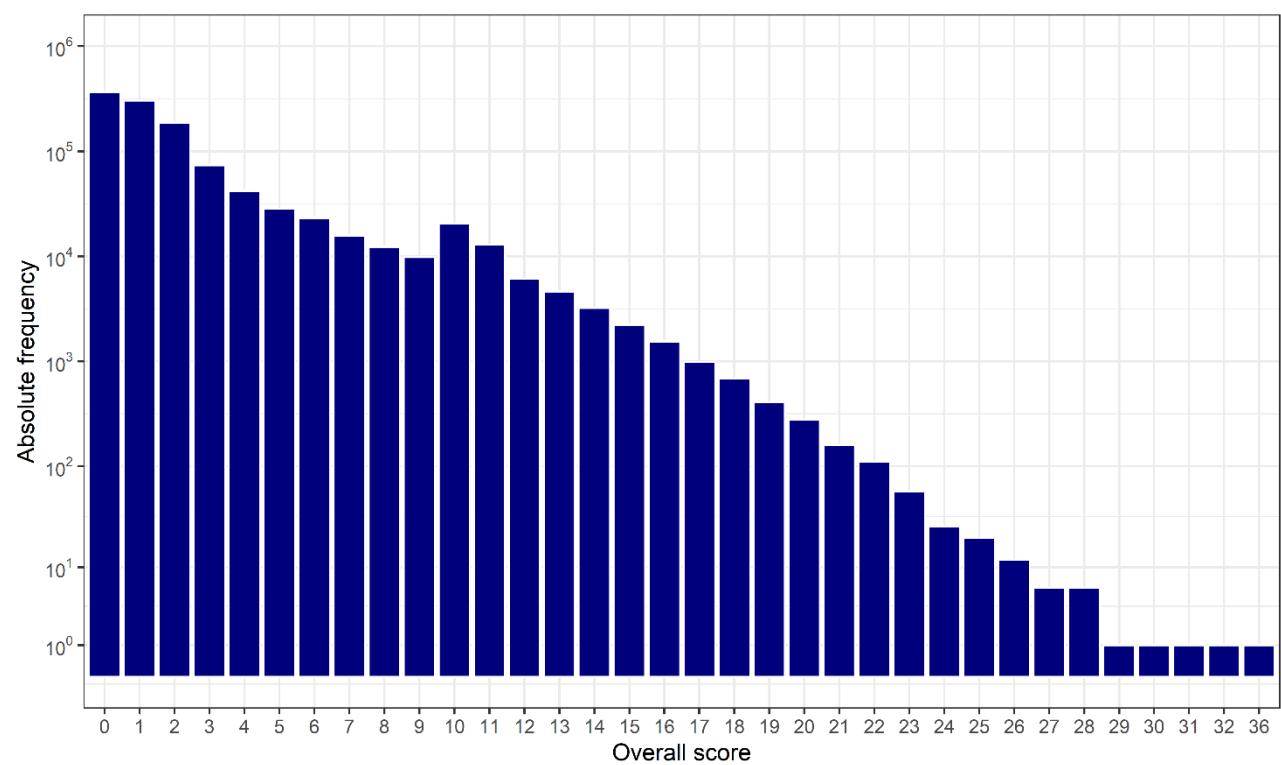


Figure 10. Histogram of the overall score ($W_{tot} = W_{tot_S1} + W_{tot_S2} + W_{tot_S3} + W_{tot_S4}$), showing its frequency distribution in the study sample. The y-axis is plotted on a logarithmic scale to improve the visibility of less frequent values and better illustrate the full range of the distribution.



References

- W. S. Al Farizi, I. Hidayah and M. N. Rizal, Isolation Forest Based Anomaly Detection: A Systematic Literature Review, 2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Semarang, Indonesia, 2021, pp. 118-122, doi: 10.1109/ICITACEE53184.2021.9617498.
- C.M. Bishop and H. Bishop (2023). Deep Learning. Foundations and Concepts. Springer Cham.
- C. Buratti, M. Barbanera, and D. Palladino (2014). An original tool for checking energy performance and certification of buildings by means of Artificial Neural Networks. *Applied Energy*, 120, 125-132. doi: 10.1016/j.epcnenergy.2014.01.053
- T. Chen, and C. Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 785–794. Association for Computing Machinery, New York, NY, USA. doi: 10.1145/2939672.2939785
- S. Hariri, M.C. Kind, and R.J. Brunner (2021). Extended Isolation Forest. *IEEE Transactions on Knowledge & Data Engineering*, 33 (4), 1479-1489. doi: 10.1109/TKDE.2019.2947676.
- H. Lin, R. Zhang, and T. Tong (2025) When Tukey Meets Chauvenet: A New Boxplot Criterion for Outlier Detection, *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2025.2520577
- F. T. Liu, K. M. Ting, and Z.-H. Zhou (2008). Isolation forest. *Proc. 8th IEEE Int. Conf. Data Mining*, 413–422. doi: 10.1109/ICDM.2008.17.
- U. Michelucci (2022). An Introduction to Autoencoders. arXiv e-prints <https://arxiv.org/abs/2201.03898>
- Shmuel, A., Glickman, O., & Lazebnik, T. (2025). A comprehensive benchmark of machine and deep learning models on structured data for regression and classification. *Neurocomputing*, 655, 131337. doi: 10.1016/j.neucom.2025.131337