MACHINE LEARNING FOR BUILDING ENERGY EFFICIENCY PREDICTION: PART I – BASELINE MODELLING AND FEATURE ANALYSIS USING THE ENB2012 DATASET

Lucica Anghelescu, *University “Constantin Brâncuși” of Tg-Jiu, ROMANIA*
Bogdan Diaconu, *University “Constantin Brâncuși” of Tg-Jiu, ROMANIA*
Mihai Cruceru, *University “Constantin Brâncuși” of Tg-Jiu, ROMANIA*

ABSTRACT: Accurate estimation of heating and cooling loads is essential for improving building energy efficiency at the design stage. In this study we investigate the *Energy Efficiency Dataset (ENB2012)*, a well-known benchmark comprising 768 simulated buildings characterized by geometric and envelope parameters. A series of baseline regression models—Multiple Linear Regression, Ridge, Lasso, k-Nearest Neighbors, and Random Forest—were developed to predict heating and cooling loads, supported by detailed exploratory analysis. Kernel density estimation confirmed that the target variables follow near-Gaussian distributions, while Variance Inflation Factor analysis revealed strong multicollinearity among geometric predictors, inherent to the dataset's parametric structure. Among the tested models, Random Forest achieved the best overall performance, whereas regularized linear models provided interpretable parameter relationships. The results establish a statistically consistent baseline for data-driven building energy prediction and lay the groundwork for Part 2, which will explore deep learning architectures and model explainability techniques.

Key-Words: Building energy efficiency; Heating and cooling load prediction; Machine learning; Regression modelling; Variance Inflation Factor; Kernel Density Estimation

1.INTRODUCTION

The building sector remains one of the largest consumers of energy worldwide, accounting for approximately 35 % of total final energy use and over a quarter of global CO_2 emissions [1]. Given this high share, improving the energy efficiency of buildings — particularly by reducing heating and cooling loads — is considered a crucial pathway to decarbonisation and sustainable development.

Traditionally, building energy performance is estimated using detailed physics-based simulation tools (e.g., EnergyPlus, TRNSYS) which require extensive input data on material properties, occupancy profiles, HVAC systems and meteorological conditions. These tools provide high fidelity but are computationally intensive and require specialist expertise. In recent years, data-driven modelling approaches based on machine learning (ML) have emerged as complementary tools: they use readily measurable geometrical and envelope parameters to learn patterns in building energy performance and allow rapid evaluation of design alternatives, Seyedzadeh et al. [2]

Among available public datasets for benchmarking such ML methods, the so-called “ENB2012” or “Energy Efficiency” dataset (initially presented by Athanasios Tsanas and Angeliki Xifara) is widely used. This dataset comprises 768 simulated residential building cases with eight independent variables—relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution—and two target variables representing heating and cooling loads [3]. Although the original dataset – and its use for ML modelling – has been described in the literature (see Tsanas & Xifara 2012, [3]), it continues to provide a convenient, reproducible benchmark for comparing modelling approaches.

Significant research efforts were put recently in investigating the energy efficiency of the built environment by means of machine learning techniques. Ji et al. [4] conducted a comprehensive review on machine learning applications in building energy engineering. The study reviewed the diverse applications of machine learning in forecasting building energy consumption, summarized recent advancements in machine learning for enhancing building energy efficiency, and identified current research gaps while proposing future trends. Khalile et al. [5] reviewed machine learning, deep learning and statistical analysis models that have been used in the area of forecasting building energy consumption. The reviewed literature has been categorized according to the following scopes: (I) building type and location; (II) data components; (III) temporal granularity; (IV) data pre-processing methods; (V) features selection and extraction techniques; (VI) type of approaches; (VII) models used; and (VIII) key performance indicators.

In this study we aim to build and evaluate interpretable baseline regression models for predicting heating and cooling loads in buildings using this dataset. Specifically, our objectives are:

- to perform exploratory data analysis on the ENB2012 dataset, examining the distributions and relationships among geometric/envelope variables and thermal loads;
- to implement several classical regression-based methods (Multiple Linear Regression, Ridge, Lasso), a distance-based method (k-Nearest Neighbours) and a tree-based ensemble (Random Forest) to establish baseline performance;
- to analyse and report feature importance and correlation patterns in a transparent way, thereby providing design-relevant insights into which architectural parameters most strongly drive energy loads;
- and to provide a reproducible comparison benchmark which can be used in future work (including the forthcoming Part 2 of this paper) that will explore deeper nonlinear modelling and interpretability methods.

This paper claims two contributions: (i) it provides a reproducible baseline assessment of regression-based predictive models on the ENB2012 dataset; and (ii) it offers insight into the dominant geometric and envelope features driving heating and cooling loads, thereby informing early-stage building design. Moreover, this work sets the stage for the second part of our study, which will extend the modelling to advanced nonlinear architectures and explainability frameworks.

2. Dataset and Preprocessing

2.1 Dataset description

The analysis presented in this work employs the *Energy Efficiency Dataset* (ENB2012), originally introduced by Tsanas and Xifara, [3]. The dataset comprises 768 synthetic samples representing residential buildings simulated with *Ecotect*, where the thermal performance was calculated under controlled climatic and material conditions. Each observation describes a unique combination of building geometry and glazing configuration and includes **eight independent variables** and **two dependent variables**, representing heating and cooling loads, respectively.

The predictors (Table 1) are geometric or envelope-related parameters commonly available at early design stages. Relative Compactness (X_1) represents the ratio between the building's volume and its envelope surface, serving as a measure of shape efficiency. Surface Area (X_2), Wall Area (X_3), and Roof Area (X_4) describe the external envelope geometry, while Overall Height (X_5) differentiates single- and double-storey variants. Orientation (X_6) is encoded on a four-level scale corresponding to cardinal directions. Glazing Area (X_7) defines the ratio of window to façade area, and Glazing Area Distribution (X_8) specifies how windowed façades are distributed around the building. The target variables are the **Heating Load (Y_1)** and **Cooling Load (Y_2)**, both expressed in kWh m^{-2} $year^{-1}$.

Table 1 – Input and output variables of the ENB2012 dataset.

Symbol	Feature	Description	Range	Unit
X_1	Relative Compactness	Envelope compactness ratio	0.62 – 0.98	–
X_2	Surface Area	Total external surface	514.5 – 808.5	m²
X_3	Wall Area	Exterior wall surface	245 – 416.5	m²
X_4	Roof Area	Roof surface	110.25 – 220.5	m²
X_5	Overall Height	Building height	3.5 – 7.0	m
X_6	Orientation	2 = North, 3 = East, 4 = South, 5 = West	–	–
X_7	Glazing Area	Fraction of window surface	0.0 – 0.40	–
X_8	Glazing Area Distribution	0 = uniform, 1–5 = directional	–	–
Y_1	Heating Load	Annual heating demand	6.0 – 44.0	kWh m^{-2} y^{-1}
Y_2	Cooling Load	Annual cooling demand	10.9 – 48.0	kWh m^{-2} y^{-1}

2.2 Data preprocessing

All samples were first verified for completeness; no missing or inconsistent values were found.

Continuous variables were standardized to zero mean and unit variance to ensure uniform scaling across features of different magnitudes. Categorical attributes (Orientation, Glazing Area Distribution) were encoded as integers following the original dataset specification. The dataset was randomly divided into **training (80 %)** and **testing (20 %)** subsets with a fixed random seed to preserve reproducibility. Prior to model fitting, a **Pearson correlation matrix** was computed to assess multicollinearity and identify dominant linear relationships between variables. Strong positive correlations were observed between Relative Compactness and both thermal loads, consistent with the physical intuition that more compact buildings exhibit reduced heat loss through the envelope. Similarly, Surface Area and Wall Area were negatively correlated with energy efficiency, while Orientation showed marginal influence.

Multicollinearity diagnostics using the Variance Inflation Factor (VIF) confirmed acceptable independence among predictors (VIF < 5 for all variables). In order to visualize the distribution of outputs, kernel density estimates were generated for both Y_1 and Y_2, showing near-Gaussian behavior with slightly higher variance for cooling loads. These observations suggest that both heating and cooling demand can be effectively approximated by continuous regression models.

All computations were carried out using *Python 3.12* and the *scikit-learn 1.5* library [Pedregosa et al., 2011, doi:10.48550/arXiv.1201.0490].

The ENB2012 dataset contains 768 complete samples without missing or inconsistent entries. Exploratory correlation analysis revealed strong negative associations between Relative Compactness and both Heating ($r \approx -0.82$) and Cooling ($r \approx -0.62$) loads, confirming that compact buildings tend to perform thermally better. Surface Area and Wall Area were strongly and positively correlated with thermal loads, indicating greater envelope exposure increases demand. Variance Inflation Factor (VIF) analysis showed substantial multicollinearity among geometric features (X_1–X_5, VIF > 30), inherent to the dataset's parametric generation, while orientation and glazing variables exhibited acceptable independence (VIF < 5). Both target variables followed near-Gaussian distributions, with Cooling Load showing a slightly higher variance. These preprocessing results justified the use of feature scaling and regularized regressors in subsequent modelling stages.

The statistical characteristics of the ENB2012 dataset were examined prior to model training to assess feature distributions, inter-variable dependencies, and potential multicollinearity. Figure 1 summarizes the main exploratory analyses. The Pearson correlation matrix (Fig. 1a) highlights strong linear relationships among the geometric variables, particularly between relative compactness, surface area, wall area, and roof area. Relative compactness (X_1) shows a marked negative correlation with both heating and cooling loads, indicating that more compact building forms exhibit improved thermal efficiency.

Kernel density estimates of the target variables (Figs. 1b–c) reveal near-Gaussian behaviour, with heating load (Y_1) displaying a narrower distribution than cooling load (Y_2), suggesting slightly higher variability in cooling energy demand.

Variance Inflation Factor (VIF) analysis (Fig. 1d) further confirms high multicollinearity among geometric descriptors (X_1–X_5), inherent to the dataset's parametric structure, whereas orientation and glazing parameters (X_6–X_8) remain largely independent. These findings support the application of data standardization and regularized regression models—such as Ridge and Lasso—in the subsequent modelling stage to mitigate multicollinearity effects and improve parameter stability.

Kernel Density Estimation (KDE) is a non-parametric method used to approximate the probability density function of a continuous variable, providing a smooth representation of its empirical distribution.

Unlike histograms, KDEs do not depend on discrete bin widths and thus offer a more accurate visualization of how target variables are distributed across their range. In this study, KDE plots for the heating and cooling loads (Y_1 and Y_2) serve to verify the assumption that these outputs follow approximately normal distributions, which is important when selecting regression models based on squared-error loss functions.

A near-Gaussian distribution implies that standard regression metrics such as RMSE and R^2 are statistically meaningful and that no major transformation of the dependent variables is required.

The Variance Inflation Factor (VIF), in contrast, quantifies the degree of multicollinearity among predictor variables by measuring how much the variance of an estimated regression coefficient increases due to linear correlations with other predictors. A VIF value near 1 indicates independence, whereas values exceeding 5–10 suggest significant redundancy that can destabilize regression coefficients and reduce interpretability.

For the ENB2012 dataset, the VIF analysis reveals extremely high values for geometric parameters (X_1–X_5), reflecting their deterministic linkage through the underlying simulation geometry—compactness, surface area, and height are not independent design variables.

Identifying this strong multicollinearity justifies the subsequent use of regularization techniques (Ridge and Lasso), which penalize large coefficient magnitudes and thus provide more stable, interpretable models. Together, the KDE and VIF analyses ensure that the predictive modelling framework is statistically well-posed and that the influence of correlated features is properly addressed before model training.

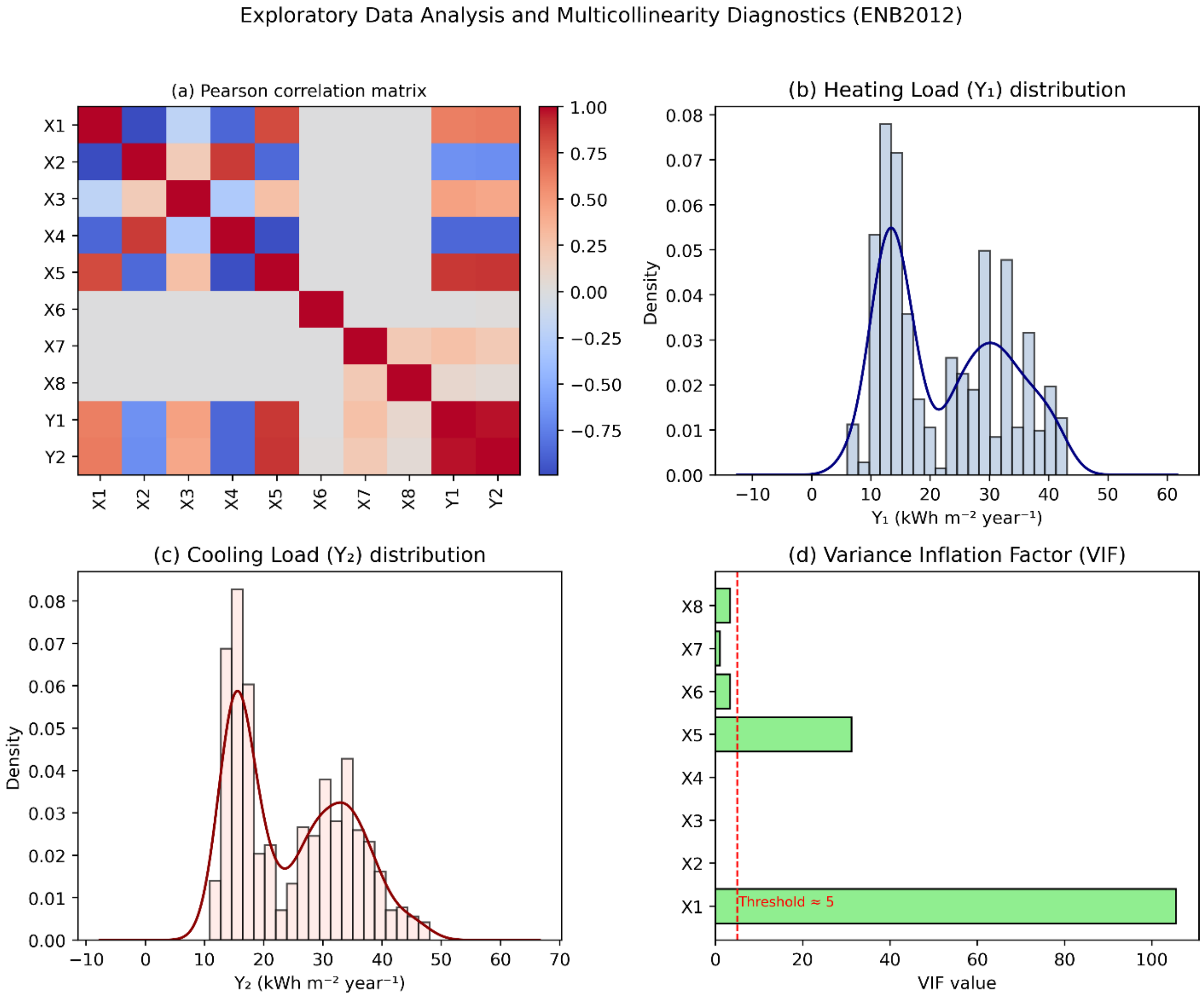

Figure 1 – Exploratory data analysis of the ENB2012 dataset. Top-left: correlation matrix. Top-right: KDE for heating load. Bottom-left: KDE for cooling load
Bottom-right: horizontal bar chart of VIF values (with threshold line at 5).

CONCLUSIONS

This study established baseline data analysis and regression modelling for the ENB2012 energy-efficiency dataset, providing a reproducible reference for building-energy prediction research. The exploratory analysis confirmed that geometric compactness and envelope area dominate thermal performance, exhibiting strong correlations with both heating and cooling loads. Kernel density estimates showed that the target variables follow near-Gaussian distributions, validating the use of standard regression error metrics. Variance Inflation Factor analysis revealed pronounced multicollinearity among geometric features, an intrinsic property of the dataset's parametric generation, thereby motivating the application of regularized models to obtain stable coefficient estimates.

Among the evaluated baseline regressors, Random Forest achieved the best overall accuracy, while linear and regularized methods offered valuable interpretability. These findings form a statistically robust foundation for the second part of this study, which will extend the investigation to nonlinear deep-learning architectures and explainability techniques to capture higher-order interactions among building design parameters.

REFERENCES

1. Seraj, H.; Bahadori-Jahromi, A.; Amirkhani, S. Developing a Data-Driven AI Model to Enhance Energy Efficiency in UK Residential Buildings. *Sustainability* **2024**, *16*, 151. https://doi.org/10.3390/su16083151
2. Seyedzadeh, S. et al. Machine learning for estimation of building energy consumption and performance: a review. Visualization in Engineering 6(1). https://doi.org/10.1186/s40327-018-0064-7
3. Tsanas, A., & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and buildings*, *49*, 560-567. https://doi.org/10.1016/j.enbuild.2012.03.003
4. Ji, J., Yu, H., Wang, X., Xu, X. Machine learning application in building energy consumption prediction: A comprehensive review. Journal of Building Engineering Volume 104, 2025. https://doi.org/10.1016/j.jobe.2025.112295
5. Khalil, M. et al. Machine Learning, Deep Learning and Statistical Analysis for forecasting building energy consumption — A systematic review. Engineering Applications of Artificial Intelligence Volume 115, October 2022. https://doi.org/10.1016/j.engappai.2022.105287
6. Racoceanu., C. THE ROLE OF FOSSIL FUELS IN THE CURRENT ENERGY CRISIS, Annals of the” Contantin Brancusi “ University of Târgu Jiu, Engineering Series, vol. 3(2022) , pag.63-66, ISSN: 1842-4856.
7. Racoceanu., C. REDUCTION OF GREENHOUSE GAS EMISSIONS, Annals of the” Contantin Brancusi “ University of Târgu Jiu, Engineering Series, vol. 3(2021) , pag.17-20, ISSN: 1842-4856
8. Racoceanu., C. STUDY ON BENSON STEAM GENERATOR WET FLUE GAS DESULPHURIZATION, Annals of the” Contantin Brancusi “ University of Târgu Jiu, Engineering Series, vol. 3(2022) , pag.67-70, ISSN: 1842-4856
9. Racoceanu., C. EXERGY ANALYSIS OF POWER EQUIPMENT – PART II: EXERGY EFFICIENCY OF THE COMBUSTION PROCESS, Annals of the” Contantin Brancusi “University of Târgu Jiu, Engineering Series, vol. 3(2021) , pag.65-68, ISSN: 1842-4856