Ensembles for Regression: A Statistical Analysis and Recommendations

1st Halcyon Davys P. de Carvalho
dept. of Computer Engineering
University of Pernambuco
Recife, Brazil
hdpc@ecomp.poli.br

2nd João Fausto L. de Oliveira
dept. of Computer Engineerin
University of Pernambuco
City, Country
fausto.lorenzato@upe.br

3rd Roberta Andrade de A. Fagundes
dept. of Computer Engineerin
University of Pernambuco
Recife, Brazil
roberta.fagundes@upe.br

***Abstract*—Ensemble techniques have proven effective in improving the performance of regression models. However, the decision between homogeneous and heterogeneous strategies is often made empirically, without considering the statistical properties of the data. This paper presents a framework that conducts a statistical analysis of datasets to recommend the most appropriate ensemble strategy. The framework leverages metrics such as coefficient of variation, skewness, kurtosis, correlation, and outlier presence to classify datasets into homogeneous or heterogeneous profiles. Ten real-world datasets with varying levels of complexity were evaluated. The results show that the proposed framework achieved the best performance in 80% of the cases, confirming that the statistical structure of the data should guide the selection of ensemble strategies. Wilcoxon statistical tests further validated the significance of the results. The proposed framework offers a systematic alternative to guide ensemble decisions in regression tasks, enhancing accuracy and reducing reliance on empirical trial-and-error procedures.**

***Index Terms*—Ensemble learning, Regression, Statistical analysis, Model recommendation, Machine learning**

I. Introduction

In recent years, ensemble learning techniques have stood out in the context of regression tasks, offering significant improvements in model robustness and generalization capabilities [1]–[3]. In particular, heterogeneous ensembles — composed of models with different architectures — have shown superior performance in scenarios characterized by high variability and structural complexity in the data [4]–[6]. However, the effectiveness of an ensemble technique strongly depends on the statistical characteristics of the dataset to which it is applied, such as distribution, correlation, dispersion, and the presence of outliers [7]–[9].

In many studies from the literature, the selection between homogeneous or heterogeneous ensembles is often carried out in a fixed or empirical manner, without explicitly considering the statistical aspects of the dataset, such as distribution, variance, correlation between variables, and local density [4], [9]–[11]. This practice overlooks the fact that certain ensemble configurations may be more appropriate for specific data profiles, directly impacting predictive performance [6]–[8].

Although there are studies that relate data characteristics to the difficulty of classification or regression [7], [8], there is still a gap regarding the formalization of strategic criteria that guide the choice of ensemble type based on the statistical properties of datasets. Works such as Tsymbal's [12] discuss how distribution shifts impact model performance, but they do not establish practical guidelines for recommending ensemble architectures.

In addition to addressing classical challenges of regression, this work is also inspired by approaches that leverage statistical data information to support modeling decisions. For instance, in industrial contexts, tree-based surrogate regression models have been successfully used in multi-objective optimization problems, combining predictive performance with statistical sensitivity in decision-making [13]. Similarly, the application of statistical preprocessing to feed neural networks has proven effective in decision support systems, even in distinct domains such as ultrasonic inspection [14]. These approaches reinforce the potential of statistical analysis as a technical foundation for defining modeling strategies that are better aligned with data profiles.

In this paper, we propose a systematic approach for the statistical analysis of regression datasets, aiming to formulate strategic recommendations regarding the most suitable ensemble type. The methodology is based on metrics such as the coefficient of variation, attribute correlation, skewness, kurtosis, N3 score, and the presence of outliers, associating these measures with the performance of different ensemble configurations across multiple datasets.

This approach may serve as a practical guide for data scientists and machine learning engineers in defining the most appropriate ensemble architecture through a preliminary analysis of the data.

The main contribution of this study is to offer a systematic and reproducible process to guide the choice between homogeneous and heterogeneous ensembles based on statistical metrics extracted from the data. In addition to enhancing adaptability and accuracy in regression tasks, this approach paves the way for integration into AutoML platforms, where automated, data-driven model selection is essential [15]–[17]. Recent frameworks such as Auto-sklearn 2.0 adopt portfolio-based meta-learning strategies to automate pipeline design, reinforcing the importance of prior knowledge for model recommendation. By reducing reliance on empirical heuristics, our framework supports more interpretable and scalable machine learning workflows.

This article is organized as follows: Section 2 presents the related work; Section 3 describes the statistical analysis methodology; Section 4 details the experiments and results; and Section 5 discusses the conclusions and future directions.

II. Related Work

The performance of regression models depends not only on their internal structure but also on the statistical characteristics of the data to which they are applied. Classical studies on the complexity of supervised learning problems, such as those by Ho and Basu [7] and Garcia et al. [8], introduced metrics that help in understanding the intrinsic difficulty of datasets, taking into account aspects such as variability, distribution, outliers, and attribute redundancy. Although initially developed for classification tasks, many of these indicators are applicable to regression, particularly in guiding model selection or combination.

In the field of ensemble learning, methods such as Bagging and Boosting have been widely used to improve the accuracy of predictive models [2], [3]. Systematic reviews indicate that heterogeneous ensembles — composed of models with different architectures — tend to perform well in domains with high structural complexity [4]–[6]. Cruz et al. [6], for instance, highlight the importance of diversity and local competence in dynamic model selection.

More recent research has focused on the use of meta-learning to guide algorithm selection based on data characteristics. Reif et al. [9] explore hyperparameter optimization through meta-level learning. Bilalli et al. [10] analyze how statistical properties such as variance, skewness, and correlation can be used to inform model selection. Alcobaça et al. [11] apply this approach to software effort estimation, suggesting that ensembles can be recommended based on simple meta-features.

Recent advances in AutoML, such as Auto-sklearn 2.0 [17], highlight how historical performance can guide model selection without relying on meta-features. Its PoSH strategy (Portfolio Success History) automates configuration under time constraints but does not explicitly use statistical descriptors from the current dataset. In contrast, our framework incorporates such descriptors to recommend ensemble strategies for regression, providing an interpretable and complementary path to automation. This underscores the growing importance of data-driven guidance in model selection—an approach central to our work.

The No Free Lunch Theorem, formulated by Wolpert and Macready [18], states that no learning algorithm is universally superior across all contexts, which supports the need for approaches that adapt to the specific properties of the data. In this regard, Woloszynski et al. [19] argue that model competence varies according to local density and the structure of the input space, indicating that different regions of the data may require distinct modeling strategies.

While most ensemble-related studies focus on structural variations or combination strategies, some recent works highlight the role of statistical analysis in model selection. In [13], the authors apply Extra Trees regression to build surrogate models in automotive simulation tasks, using metrics such as correlation and sensitivity to guide the process. In [14], techniques such as PCA, DCT, and HHT are employed to transform the data before applying neural networks, demonstrating how statistical preprocessing can influence the performance of machine learning models. Although these works do not directly address the selection between ensemble strategies, they support the relevance of statistical features in predictive decision-making.

Finally, although studies such as that by Tsymbal [12] acknowledge that distribution shifts in data directly impact model performance, there remains a gap in the formulation of practical rules that associate statistical data properties with the choice of ensemble type. This work aims to address that gap by proposing a statistical analysis and empirical evaluation-based approach to support strategic decisions in building regression ensembles.

III. Statistical Analysis Methodology

This section describes the proposed methodology for analyzing the statistical characteristics of regression datasets and defining strategic recommendations for selecting the most suitable ensemble type. The approach is based on the hypothesis that different statistical profiles in the data require distinct model combination strategies. This premise is supported by studies that associate data structure with learning task complexity [7], [8], and by the principle that algorithm performance varies according to data distribution and variability [18], [19].

Figure 1 illustrates the methodological workflow adopted in this study, divided into five sequential and interdependent steps that reflect the recommendation logic of the proposed framework:

- **A) Regression Dataset** — represents the datasets used in this study, detailed in Section III-A. Ten datasets were selected with varying levels of complexity and origin, to ensure both statistical diversity and practical relevance;
- **B) Extraction of Statistical Features** — discussed in Section III-B, involves the collection of descriptive metrics that characterize the internal structure of the data, such as coefficient of variation, skewness, kurtosis, average attribute correlation, N3 score, and number of outliers;
- **C) Dataset Classification** — presented in Section III-C, categorizes each dataset based on the extracted metrics, classifying them as *homogeneous* or *heterogeneous*, according to predefined criteria;
- **D) Recommended Strategy** — described in Section III-D, where the most appropriate ensemble type — homogeneous or heterogeneous — is suggested based on the identified statistical profile;
- **E) Strategy Application in the Framework** — corresponds to the experimental stage presented in Section IV, in which the recommended strategies are applied and their performance is compared to fixed approaches, aiming to validate the effectiveness of the recommendations.

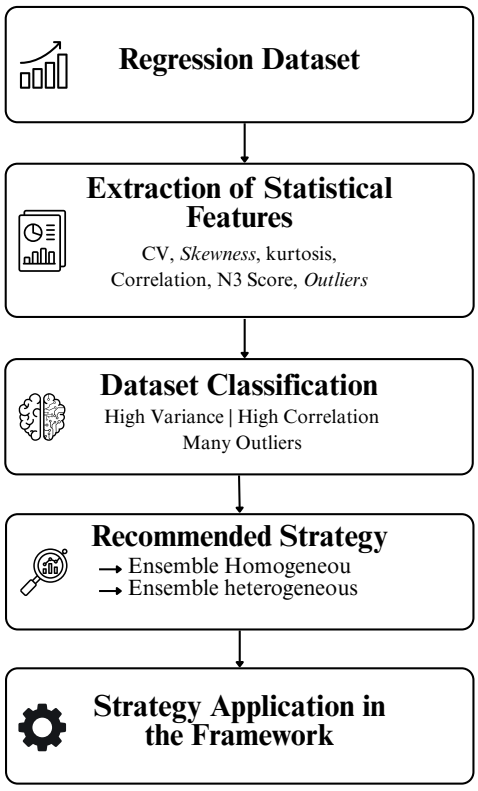

Fig. 1. Strategic Ensemble Recommendation Workflow Based on Dataset Statistical Analysis.

A. Regression Datasets

The experiments were conducted using **10 regression datasets**, selected to represent different levels of complexity, statistical variability, and application domains. The datasets span areas such as software engineering, energy efficiency, acoustics, and chemistry, and are widely used in the scientific literature.

Table I summarizes the main characteristics of each dataset, including the number of instances, predictive attributes, and source — with data obtained from the *UCI Machine Learning Repository* [20] and the *PROMISE repository* [21].
The dataset diversity enables the evaluation of the proposed framework's robustness across different statistical structures — from datasets with low variability and symmetry to those with strong asymmetry, high dispersion, and significant presence of outliers. This selection is essential for validating the generalization capacity of the strategic ensemble recommendations in multiple regression scenarios.

B. Statistical Feature Extraction

In the first stage (see Figure 1), statistical metrics are extracted from the datasets in order to characterize their internal structure and variability. The following metrics were used:

TABLE I
CHARACTERISTICS OF THE DATASETS USED IN THE PROPOSED FRAMEWORK

Dataset	Instances	Attributes	Source
Abalone	4177	8	UCI [20]
Airfoil Self Noise	1503	5	UCI [20]
Bike Sharing	731	14	UCI [20]
Cocomo81	63	7	PROMISE [21]
Concrete	1030	8	UCI [20]
Desharnais	81	6	PROMISE [21]
Energy Efficiency	768	9	UCI [20]
Housing	506	13	UCI [20]
NASA93	93	22	PROMISE [21]
Wine Quality (Red)	1599	12	UCI [20]

Source: Prepared by the authors based on UCI and PROMISE

- **Coefficient of Variation (CV)**: measures the relative dispersion of the target variable values, calculated as the ratio between the standard deviation and the mean [8].
- **Skewness**: quantifies the degree of symmetry in the data distribution [8].
- **Kurtosis**: indicates the presence of heavy tails and outliers [8].
- **Mean Correlation Between Attributes**: average Pearson correlation coefficients among the independent variables [7].
- **N3 Score**: inspired by Ho and Basu [7], estimates the local complexity of the input space based on nearest neighbors.
- **Number of Outliers**: calculated using the interquartile range (IQR), counting instances that fall outside $1.5 \times IQR$ [22].

These metrics provide a broad view of the statistical structure of the data, enabling the identification of relevant patterns for the next stage.

C. Structural Classification of the Dataset

Based on the extracted metrics, each dataset was classified according to its level of statistical complexity, as also depicted in Figure 1. The classification follows empirical criteria based on the combined behavior of the metrics, considering relative thresholds and insights from the literature on data complexity and learning difficulty [7], [8].

- **Homogeneous**: characterized by low variability (CV < 0.5), approximately symmetric distribution (skewness close to zero), low kurtosis (no heavy tails), and a small number of outliers. These datasets tend to exhibit a more stable and predictable statistical structure, favoring the use of simple ensemble strategies such as homogeneous ensembles, which are less sensitive to noise and local fluctuations.
- **Heterogeneous**: characterized by high variability (CV > 0.5), strong asymmetry (skewness > 1 or < -1), high kurtosis (indicating heavy tails), a significant number of outliers, and often high correlation between attributes. Datasets with this profile demand more flexible and robust models, such as heterogeneous ensembles, which

can capture distinct patterns and complex structural variations.

The classification was conducted based on a combined exploratory analysis, taking into account the extracted metrics, graphical visualizations, and descriptive statistics. This categorization serves as the foundation for the strategic recommendation of the most suitable ensemble type, as discussed in the next subsection. By aligning the statistical profile of the data with the ensemble architecture, the goal is to promote greater adaptability to the problem and improved predictive performance.

It is worth noting that the threshold values used for classification (e.g., CV > 0.5, skewness > 1, kurtosis > 3) were defined empirically based on literature insights. Although effective in our experiments, the sensitivity of these thresholds may vary depending on the application domain or dataset scale. A more robust and adaptive thresholding mechanism—potentially learned through meta-learning—remains an open direction for future research.

D. Strategic Ensemble Recommendation

The final step consists of associating each identified statistical profile with a specific ensemble strategy recommendation. The mapping logic, also shown in Figure 1, is summarized in Table II. The central hypothesis is that different statistical profiles require different combination strategies in order to maximize performance and generalization.

TABLE II
STRATEGIC RECOMMENDATION BASED ON THE STATISTICAL STRUCTURE OF THE DATA

Dataset Type	Recommended Strategy	Justification
Homogeneous	Homogeneous Ensemble	Promotes stability and avoids introducing noise in simple and predictable data.
Heterogeneous	Heterogeneous Ensemble	Leverages model complementarity to handle structural diversity and distinct local patterns.

Source: Proposed by the authors, inspired by [7]–[9], [12].

To provide a clear overview of the proposed strategy, Algorithm 1 presents the step-by-step logic implemented in the framework. It reflects the sequential operations described in the methodological workflow of Figure 1, from the extraction of statistical descriptors to the application of the recommended ensemble strategy.

This recommendation strategy will be empirically validated in the next section, which compares the results obtained using different ensemble configurations based on the datasets' statistical profiles.

The proposed methodology is easily reproducible, and its logic can be integrated into automated systems for ensemble configuration recommendation, with potential applications in *AutoML* platforms and data-driven preprocessing workflows.

Algorithm 1 Step-by-step procedure for dataset profiling, ensemble recommendation, and application within the proposed framework

Require: Dataset $\mathcal{D} = \{X, y\}$
Compute CV, Skewness and Kurtosis of y
Compute mean correlation among features in X
Compute N3 Score
Compute number of outliers in y (IQR method)
if CV > 0.5 **or** $|\text{Skewness}| > 1$ **or** Kurtosis > 3 **or** Outliers $> 5\%$ **then**
    Label dataset as **Heterogeneous**
else
    Label dataset as **Homogeneous**
end if
for all base models **do**
    Train model using 10-fold cross-validation (20 runs)
    Compute average MSE
end for
Select best model (lowest MSE)
Select top-3 models (lowest MSEs)
if Dataset is Homogeneous **then**
    Apply Bagging using the best model (5 estimators)
else
    Apply Average Ensemble using the top-3 models
end if
Evaluate ensemble on test data (MSE, MAE, R^2)
Ensure: Recommended ensemble type and performance metrics

IV. EXPERIMENTS AND RESULTS

A. Datasets

In this stage, ten regression datasets widely used in the literature were selected, originating from public repositories such as the *UCI Machine Learning Repository* [20] and *PROMISE* [21]. The selection was guided by statistical diversity criteria, aiming to represent various scenarios commonly found in real-world applications.

The datasets include cases with low, medium, and high variability, as well as different degrees of skewness, kurtosis, attribute correlation, and presence of outliers. This diversity is essential to assess the robustness of the proposed framework across different structural data profiles and to validate its ability to formulate strategic recommendations adapted to the statistical context of each problem.

Table III presents the actual values of the statistical metrics extracted from each dataset. Based on these values and the classification criteria defined in the methodology, each dataset was categorized according to its structural profile.

Note: The values presented were obtained using the methodology described in Section III-B, employing the scripts provided in the proposed framework. Data processing was carried out using 10-fold cross-validation repeated 20 times.

For instance, the **Airfoil Self Noise** dataset exhibits an extremely low coefficient of variation (CV = 0.06), slightly

TABLE III
STATISTICAL SUMMARY OF THE DATASETS USED

Dataset	CV	Skew.	Kurt.	Corr.	Outliers
Airfoil Self Noise	0.06	-0.42	-0.32	0.22	2.5%
Bike Sharing	0.43	-0.05	-0.81	0.17	0.0%
Wine Quality (Red)	0.22	0.22	0.29	0.20	5.6%
Energy Efficiency	0.45	0.36	-1.25	0.24	0.0%
Desharnais	0.87	1.97	4.36	0.67	6.7%
COCOMO81	2.64	4.37	20.08	0.24	7.4%
NASA93	1.81	4.19	21.81	0.17	5.1%
Housing	0.41	1.10	1.47	0.39	19.6%
Abalone	0.32	1.11	2.33	0.64	14.8%
Concrete	0.47	0.42	-0.32	0.21	1.7%

Source: Prepared by the authors

TABLE IV
ENSEMBLE RECOMMENDATION BASED ON STATISTICAL PROFILE

Dataset	Recommended Strategy
Airfoil Self Noise	Homogeneous Ensemble
Bike Sharing	Homogeneous Ensemble
Concrete	Homogeneous Ensemble
Energy Efficiency	Homogeneous Ensemble
Abalone	Heterogeneous Ensemble
COCOMO81	Heterogeneous Ensemble
Desharnais	Heterogeneous Ensemble
NASA93	Heterogeneous Ensemble
Housing	Heterogeneous Ensemble
Wine Quality (Red)	Heterogeneous Ensemble

Source: Prepared by the authors

negative skewness (-0.42), and negative kurtosis, along with a low incidence of outliers (2.5%). This profile suggests high stability, low dispersion, and absence of heavy tails, characterizing it as a typical case of a **homogeneous dataset**.

In contrast, the **COCOMO81** dataset has a very high CV (2.64), skewness of 4.37, and kurtosis of 20.08, in addition to a significant proportion of outliers (7.4%). These values reflect a highly skewed distribution with extreme values, representing a case of **pronounced statistical heterogeneity**.

The **Wine Quality (Red)** dataset, on the other hand, has a relatively low CV (0.22), mild skewness (0.22), and kurtosis close to normality (0.29), but a higher proportion of outliers (5.6%). Although some metrics suggest stability, the presence of outliers and local variability justify its classification as **heterogeneous** according to the criteria defined in this work.

This binary classification of statistical profiles is fundamental to the operation of the proposed framework, as it guides the strategic selection of the most appropriate ensemble type — homogeneous or heterogeneous — as discussed in the next subsection.

B. Experimental Settings

The evaluation metrics used were: Mean Squared Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). All experiments were conducted using 10-fold cross-validation, repeated 20 times, and the results were normalized to enable consistent comparisons.

C. Results and Comparative Analysis

This section presents the results obtained by applying the ensemble strategies to the various datasets according to their statistical profiles.

Based on the statistical analysis presented in Section III-D, Table IV summarizes the ensemble strategy recommended for each dataset. These recommendations are reflected in the second column of Table VI, allowing for a direct comparison between the recommended ensemble and the homogeneous and heterogeneous ensembles applied uniformly to all datasets.

Table VI summarizes the average performance (MSE) of the three evaluated approaches: homogeneous ensemble, heterogeneous ensemble, and the strategy recommended by the proposed framework based on dataset statistical analysis.

*a) **Performance of Base Models**:* Next, we present an individual analysis of the models used in the ensemble compositions. Table V shows the Mean Squared Error (MSE) and standard deviation obtained by each base model across the datasets. Values were calculated from 20 repeated executions using 10-fold cross-validation to ensure statistical robustness. The lowest MSE for each dataset is highlighted in bold.

The analysis of Table V reveals consistent patterns between the statistical profile of the data and the performance of the base models:

- In datasets with a **homogeneous profile**, such as **Airfoil Self Noise** and **Bike Sharing**, the lowest errors were achieved by simple, low-complexity models such as `CART` and `LINEAR`, as well as shallow neural networks like `FANN-1` and `FANN-2`. These results indicate that in scenarios with low variability and stable statistical structure, models with reduced parametric capacity are sufficient to achieve good generalization and low error.
- In datasets with greater complexity and asymmetry — such as **COCOMO81** and **Desharnais** — the best performances were achieved by more flexible models, including neural networks (`FANN-1`), linear regression (`LINEAR`), and kernel-based methods like `SVR-POLY3`. The `RBF` model also performed well in several of these datasets. These datasets, classified as **heterogeneous**, showed higher standard deviations, reflecting the instability of less robust methods when facing structural data diversity.
- In other datasets also classified as **heterogeneous**, such as **Energy Efficiency**, **Wine Quality (Red)**, **Housing**, and **Concrete**, the lowest errors were often obtained by tree-based models (`CART`) and simple neural networks (`FANN-1` and `FANN-2`). These results suggest that even in scenarios with moderate heterogeneity, models with nonlinear approximation capability and low complexity can achieve excellent performance, especially when there is a balance between variability and local stability in the data.
- In terms of **robustness**, models such as `FANN-1`, `FANN-2`, `CART`, and `RBF` demonstrated stable and competitive performance across different scenarios. Although they did not always yield the lowest absolute error, these models stood out for their consistency, justifying their

TABLE V
MEAN SQUARED ERROR (MSE) AND STANDARD DEVIATION PER MODEL AND DATASET

Dataset	CART	LINEAR	FANN-1	FANN-2	SRV-RBF	SVR-POLY1	SVR-POLY3	RBF	3NN	5NN
abalone	0.01185 (0.02668)	0.00622 (0.01502)	**0,00571 (0.01405)**	0,00588 (0.02083)	0.00616 (0.01562)	0.00700 (0.01449)	0.00653 (0.01359)	0.00584 (0.01382)	0,00733 (0,01703)	0.00665 (0.01617)
airfoil_self_noise	**0.00521 (0.01278)**	0.01658 (0.02650)	0,00530 (0.01024)	0,00674 (0.01416)	0.01304 (0.02536)	0.01667 (0.02626)	0.01281 (0.02276)	0.01236 (0.02191)	0,00741 (0.01588)	0,01058 (0.01875)
bike_sharing	0.00074 (0.00223)	**0.00000 (0.00000)**	0,00000 (0,00001)	0,00001 (0.00003)	0.00080 (0.00120)	0.00243 (0.00376)	0.00290 (0.00302)	0.00017 (0.00050)	0,00631 (0.01441)	0,00620 (0.01583)
cocomo81	0.01811 (0.08708)	0.02371 (0.07686)	**0,01055 (0.04551)**	0,01154 (0.04551)	0.01982 (0.09646)	0.02353 (0.09495)	0.01817 (0.07144)	0.01632 (0.07401)	0,02058 (0.08630)	0,02179 (09996)
concrete	0.00665 (0.01695)	0.01725 (0.02493)	**0,00550 (0.01078)**	0,00566 (0.01085)	0.01192 (0.01970)	0.01743 (0.02479)	0.01215 (0.01872)	0.01014 (0.01699)	0,013889 (0.02755)	0,01379 (0.02345)
desharnais	0.02823 (0.08256)	**0.01923 (0.04115)**	0,06771 (0.24527)	0,07628 (0.30161)	0.02434 (0.05996)	0.01969 (0.04500)	0.01942 (0.05176)	0.02354 (0.06361)	0,02722 (0.06749)	0,02601 (0.06441)
energy_efficiency	**0.00026 (0.00059)**	0.00628 (0.01101)	0,00042 (0.00084)	0,00067 (0.00107)	0.00527 (0.00995)	0.00679 (0.01150)	0.00639 (0.00825)	0.00471 (0.00883)	0,00589 (0.01275)	0,00595 (0.01203)
housing	0.01169 (0.04120)	0.01149 (0.03045)	0,00949 (0.11494)	0,00737 (0.04356)	0.01057 (0.03735)	0.01358 (0.04061)	0.00993 (0.02903)	**0.00707 (0.02192)**	0,01069 (0.03118)	0,01220 (0.03662)
nasa93	0.02349 (0.11590)	0.01273 (0.04840)	0,01107 (0.03209)	**0,01002 (0.02441)**	0.01262 (0.05869)	0.01615 (0.06381)	0.01426 (0.04901)	0.01309 (0.05152)	0,01417 (0.05521)	0,01405 (0.06580)
wineq-red	0.02547 (0.04542)	0.01724 (0.02875)	0,01855 (0.006925)	0,01704 (0.03008)	0.01695 (0.02838)	0.01769 (0.02954)	0.01758 (0.02955)	**0.01661 (0.02728)**	0,01982 (0.03505)	0,01889 (0.03298)

Source: Prepared by the authors

frequent inclusion in heterogeneous ensemble compositions.

These observations reinforce the central hypothesis of this work: different statistical data profiles require different modeling strategies. In addition to identifying the models with the lowest average error per dataset, the analysis also highlighted the consistent performance of certain algorithms — such as `FANN-1`, `FANN-2`, `CART`, and `RBF` — across multiple contexts. Based on these findings, the three most promising models were used to build the **heterogeneous ensembles**. This configuration will be compared, in the next subsection, with the homogeneous ensembles and the strategic recommendations generated by the proposed framework.

*b) **Performance of Ensemble Strategies**:* Table VI presents a direct comparison among three ensemble approaches: (i) *homogeneous* — using the best individual model; (ii) *heterogeneous* — average prediction of the three best models; and (iii) *recommended strategy* — defined based on the statistical profile of each dataset. The results include the means and standard deviations for MSE, MAE, and R^2, obtained from 20 runs using 10-fold cross-validation.

The best performances for each metric are highlighted in the table. The "Recom." column indicates which strategy was selected by the proposed framework, allowing for the assessment of its effectiveness in guiding the choice of the most appropriate ensemble technique.

Performance on Homogeneous Datasets. For the three datasets identified as homogeneous — *airfoil self noise*, *bike sharing*, and *energy efficiency* — the framework correctly recommended the homogeneous strategy, which indeed yielded the best results across all metrics. The accuracy rate in this category was **100%**, reinforcing that in domains with low variability and linear structure, well-tuned individual models are more effective than diverse combinations.

Performance on Heterogeneous Datasets. Among the seven datasets classified as heterogeneous, the framework provided correct recommendations in **5 cases** (*abalone*, *cocomo81*, *concrete*, *housing*, and *wineq-red*). In these scenarios, the heterogeneous strategy consistently outperformed in terms of MSE and MAE and also showed gains in R^2. However, in *desharnais* and *nasa93*, the homogeneous ensemble achieved slightly better performance. In both cases, negative R^2 values and high standard deviations indicate elevated model instability — likely caused by high data dispersion and low sample density, which may compromise the expected advantage of heterogeneous strategies.

Summary of Results. The proposed framework provided correct recommendations for **8 out of 10 datasets**, reaching an overall accuracy of **80%**. In the two remaining cases, the difference between strategies was small, indicating that even when the choice was not optimal, it remained within an acceptable performance margin. Furthermore, the observed standard deviation values suggest that the recommended strategies tend to yield more stable results in most scenarios.

These findings reinforce the central hypothesis of this study: **ensemble strategies should be adapted to the statistical structure of the data**. The proposed framework offers a metric-driven, less empirical approach to this decision, promoting performance gains and enhanced robustness in regression tasks.

These findings reinforce the main hypothesis of this work: different statistical data profiles require different modeling strategies. The proposed framework provides a systematic means to guide this decision, reducing the empirical nature of the choice and promoting better performance in

TABLE VI
PERFORMANCE COMPARISON BY DATASET PROFILE AND ENSEMBLE STRATEGY

Dataset	Perfil Proposto	MSE			MAE		R2	
		Homog.	Heterog.	Recom.	Homog.	Heterog.	Homog.	Heterog.
abalone	Heterogêneo	0.00572 (0.0090)	0.00565 (0.0080)	**0.00565** (Heterog.)	0.05343 (0.00334)	0.05343 (0.00310)	0.56655 (0.03458)	0.57099 (0.03161)
airfoil self noise	Homogêneo	0.00383 (0.00086)	0.00400 (0.00066)	**0.00383** (Homog.)	0.04460 (0.00405)	0.04601 (0.00320)	0.88412 (0.03114)	0.87908 (0.04970)
bike sharing	Homogêneo	0.00000 (0.00000)	0.00002 (0.00002)	**0.00000** (Homog.)	0.00000 (0.00000)	0.00000 (0.00112)	1.00000 (0.00000)	0.99996 (0.00002)
cocomo81	Heterogêneo	0.01362 (0.03086)	0.01333 (0.03042)	**0.01333** (Heterog.)	0.05218 (0.03764)	0.05010 (0.04062)	-2.11150 (2.88569)	-4.46123 (9.90601)
concrete	Heterogêneo	0.00486 (0.00111)	0.00428 (0.00080)	**0.00428** (Heterog.)	0.05176 (0.00447)	0.04715 (0.00346)	0.88608 (0.02548)	0.90016 (0.01479)
desharnais	Heterogêneo	0.01809 (0.01430)	0.01846 (0.01409)	**0.01809** (Homog.)	0.09428 (0.03532)	0.09120 (0.03532)	-0.14582 (1.86408)	-0.12852 (1.85075)
energy efficiency	Homogêneo	0.00020 (0.00004)	0.00028 (0.00005)	**0.00020** (Homog.)	0.00947 (0.00092)	0.01295 (0.00144)	0.99723 (0.00050)	0.99611 (0.00112)
housing	Heterogêneo	0.00735 (0.0261)	0.00534 (0.0291)	**0.00534** (Heterog.)	0.05799 (0.00685)	0.04781 (0.00577)	0.82238 (0.05290)	0.86805 (0.07251)
nasa93	Heterogêneo	0.01212 (0.01589)	0.01266 (0.01883)	**0.01212** (Homog.)	0.05076 (0.02203)	0.05518 (0.03046)	-1.43123 (3.73155)	-1.37458 (3.52228)
wineq-red	Heterogêneo	0.01660 (0.00265)	0.01629 (0.00246)	**0.01629** (Heterog.)	0.10035 (0.00782)	0.09907 (0.00772)	0.35530 (0.08641)	0.36675 (0.08376)

Source: Prepared by the authors

regression problems.

To assess whether the performance differences between homogeneous and heterogeneous strategies are statistically significant, the paired non-parametric Wilcoxon test was applied with a 5% significance level.

*c) **Statistical Hypotheses**:* Since the objective is to validate whether the recommendation was correct, the following hypotheses were formulated:

- **For datasets with a homogeneous profile:**
  - **Null hypothesis (H_0):** The heterogeneous strategy has equal or lower average error compared to the homogeneous strategy.
    $H_0 : \mu_{\text{Heterogeneous}} \leq \mu_{\text{Homogeneous}}$
  - **Alternative hypothesis (H_1):** The homogeneous strategy has a significantly lower average error.
    $H_1 : \mu_{\text{Homogeneous}} < \mu_{\text{Heterogeneous}}$
- **For datasets with a heterogeneous profile:**
  - **Null hypothesis (H_0):** The homogeneous strategy has equal or lower average error compared to the heterogeneous strategy.
    $H_0 : \mu_{\text{Homogeneous}} \leq \mu_{\text{Heterogeneous}}$
  - **Alternative hypothesis (H_1):** The heterogeneous strategy has a significantly lower average error.
    $H_1 : \mu_{\text{Heterogeneous}} < \mu_{\text{Homogeneous}}$

For the analysis, we considered the results from 20 executions using 10-fold cross-validation, which generated a robust set of Mean Squared Error (MSE) values for each strategy and dataset.

D. Wilcoxon Test Results

To statistically validate the differences between the homogeneous and heterogeneous ensemble strategies, paired Wilcoxon tests were performed at a 5% significance level. The tests were conducted separately according to the dataset classification, as shown in Table VII and Table VIII.

The obtained p-values indicate the statistical significance of the superiority of the recommended technique. Values below $\alpha = 0.05$ indicate that the recommendation provided by the proposed framework resulted in a statistically superior performance compared to the alternative strategy.

a) Homogeneous Datasets

TABLE VII
WILCOXON TEST FOR HOMOGENEOUS DATASETS (MSE)

Dataset	Compared Strategies	p-value
Airfoil Self Noise	Homog. vs Heterog.	0.037
Bike Sharing	Homog. vs Heterog.	0.002
Energy Efficiency	Homog. vs Heterog.	0.000

b) Heterogeneous Datasets

TABLE VIII
WILCOXON TEST FOR HETEROGENEOUS DATASETS (MSE)

Dataset	Compared Strategies	p-value
Abalone	Heterog. vs Homog.	0.002
Cocomo81	Heterog. vs Homog.	0.003
Concrete	Heterog. vs Homog.	0.000
Desharnais	Heterog. vs Homog.	0.207
Housing	Heterog. vs Homog.	0.004
NASA93	Heterog. vs Homog.	0.109
Wine Quality (Red)	Heterog. vs Homog.	0.043

Table VII presents the results of the Wilcoxon tests applied to the datasets classified as homogeneous. For all cases, the p-values were below the 0.05 significance threshold, indicating statistical evidence that the homogeneous strategy significantly outperformed the heterogeneous one in terms of Mean Squared Error (MSE). These results reinforce the hypothesis that, in scenarios with low variability and predictable structure, homogeneous ensembles offer greater stability and predictive accuracy.

Conversely, Table VIII presents the Wilcoxon tests for the heterogeneous datasets. In this category, the heterogeneous strategy showed statistically superior performance ($p < 0.05$) in 5 out of the 7 analyzed datasets: *abalone*, *cocomo81*, *concrete*, *housing*, and *wine quality (red)*. For the remaining two — *desharnais* and *nasa93* — the p-values (0.207 and 0.109, respectively) do not indicate a statistically significant difference, which may be attributed to the presence of simpler local patterns or data sparsity in complex regions of the input space.

These findings strengthen the central premise of this work: **the statistical profile of the data directly influences the choice of the most appropriate ensemble strategy**. The use of the Wilcoxon test validated that a recommendation based on prior statistical analysis of the data can lead to more effective and evidence-based decisions, outperforming fixed and empirical approaches in predictive modeling.

V. Conclusion

This work proposed a statistical framework to recommend ensemble strategies for regression tasks, based on the structural characteristics of the data. By integrating metrics such as coefficient of variation, skewness, kurtosis, N3 Score, and others, the framework classified datasets as homogeneous or heterogeneous and recommended the corresponding ensemble strategy accordingly.

Experimental results on 10 real-world datasets demonstrated that the framework achieved 100% accuracy for homogeneous datasets and 71% for heterogeneous ones. Wilcoxon statistical tests confirmed that in cases with significant differences, the framework's recommendation led to statistically better or equivalent results, reinforcing its robustness and applicability.

From a practical perspective, the proposed framework contributes to more assertive regression tasks by providing a data-driven method for selecting appropriate model combinations. This reduces the reliance on empirical guesswork and facilitates the effective implementation of regression algorithms within machine learning systems.

Future directions include:

- Refining the dataset classification criteria using meta-learning techniques;
- Investigating the influence of normalization and preprocessing techniques on strategic recommendations;
- Exploring more sophisticated ensemble strategies (e.g., stacking, blending, dynamic selection) to improve adaptability across different dataset profiles.

In summary, this study takes a relevant step toward adaptive and intelligent frameworks for automated model selection in regression, promoting improved performance, interpretability, and scalability.

References

[1] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, vol. 1857, pp. 1–15, 2000.

[2] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[3] D. W. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[4] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 1–39, 2010.

[5] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," *Information Fusion*, vol. 16, pp. 3–17, 2014.

[6] R. M. O. Cruz, D. V. Oliveira, G. D. C. Cavalcanti, and R. Sabourin, "Dynamic classifier selection: Recent advances and perspectives," *Information Fusion*, vol. 41, pp. 195–216, 2018.

[7] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 289–300, 2002.

[8] S. Garcia, A. Fernandez, J. Luengo, and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability," *Soft Computing*, vol. 13, no. 10, pp. 959–977, 2009.

[9] M. Reif, F. Shafait, and A. Dengel, "Meta-learning for evolutionary parameter optimization of classifiers," *Machine Learning*, vol. 87, no. 3, pp. 357–380, 2012.

[10] B. Bilalli, A. Abelló, T. Aluja-Banet, and R. Wrembel, "Predictive analytics with regression and classification: A survey of algorithm selection techniques," *ACM Computing Surveys*, vol. 51, no. 1, pp. 1–35, 2017.

[11] E. Alcobaça, M. Kalinowski, and E. Mendes, "Meta-feature analysis for selecting ensembles in software effort estimation," *Journal of Systems and Software*, vol. 181, 111041, 2021.

[12] A. Tsymbal, "The problem of concept drift: Definitions and related work," Technical Report TCD-CS-2004-15, Department of Computer Science, Trinity College Dublin, 2004.

[13] B. da Silva *et al.*, "Surrogate Model and Multi-objective Evolutionary Algorithm Applied to Automotive Stamping," in *Anais do XVI Congresso Brasileiro de Inteligência Computacional (CBIC)*, Salvador, 2023.

[14] L. T. Macedo *et al.*, "Avaliação de Diferentes Técnicas de Pré-Processamento para um Sistema Neural de Apoio à Decisão em Inspeções por Ultrassom," in *Anais do XVI Congresso Brasileiro de Inteligência Computacional (CBIC)*, Salvador, 2023.

[15] F. Hutter, L. Kotthoff, and J. Vanschoren, "Automated Machine Learning: Methods," *Systems, Challenges, Springer*, 2019.

[16] M. Feurer et al., "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 2962–2970, 2015.

[17] M. Feurer et al., "Auto-sklearn 2.0: Hands-free AutoML via Meta-Learning," in *The Journal of Machine Learning Research*, vol. 23, , Issue 1 Article No.: 261, pp. 11936 - 11996, 2022.

[18] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[19] T. Woloszynski, M. Kurzynski, and R. M. O. Cruz, "A measure of competence based on random classification for dynamic ensemble selection," *Information Fusion*, vol. 13, no. 3, pp. 207–213, 2012.

[20] M. Kelly, R. Longjohn, e K. Nottingham, "The UCI Machine Learning Repository," 2025. Disponível em: https://archive.ics.uci.edu/datasets. Acesso em: fev. 2025.

[21] PROMISE, "The PROMISE Repository of Software Engineering Databases," *School of Information Technology and Engineering, University of Ottawa, Canada*, 2005. Disponível em: http://promise.site.uottawa.ca/SERepository. Acesso em: 24 out. 2020.

[22] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.