Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5, 2750-2764
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i5.7589

Multi-K KNN regression with bootstrap aggregation: Accurate predictions and alternative prediction intervals

Soraida Sriprasert[1], Patchanok Srisuradetchai[1*]
[1]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand; patchanok@mathstat.sci.tu.ac.th (P.S.).

Abstract: The k-nearest neighbors (KNN) algorithm is widely recognized for its simplicity and flexibility in modeling complex, non-linear relationships; however, standard KNN regression does not inherently provide prediction intervals (PIs), presenting a persistent challenge for uncertainty quantification. This study introduces a bootstrap-based multi-K approach specifically designed to construct robust prediction intervals in KNN regression. By systematically aggregating predictions across multiple neighborhood sizes through ensemble techniques and bootstrap resampling, the method effectively quantifies prediction uncertainty, particularly in challenging high-dimensional scenarios. Evaluations conducted on 15 diverse datasets spanning education, healthcare, chemistry, economics, and social sciences reveal that the proposed approach consistently achieves competitive predictive accuracy compared to traditional regression methods. Although traditional regression produces wider intervals with higher coverage probabilities, the proposed bootstrap-based KNN method generates notably tighter intervals, enhancing interpretability and practical utility. Despite occasionally reduced coverage probabilities, especially in high-dimensional contexts, the proposed methodology effectively balances precision and predictive coverage. Practically, this multi-K bootstrap approach provides researchers and practitioners with an effective and interpretable method for robust uncertainty quantification in complex predictive modeling tasks.

Keywords: *Bootstrapping, Ensemble methods, K-nearest neighbors, KNN regression, Machine learning, Non-parametric regression, Prediction intervals, Uncertainty quantification.*

1. Introduction

The k-nearest neighbors (KNN) algorithm is widely recognized for its simplicity, flexibility, and effectiveness in modeling complex, nonlinear relationships across diverse fields such as finance, environmental monitoring, medical diagnostics, and engineering. Its intuitive nature and minimal assumptions make KNN particularly valuable for real-world predictive tasks requiring both accuracy and interpretability.

In the financial domain, KNN has demonstrated substantial potential in tasks such as stock price prediction and market analysis Easton, et al. [1]. For example, Alkhatib, et al. [2] utilized KNN to forecast stock prices for companies listed on the Jordanian stock exchange, producing predictions closely aligned with actual market values. Additionally, Kim [3] compared KNN against support vector machines (SVM) and other machine learning (ML) techniques for financial time series forecasting, highlighting the method's competitive performance and robustness in modeling intricate financial behaviors.

Beyond finance, KNN has shown significant utility in environmental monitoring and remote sensing Chao and Guang Qiu [4]. Thanh Noi and Kappas [5] compared KNN with random forests (RF) and SVMs for land cover classification using Sentinel-2 satellite imagery, demonstrating KNN's strong performance. KNN has also effectively contributed to air quality prediction and monitoring [6,

History: Received: 24 March 2025; Revised: 28 April 2025; Accepted: 1 April 2025; Published: 27 May 2025
* Correspondence: patchanok@mathstat.sci.tu.ac.th

7] and has proven adaptable in forecasting COVID-19 cases Sejuti and Islam [8] thus underscoring its applicability to pressing public health and environmental issues.

In engineering, KNN has been successfully applied to various predictive tasks. For instance, Al-Dosary, et al. [9] demonstrated the effectiveness of KNN in predicting fuel consumption for tractor-chisel plow systems, achieving superior outcomes compared to traditional multiple linear regression (MLR) models. Zhang, et al. [10] introduced a hybrid metric-based KNN approach incorporating both distance and direction information, enhancing forward gait stability control in bipedal robots and effectively identifying faults associated with abnormal motion. Similarly, Jegorowa, et al. [11] developed a non-invasive, automated system employing KNN to evaluate drill-bit conditions based on operational data such as feed force, cutting torque, jig vibrations, acoustic emissions, and noise.

The medical field has also benefited significantly from KNN, especially in automated diagnostic systems. Onan [12] proposed a fuzzy-rough nearest neighbor classifier enhanced with consistency-based subset evaluation and instance selection, resulting in improved accuracy for breast cancer diagnosis. Shouman, et al. [13] applied KNN to diagnose heart disease, demonstrating the method's reliability and effectiveness. Liu, et al. [14] employed KNN for predicting early recurrence in hepatocellular carcinoma patients, illustrating its strength in oncology. Additionally, Archana and Komarasamy [15] introduced an innovative ensemble method combining bagging techniques with KNN, significantly enhancing brain malignancy detection accuracy.

A comprehensive review by Halder, et al. [16] summarized recent advancements in KNN methodologies, highlighting the algorithm's increasing adaptability to various predictive modeling challenges. Notable developments include feature randomization and bootstrapping for kernel KNN proposed by Srisuradetchai and Suksrikran [17] adaptive soft KNN classifiers capable of estimating posterior probabilities introduced by Bermejo and Cabestany [18] and hybrid approaches combining K-means clustering and KNN classification by Deng, et al. [19]. Further methodological innovations encompass dimensionality reduction algorithms such as Q-SNE by Ingram and Munzner [20] structured data processing strategies addressing big data challenges by Pramanik, et al. [21] efficient data partitioning methods proposed by Saadatfar, et al. [22] NCP-KNN techniques introduced by Abdalla and Amer [23] to reduce computational complexity, and advanced exact KNN queries for high-dimensional data discussed by Ukey, et al. [24].

Prediction intervals significantly enhance the practical utility of KNN regression by quantifying predictive uncertainty, thereby providing intervals within which true outcomes likely reside, rather than offering single point estimates [25]. Such intervals are particularly crucial in decision-making scenarios, including aircraft trajectory prediction or parking availability forecasting, where understanding outcome variability is essential. Additionally, prediction intervals facilitate more accurate assessments of model reliability, especially in scenarios involving heteroscedastic data or complex nonlinear relationships [26]. These intervals are especially important in sectors like agriculture and healthcare, where managing uncertainty induced by climate variability is paramount. For instance, Srisuradetchai [27] demonstrated effective interval forecasting for time-series KNN regression.

Despite the widespread adoption and continuous advancements in KNN regression methods, reliably constructing PIs remains challenging. This study addresses this critical gap by integrating ensemble and bootstrap techniques specifically tailored to KNN regression. To our knowledge, explicitly incorporating multiple neighborhood sizes (multi-K) within a KNN regression framework to systematically construct robust bootstrap-based PIs has not been previously explored.

Figure 1 illustrates the regression surfaces generated by KNN regression applied to a synthetic dataset with two predictors, x_1 and x_2, and a response variable modeled as $y = 2\sin(x_1) + 0.5x_2 + \varepsilon$, where $\varepsilon \sim N(0,1)$. The visualizations clearly demonstrate the influence of the number of neighbors (k) on the regression outcome. With smaller values of k (e.g., k = 3), the regression surface captures detailed local variations, potentially increasing prediction variance. Conversely, larger k (e.g., k = 15) yields smoother surfaces, effectively reducing variance but at the expense of increased bias. This trade-

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

off highlights the complexity inherent in selecting an optimal k, emphasizing the advantage of adopting a multi- k ensemble approach. These patterns validate the rationale for employing bootstrap-based prediction intervals, as combining predictions from multiple k-values can effectively balance predictive accuracy, interpretability, and uncertainty quantification.

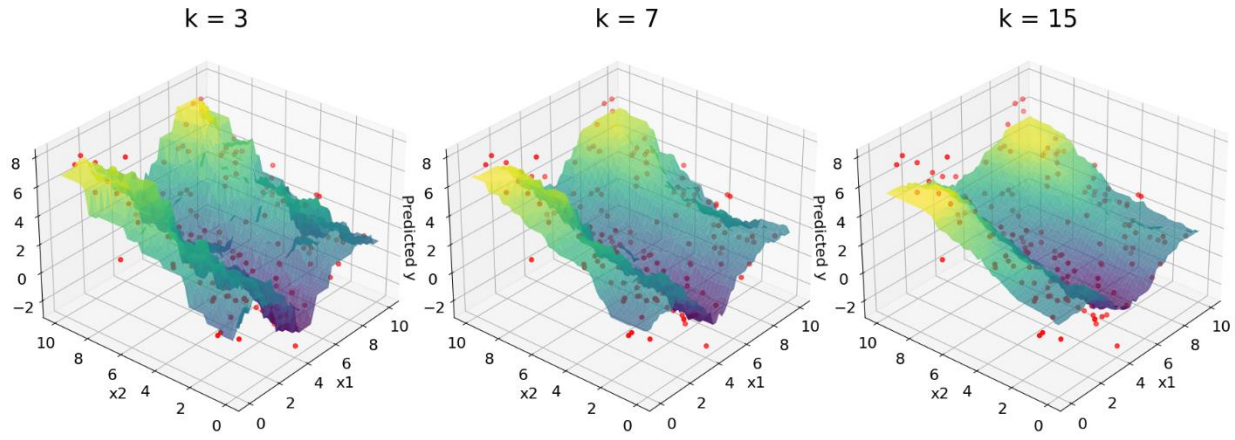

Figure 1.
KNN regression surfaces for synthetic data at different neighborhood sizes.

2. K-Nearest Neighbors (KNN) Regression

KNN regression is a non-parametric, data-driven algorithm extensively used for predicting continuous variables. Unlike parametric models that assume predefined functional relationships, KNN relies solely on similarity among data points, generating predictions by averaging the target values of the K nearest neighbors within the training dataset [17]. This straightforward approach, combined with its ability to effectively model complex, nonlinear relationships, has made KNN popular across diverse fields such as finance, healthcare, and environmental science.

The theoretical foundation of KNN regression is succinctly captured by the formula:

$$\hat{y}(x) = \frac{1}{K}\sum_{i=1}^{K} y_i ,$$

where y_i denotes the target values corresponding to the K nearest neighbors of the query point x [28]. Selecting an optimal K significantly impacts the algorithm's performance by affecting the bias-variance trade-off. Smaller K values capture intricate local patterns, reducing bias but increasing variance and sensitivity to noise. Conversely, larger K smoothes variations, increasing bias but effectively reducing variance. Therefore, selecting an optimal K is critical for robust predictive performance.

The accuracy of KNN regression also heavily depends on the choice of distance metric utilized to determine the proximity of points. Commonly employed metrics include the Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2} ,$$

the Manhattan distance:

$$d(x, x_i) = \sum_{j=1}^{p} \left| x_j - x_{ij} \right| ,$$

and the Minkowski distance:

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

$$d(x, x_i) = \left(\sum_{j=1}^{p} \left| x_j - x_{ij} \right|^q \right)^{1/q},$$

where p represents the number of features, and q determines the type of distance measure ($q = 2$ corresponds to Euclidean distance, while $q = 1$ corresponds to Manhattan distance) [29].

Recent advancements in KNN regression focus on enhancing its adaptability and predictive power. Adaptive KNN dynamically adjusts the neighborhood size K based on local data characteristics, optimizing predictive performance for individual query points. Weighted KNN assigns greater influence to closer neighbors by applying weights based on distance from the query point. The weighted prediction formula is:

$$\hat{y}(x) = \frac{\sum_{i=1}^{K} w_i y_i}{\sum_{i=1}^{K} w_i},$$

where the weights w_i are typically defined as $1/\left(d(x, x_i)\right)^2$. Kernel KNN further enhances predictive accuracy by using kernel functions to weight neighbor contributions, enabling improved modeling of non-linear relationships. Local linear regression integrates linear modeling within the neighborhood defined by the K nearest neighbors, thus improving local prediction accuracy. Additionally, random KNN methods introduce randomness in feature selection and distance calculations to enhance robustness and computational efficiency. Ensemble KNN combines predictions from multiple KNN models, each employing different values of K or distance metrics, thereby creating a more stable and reliable predictive model [30].

Theoretical guarantees underpin the reliability of KNN regression under specific conditions. Specifically, as the sample size n grows toward infinity and the neighborhood size K increases in such a manner that the ratio K/n → 0, the KNN estimator converges to the true regression function. Under these conditions, the asymptotic mean squared error (MSE) for KNN regression at a query point xxx can be decomposed into its bias and variance components as follows:

$$MSE(x) \approx \left(\frac{K}{n} \right)^{4/d} B(x) + \frac{1}{K} V(x),$$

where B(x) denotes the squared bias, V(x) represents the variance at point x, and d is the dimensionality of the feature space. This decomposition highlights the critical balance between bias and variance. A smaller K results in higher variance but lower bias, whereas a larger value of K produces lower variance but higher bias. Thus, careful selection of the neighborhood size K relative to the sample size n is essential for optimal predictive accuracy [31, 32].

3. Cross-Validation Error

Cross-validation is a widely adopted technique used to assess the predictive performance of statistical and machine learning models. Among various methods, V-fold cross-validation is particularly popular due to its optimal balance between computational efficiency and reliable error estimation.

In $V-$fold cross-validation, the dataset $\mathbf{S} = \left\{ \left(\mathbf{x}_i, y_i \right) \right\}_{i=1}^{n}$ is divided into V equally sized subsets known as folds. Each fold sequentially serves as a test set, while the remaining $V-1$ folds collectively form the training set. This ensures each data point contributes to both training and testing, enhancing

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

the robustness and generalizability of evaluation. Formally, the dataset $\mathbf{S}$ is partitioned into V subsets, Fold_v for $v = 1, 2, \ldots, V$, satisfying:

$$\bigcup_{v=1}^{V} \text{Fold}_v = S \text{ and } \text{Fold}_i \cap \text{Fold}_j = \varnothing, \quad \forall i \neq j$$

In each iteration v, the model is trained on $S \setminus \text{Fold}_v$, the complement of the v-th fold, producing surrogate model $\hat{\psi}_v(\mathbf{x}_i)$. The predictive error e_v for each fold is calculated as:

$$e_v = \frac{1}{|\text{Fold}_v|} \sum_{(\mathbf{x}_i, y_i) \in \text{Fold}_v} \ell(y_i, \hat{\psi}_v(\mathbf{x}_i)),$$

where ℓ denotes an appropriate loss function such as mean squared error (MSE). After iterating over all folds, overall CV error $\hat{\text{o}}_{\text{CV}}^{K\text{-fold}}$ is computed as the average of fold-specific errors:

$$\hat{\text{o}}_{\text{CV}}^{K\text{-fold}} = \frac{1}{V} \sum_{v=1}^{V} e_v .$$

This process ensures comprehensive error estimates, highlighting the necessity for meticulous design and validation when applying V-fold CV, particularly in complex or heterogeneous datasets [33-35].

However, the implementation of V-fold cross-validation can face several challenges. One key issue is class imbalance, where certain classes might be underrepresented or entirely excluded from some folds. For instance, if Fold_k contains only data from one class, the surrogate model $\hat{\psi}_v(\mathbf{x})$ may fail to generalize to unseen classes. To mitigate this, it is common practice to shuffle the data before partitioning it into folds. Shuffling assumes that the samples are independent and identically distributed (i.i.d.), allowing for a more balanced representation of classes across folds. Nevertheless, this assumption may not hold for time-series or correlated data, where shuffling could result in overly optimistic error estimates due to the violation of data dependencies [36-38].

Another critical consideration is ensuring representative data distribution across folds. If certain regions of the feature space are inadequately represented in training folds, the model may fail to generalize effectively when predicting the test fold. These concerns highlight the importance of meticulous design and validation when applying V-fold cross-validation, particularly in complex or heterogeneous datasets.

4. Performance Criteria

Evaluating predictive models typically involves metrics that fall into two broad categories: point estimation metrics and prediction interval metrics. These metrics provide a comprehensive assessment of a model's accuracy, reliability, and uncertainty quantification capabilities.

4.1. Point Estimation Metrics

4.1.1. Root Mean Squared Error

Root mean squared error (RMSE) is widely utilized for measuring prediction error magnitude. Mathematically, RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} ,$$

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

where y_i represents the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. RMSE emphasizes large prediction errors, making it particularly useful in applications where minimizing substantial deviations is crucial [38].

4.1.2. Mean Absolute Error

Mean absolute error (MAE) measures the average magnitude of prediction errors, providing an intuitive metric for predictive performance. MAE is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|,$$

Compared to RMSE, MAE is less sensitive to extreme values, making it robust to outliers and suitable when reducing large deviations is not critical [39].

4.1.3. R-Squared

The coefficient of determination (R^2) quantifies the proportion of variability in the dependent variable explained by the predictive model. Mathematically, it is expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of the observed values. The R^2 metric ranges between negative infinity and 1, where values closer to 1 indicate better predictive performance [40, 41].

4.2. Prediction Interval Metrics

4.2.1. Coverage Probability

Coverage probability (CP) evaluates the proportion of actual observed values that fall within predicted intervals. Ideally, for a nominal confidence level $(1-\alpha)100\%$, empirical CP should closely match this nominal level, thus reflecting the interval's reliability. CP serves as a critical metric for assessing the accuracy of interval estimates in various statistical contexts, including zero-inflated models and regression frameworks [42-44]. It is calculated as:

$$\mathrm{CP} = \frac{1}{n}\sum_{i=1}^{n} I(L_i \le y_i \le U_i),$$

where L_i and U_i denote the lower and upper bounds of the prediction interval, respectively, and $I(\cdot)$ is an indicator function.

4.2.2. Expected Length

Expected length (EL) measures the average width of prediction intervals, calculated as:

$$\mathrm{EL} = \frac{1}{n}\sum_{i=1}^{n}(U_i - L_i).$$

Generally, narrower intervals are preferred, provided they maintain an appropriate CP consistent with the desired confidence level. Balancing interval width and CP is essential for effective uncertainty quantification [45, 46].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

5. Proposed Methodology

This study introduces a novel methodology termed multi-K KNN regression, explicitly designed to enhance prediction intervals (PIs) in K-nearest neighbor (KNN) regression. Although ensemble methods, bootstrapping, and cross-validation (CV) are individually well-established, their integrated application—specifically tailored to leverage multiple neighborhood sizes (multi-K)—represents a significant methodological advancement in KNN regression. By combining predictions from various neighborhood sizes with V-fold CV and bootstrap resampling, the proposed framework generates robust PIs, accurately quantifying prediction uncertainty and variability. The methodology comprises three primary stages: data preprocessing, multi-K prediction generation, and bootstrap-based PI construction.

5.1. Data Preprocessing

Consider a dataset S containing n observations:

$$S=\{(\mathbf{x}_i,y_i)\}_{i=1}^{n},\quad \mathbf{x}_i\in\mathbb{R}^{d},y_i\in\mathbb{R},$$

where $\mathbf{x}_i$ are input features and y_i are target values. The dataset is partitioned into V -fold cross-validation sets $\{F_1,F_2,\ldots,F_V\}$. In iteration v, fold F_V serves as the validation set, and the remaining $V-1$ folds form the training set $S_{\mathrm{tr},v}$. Metrics are subsequently aggregated across folds to ensure robustness.

For consistency and comparability across features, standardization is performed:

$$x'_{i,j}=\frac{x_{i,j}-\mu_j}{\sigma_j},$$

where μ_j and σ_j denote the mean and standard deviation computed from the training set, respectively.

5.2. Multi-K Prediction Generation

KNN regression predictions are iteratively computed for multiple neighborhood sizes, defined as:

$$K=\left\{1,2,\ldots,\left\lfloor\sqrt{n}\right\rfloor\right\}.$$

For each validation fold F_v and test point $\mathbf{x}\in F_v$, predictions for each neighborhood size $k\in K$ are obtained via:

$$\hat{y}_k(\mathbf{x})=\frac{1}{k}\sum_{i=1}^{k}y_{(i)},$$

where $y_{(i)}$ represents the target values of the k nearest neighbors identified in $S_{\mathrm{tr},v}$. Predictions for all k are collected into the prediction vector:

$$\mathbf{P}_{\mathbf{x}}=\{\hat{y}_1(\mathbf{x}),\hat{y}_2(\mathbf{x}),\ldots,\hat{y}_{\left\lfloor\sqrt{n}\right\rfloor}(\mathbf{x})\},$$

thereby promoting diversity and enhancing predictive robustness.

5.3. Bootstrap Interval Construction

Bootstrap resampling generates prediction intervals from $\mathbf{P}_{\mathbf{x}}$. For each test point $\mathbf{x}\in F_k$:

1. Bootstrap Sampling: Generate B bootstrap samples from $\mathbf{P}_{\mathbf{x}}$ with replacement. Each bootstrap sample is denoted as:

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

$\mathbf{P}_{\mathbf{x}}^{(b)}$, $b = 1, 2, ..., B$.

2. Bootstrap Mean: Compute the mean prediction for each bootstrap sample:

$$\mu_{\mathbf{x}}^{(b)} = \frac{1}{|\mathbf{P}_{\mathbf{x}}^{(b)}|} \sum_{j \in \mathbf{P}_{\mathbf{x}}^{(b)}} y_j .$$

3. Quantile Estimation: Estimate the $\alpha/2$-th and $(1-\alpha/2)$-th quantiles of the bootstrap means:

$$L_{\mathbf{x}} = Q_{\alpha/2}\left(\{\mu_{\mathbf{x}}^{(b)}\}_{b=1}^{B}\right), \quad U_{\mathbf{x}} = Q_{1-\alpha/2}\left(\{\mu_{\mathbf{x}}^{(b)}\}_{b=1}^{B}\right).$$

In this paper, we use $\alpha = 0.05$.

4. Prediction Interval: Construct the PI for $\mathbf{x}$:

$$\mathrm{PI}_{\mathbf{x}} = [L_{\mathbf{x}}, U_{\mathbf{x}}] .$$

The complete procedure is succinctly summarized in the following algorithm:
Input:

- Dataset: $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Number of CV folds: V
- Neighborhood sizes: $K = \left\{1, 2, \ldots, \left\lfloor \sqrt{n} \right\rfloor\right\}$
- Number of bootstrap samples: B
- Confidence level: $(1-\alpha)$

Procedure:

1. Partition the dataset into V folds.
2. For each fold F_v:

- Standardize features based on the training set.
- For each test point $\mathbf{x} \in F_v$, compute predictions using KNN for all $k \in K$.
- Perform bootstrap resampling on predictions to construct PIs.

Output: Prediction intervals $[L_{\mathbf{x}}, U_{\mathbf{x}}]$ for each test point.

6. Evaluation Datasets and Results

This section presents the datasets used for benchmarking and evaluates the proposed bootstrap multi-K KNN regression approach. Results are compared with those obtained from standard regression methods. Both point prediction accuracy and interval estimation effectiveness are assessed using the metrics RMSE, MAE, MAPE, R^2, CP, and EL.

6.1. Datasets for Benchmarking

To ensure rigorous evaluation, 15 datasets spanning various fields—including education, healthcare, chemistry, economics, and social sciences—were selected. These datasets substantially differ in sample size (n), feature dimensionality (p), and response variables, thus enabling comprehensive validation of the proposed methodology. Table 1 summarizes these structured datasets, which range from low-dimensional to high-dimensional settings, covering simple and complex predictive tasks.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

Table 1.
Summary of datasets used for benchmarking.

Datasets	n	p	**Responses**
Student Performance (MATH) Cortez [47]	395	42	Math Scores
Student Performance (POR) Cortez [47]	649	30	Portuguese Scores
Wisconsin Prognostic Breast Cancer (CANCER) Wolberg, et al. [48]	198	34	Recurrence Time
Polycyclic Aromatic Hydrocarbon (PAH)Todeschini, et al. [49]	80	113	Effectiveness of PDGFH inhibitors
Platelet-Derived Growth Factor Receptor (PDFGR) Guha and Jurs [50]	79	304	IC50 Biological Activity
Boston House (BOSTON) Harrison Jr and Rubinfeld [51]	506	16	Median House Value ($1000's)
Phenethyl (PHEN) Kubinyi [52]	22	629	Phenethyl Derivatives
TECATOR Borggaard and Thodberg [53]	240	125	Fat Content of Meat Sample
FRIC4 Friedman [54]	1,000	100	Artificially Generated Model Responses
Happiness Rank (HAPPY) Helliwell, et al. [55]	235	9	Happiness Scores (World Happiness Report)
Auto Horse (AUTO) OpenML. dataset-autoHorse_fixed [56]	201	69	Price
Residential Building (RESIDENT) Rafiei [57]	372	108	Actual Selling Price
Communities and Crime (CRIME) Redmond [58]	1,994	100	Violent Crimes per 100,000 Population
DETROIT Fisher [59]	13	14	Homicides per 100,000 Population
qsbr_rw1 (QSBR) Damborsky, et al. [60]	14	51	Structure–Biodegradability Relationships

6.2. Results

The proposed bootstrap multi-K KNN regression method was evaluated across diverse datasets, with results summarized in Figures 2–7. Predictive accuracy is assessed through Figures 2–5, and prediction interval quality is analyzed in Figures 6–7.

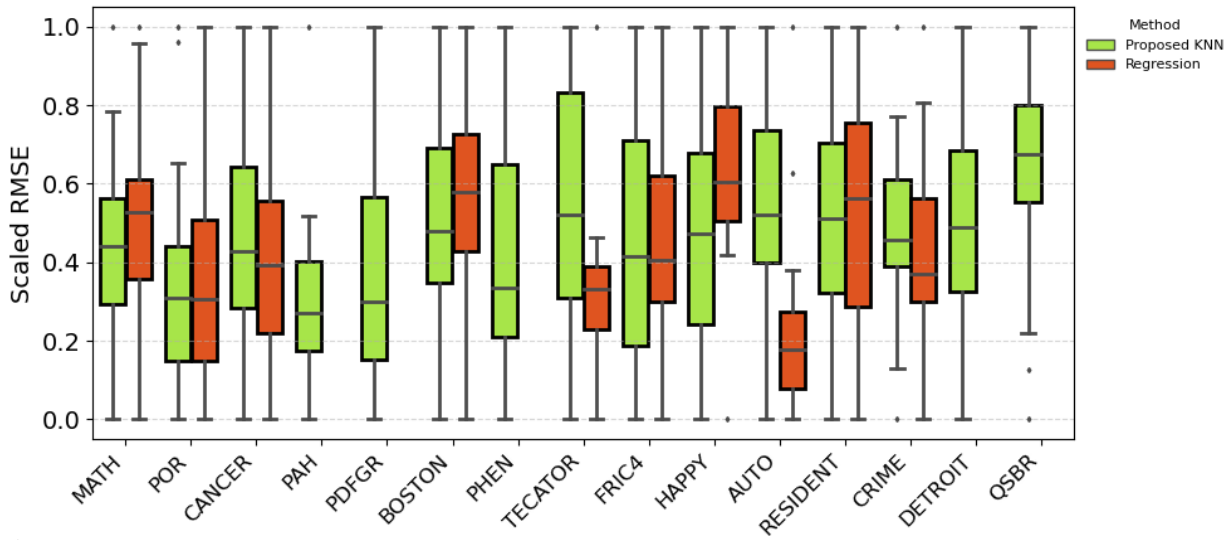

Figure 2.
Min-max scaled RMSE comparison across datasets. For high-dimensional datasets, only the proposed KNN method is applicable.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

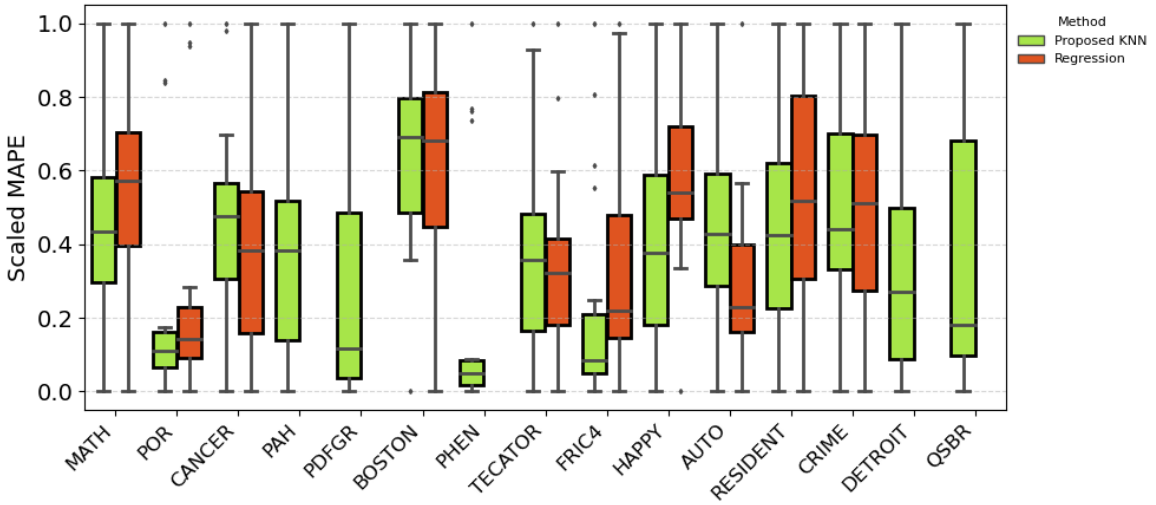

Figure 3.
Min-max scaled RMSE comparison across datasets.

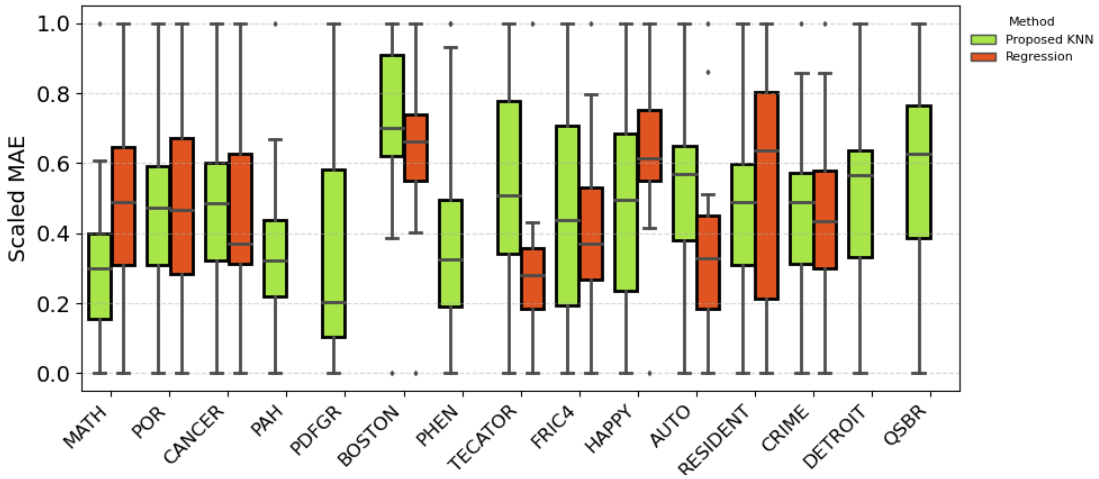

Figure 4.
Min-max scaled MAE comparison across datasets.

For low-dimensional datasets (where $n > p$), results for both the proposed KNN and regression methods are presented. Overall, both methods show competitive performance, though the proposed KNN consistently yields lower RMSE and MAE in several cases (e.g., MATH, BOSTON, HAPPY, and RESIDENT), indicating robust predictive accuracy, especially when modeling complex relationships. Figure 5 highlights the strength of the proposed KNN in providing superior explanatory power (higher R^2) compared to regression, particularly in non-linear scenarios such as AUTO and CRIME.

In high-dimensional settings (PAH, PDFGR, PHEN, DETROIT, and QSBR), where traditional regression fails due to the number of predictors exceeding observations, only the proposed KNN results are presented. Despite the absence of regression results for comparison, KNN achieves consistently high predictive accuracy ($R^2 > 0.6$), demonstrating its utility and robustness in challenging high-dimensional scenarios.

Figure 6 presents CP comparisons. Regression generally achieves higher CP, especially in datasets like CANCER, TECATOR, RESIDENT, and CRIME, due to excessively wide intervals. Conversely, the proposed KNN maintains tighter intervals, resulting in lower CP for some cases (PHEN,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

TECATOR). Notably, for datasets such as BOSTON and HAPPY, the proposed KNN intervals provide higher or comparable CP, indicating scenarios where regression intervals may be overly conservative or unnecessarily wide.

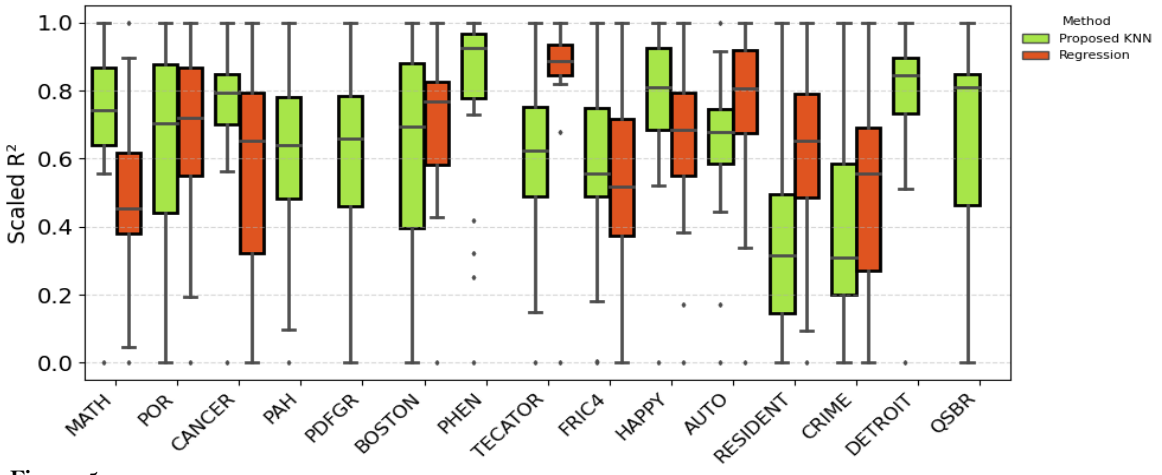

Figure 5.
Min-max scaled R^2 comparison across datasets. For high-dimensional datasets, only the proposed KNN method is applicable.

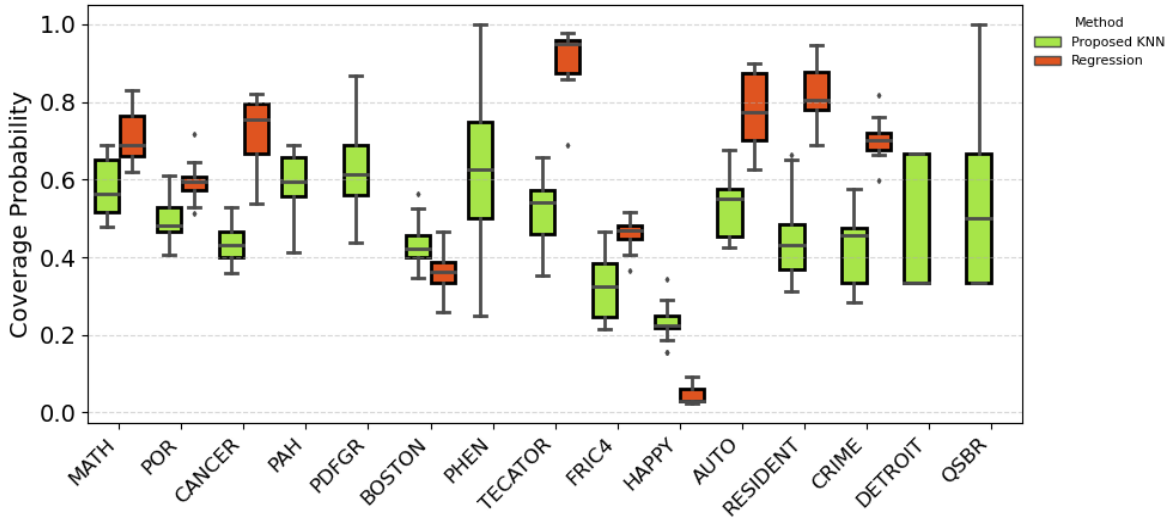

Figure 6.
Coverage probability comparison across datasets.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

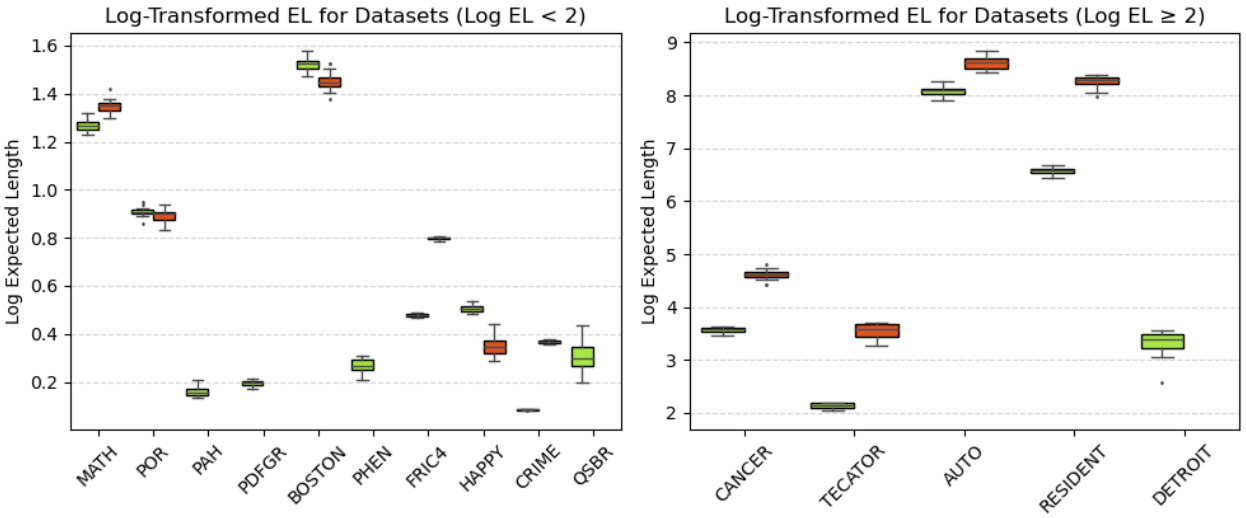

Figure 7.
Log-transformed expected length comparison.

The log-transformed EL results in Figure 7 clarify differences across datasets with varying interval widths. Regression intervals consistently exhibit larger EL values across datasets (e.g., MATH, FRIC4, CRIME, CANCER, TECATOR, AUTO, and RESIDENT). By contrast, the proposed KNN consistently provides narrower intervals, effectively balancing precision and predictive coverage.

Overall, these results demonstrate the methodological strength of the proposed multi-K bootstrap approach, particularly its applicability in high-dimensional contexts and its effectiveness in providing compact yet reliable intervals, offering valuable improvements over traditional regression-based methods.

7. Conclusion and Discussion

This study proposed a novel bootstrap-based multi-K approach for constructing robust prediction intervals in KNN regression, particularly addressing challenges posed by high-dimensional datasets. Traditional regression-based methods fail when predictors exceed observations; however, the proposed methodology remains effective under such scenarios, consistently achieving substantial predictive accuracy ($R^2 > 0.6$).

Although the proposed KNN intervals often yield lower CP compared to linear regression, particularly evident in datasets like TECATOR and PHEN, they significantly outperform regression in terms of EL, especially in challenging datasets such as AUTO, RESIDENT, and CANCER. This narrower interval enhances interpretability and practical decision-making utility, though it highlights the inherent trade-off between interval tightness and coverage reliability.

Furthermore, this study demonstrates the methodological flexibility of combining ensemble, cross-validation, and bootstrap methods within a unified multi-K framework, significantly improving interval estimation compared to traditional approaches. Future research may explore adaptive weighting or feature selection strategies to further optimize coverage probability without substantially sacrificing interval compactness. Extending this framework to other machine learning methods and evaluating its performance on large-scale datasets may also provide valuable insights for practitioners seeking reliable uncertainty quantification tools.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 5: 2750-2764, 2025
DOI: 10.55214/25768484.v9i5.7589

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] P. D. Easton, M. M. Kapons, S. J. Monahan, H. H. Schütt, and E. H. Weisbrod, "Forecasting earnings using k-nearest neighbors," *The Accounting Review*, vol. 99, no. 3, pp. 115-140, 2024. https://doi.org/10.2308/TAR-2021-0478

[2] K. Alkhatib, H. Najadat, I. Hmeidi, and M. K. A. Shatnawi, "Stock price prediction using k-nearest neighbor (kNN) algorithm," *International Journal of Business, Humanities and Technology*, vol. 3, no. 3, pp. 32-44, 2013.

[3] K.-j. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1-2, pp. 307-319, 2003.

[4] B. Chao and H. Guang Qiu, "Air pollution concentration fuzzy evaluation based on evidence theory and the K-nearest neighbor algorithm," *Frontiers in Environmental Science*, vol. 12, p. 1243962, 2024. https://doi.org/10.3389/fenvs.2024.1243962

[5] P. Thanh Noi and M. Kappas, "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery," *Sensors*, vol. 18, no. 1, p. 18, 2017. https://doi.org/10.3390/s18010018

[6] K. V. V. Ganesh, G. Sheetal, M. S. Amith, L. H. Goyal, and K. S. Rao, "Air quality prediction using KNN and LSTM," *Int. J. Innov. Sci. Res. Technol*, vol. 9, no. 4, 2024. https://doi.org/10.38124/IJISRT24APR1706

[7] P. Srisuradetchai and W. Panichkitkosolkul, "Using ensemble machine learning methods to forecast particulate matter (PM2. 5) in Bangkok, Thailand," presented at the In International Conference on Multi-disciplinary Trends in Artificial Intelligence (pp. 204-215). Cham: Springer International Publishing, 2022.

[8] Z. A. Sejuti and M. S. Islam, "A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation," *Sensors International*, vol. 4, p. 100229, 2023. https://doi.org/10.1016/j.sintl.2023.100229

[9] N. M. N. Al-Dosary, S. A. Al-Hamed, and A. M. Aboukarima, "K-nearest Neighbors method for prediction of fuel consumption in tractor-chisel plow systems," *Engenharia Agrícola*, vol. 39, no. 6, pp. 729-736, 2019. https://doi.org/10.1590/1809-4430-Eng.Agric.v39n6p729-736/2019

[10] C. Zhang, P. Zhong, M. Liu, Q. Song, Z. Liang, and X. Wang, "Hybrid metric k-nearest neighbor algorithm and applications," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 8212546, 2022. https://doi.org/10.1155/2022/8212546

[11] A. Jegorowa, J. Górski, J. Kurek, and M. Kruk, "Use of nearest neighbors (k-NN) algorithm in tool condition identification in the case of drilling in melamine faced particleboard," *Maderas. Ciencia y tecnologia*, vol. 22, no. 2, pp. 189-196, 2020. https://doi.org/10.4067/S0718-221X2020005000205

[12] A. Onan, "A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6844-6852, 2015. https://doi.org/10.1016/j.eswa.2015.05.006

[13] M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," *International Journal of Information and Education Technology*, vol. 2, no. 3, pp. 220-223, 2012. https://doi.org/10.7763/IJIET.2012.V2.114

[14] C. Liu *et al.*, "A K-nearest neighbor model to predict early recurrence of hepatocellular carcinoma after resection," *Journal of clinical and translational hepatology*, vol. 10, no. 4, p. 600, 2022. https://doi.org/10.14218/JCTH.2021.00348

[15] K. Archana and G. Komarasamy, "A novel deep learning-based brain tumor detection using the Bagging ensemble with K-nearest neighbor," *Journal of Intelligent Systems*, vol. 32, no. 1, p. 20220206, 2023. https://doi.org/10.1515/jisys-2022-0206

[16] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 1, p. 113, 2024. https://doi.org/10.1186/s40537-024-00973-y

[17] P. Srisuradetchai and K. Suksrikran, "Random kernel k-nearest neighbors regression," *Frontiers in big Data*, vol. 7, p. 1402384, 2024. https://doi.org/10.3389/fdata.2024.1402384

[18] S. Bermejo and J. Cabestany, "Adaptive soft k-nearest-neighbour classifiers," *Pattern Recognition*, vol. 33, no. 12, pp. 1999-2005, 2000. https://doi.org/10.1016/S0031-3203(99)00186-7

[19] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143-148, 2016. https://doi.org/10.1016/j.neucom.2015.08.112

[20] S. Ingram and T. Munzner, "Dimensionality reduction for documents with nearest neighbor queries," *Neurocomputing*, vol. 150, pp. 557-569, 2015. https://doi.org/10.1016/j.neucom.2014.07.007

[21] P. K. D. Pramanik, M. Mukhopadhyay, and S. Pal, *Big data classification: Applications and challenges. In Artificial Intelligence and IoT; Manoharan, K. G., Nehru, J. A., Balasubramanian, S., Eds.; Studies in Big Data*. Singapore: Springer, 2021.

[22] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A new K-nearest neighbors classifier for big data based on efficient data pruning," *Mathematics*, vol. 8, no. 2, p. 286, 2020. https://doi.org/10.3390/math8020286

[23] H. I. Abdalla and A. A. Amer, "Towards highly-efficient k-nearest neighbor algorithm for big data classification," in *In Proceedings of the 2022 5th International Conference on Networking, Information Systems and Security; IEEE: New York City, NY, USA, 2022; pp. 1–5*, 2022.

[24] N. Ukey, Z. Yang, B. Li, G. Zhang, Y. Hu, and W. Zhang, "Survey on exact knn queries over high-dimensional data space," *Sensors*, vol. 23, no. 2, p. 629, 2023. https://doi.org/10.3390/s23020629

[25] M. G. Hamed, M. Serrurier, and N. Durand, "Possibilistic KNN regression using tolerance intervals," in *In Proceedings of the 14th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2012); Springer: Catania, Italy, 2012; pp. 410–419. https://doi.org/10.1007/978-3-642-31718-7_43*, 2012.

[26] M. G. Hamed, M. Serrurier, and N. Durand, "Simultaneous interval regression for K-nearest neighbor," in *In Proceedings of the 25th Australasian Joint Conference on Artificial Intelligence (AI 2012); Springer: Sydney, Australia, 2012; pp. 602–613. https://doi.org/10.1007/978-3-642-35101-3_51*, 2012.

[27] P. Srisuradetchai, "A novel interval forecast for k-nearest neighbor time series: a case study of durian export in Thailand," *IEEE Access*, vol. 12, pp. 2032-2044, 2023. https://doi.org/10.1109/ACCESS.2023.3348078

[28] E. Ozturk Kiyak, B. Ghasemkhani, and D. Birant, "High-level K-nearest Neighbors (HLKNN): a supervised machine learning model for classification analysis," *Electronics*, vol. 12, no. 18, p. 3828, 2023. https://doi.org/10.3390/electronics12183828

[29] P. G. Giannopoulos, T. K. Dasaklis, and N. Rachaniotis, "Development and evaluation of a novel framework to enhance k-NN algorithm's accuracy in data sparsity contexts," *Scientific Reports*, vol. 14, no. 1, p. 25036, 2024. https://doi.org/10.1038/s41598-024-76909-6

[30] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports*, vol. 12, no. 1, p. 6256, 2022. https://doi.org/10.1038/s41598-022-10358-x

[31] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction* (Springer). Cham, Switzerland, 2009.

[32] D. W. Scott, *Multivariate density estimation: Theory, practice, and visualization*. New York: John Wiley & Sons, 2015.

[33] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An introduction to statistical learning: With applications in python*. Cham, Switzerland: Springer, 2023.

[34] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge, UK: Cambridge University Press, 2020.

[35] A. Zollanvari, *Machine learning with python: Theory and implementation*. Cham, Switzerland: Springer, 2023.

[36] J. White and S. D. Power, "K-fold cross-validation can significantly over-estimate true classification accuracy in common EEG-based passive BCI experimental designs: an empirical investigation," *Sensors*, vol. 23, no. 13, p. 6077, 2023. https://doi.org/10.3390/s23136077

[37] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'95) Morgan Kaufmann Publishers San Francisco 1995.

[38] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *Journal of Econometrics*, vol. 187, no. 1, pp. 95-112, 2015. https://doi.org/10.1016/j.jeconom.2015.02.006

[39] U. Kummaraka and P. Srisuradetchai, "Time-series interval forecasting with dual-output Monte Carlo dropout: A case study on durian exports," *Forecasting*, vol. 6, no. 3, pp. 616-636, 2024. https://doi.org/10.3390/forecast6030033

[40] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247-1250, 2014. https://doi.org/10.5194/gmd-7-1247-2014

[41] K. Suksrikran and P. Srisuradetchai, "Enhanced kernel k-nearest neighbors regression with backward feature selection," in *In Proceedings of the 2024 22nd International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 2024; pp. 1–6. https://doi.org/10.1109/ICTKE62841.2024.10787198*, 2024.

[42] D. L. Alexander, A. Tropsha, and D. A. Winkler, "Beware of R 2: Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models," *Journal of chemical information and modeling*, vol. 55, no. 7, pp. 1316-1322, 2015. https://doi.org/10.1021/acs.jcim.5b00206

[43] A. Niyomdecha and P. Srisuradetchai, "Complementary gamma zero-truncated Poisson distribution and its application," *Mathematics*, vol. 11, no. 11, p. 2584, 2023. https://doi.org/10.3390/math11112584

[44] P. Srisuradetchai and K. Dangsupa, "On interval estimation of the geometric parameter in a zero–inflated geometric distribution," *Thailand Statistician*, vol. 21, no. 1, pp. 93-109, 2023.
[45] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Comprehensive review of neural network-based prediction intervals and new advances," *IEEE Transactions on neural networks*, vol. 22, no. 9, pp. 1341-1356, 2011. https://doi.org/10.1109/TNN.2011.2162110
[46] T. Pearce, A. Brintrup, M. Zaki, and A. Neely, "High-quality prediction intervals for deep learning: A distribution-free, ensembled approach," in *In Proceedings of the 35th International Conference on Machine Learning; Proceedings of Machine Learning Research, Vol. 80, 2018, pp. 4075–4084. https://proceedings.mlr.press/v80/pearce18a.html*, 2024.
[47] P. Cortez, "Student performance data set," Retrieved: https://archive.ics.uci.edu/ml/datasets/Student, 2014.
[48] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast cancer wisconsin (diagnostic)," UCI Machine Learning Repository, 10, C5DW2B. https://doi.org/10.24432/C5DW2B, 1995.
[49] R. Todeschini, P. Gramatica, R. Provenzani, and E. Marengo, "Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons," *Chemometrics and Intelligent Laboratory Systems*, vol. 27, no. 2, pp. 221-229, 1995. https://doi.org/10.1016/0169-7439(95)80026-6
[50] R. Guha and P. C. Jurs, "Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors," *Journal of chemical information and computer sciences*, vol. 44, no. 6, pp. 2179-2189, 2004. https://doi.org/10.1021/ci049849f
[51] D. Harrison Jr and D. L. Rubinfeld, "Hedonic housing prices and the demand for clean air," *Journal of environmental economics and management*, vol. 5, no. 1, pp. 81-102, 1978. https://doi.org/10.1016/0095-0696(78)90006-2
[52] H. Kubinyi, "QSAR: Hansch analysis and related approaches methods and principles in medicinal chemistry." Germany: Wiley-VCH, 1993, p. 438.
[53] C. Borggaard and H. H. Thodberg, "Optimal minimal neural interpretation of spectra," *Analytical chemistry*, vol. 64, no. 5, pp. 545-551, 1992. https://doi.org/10.1021/ac00029a018
[54] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, pp. 1189-1232, 2001.
[55] J. Helliwell, R. Layard, and J. Sachs, *World happiness report*. New York: Sustainable Development Solutions Network, 2017.
[56] OpenML. dataset-autoHorse_fixed, "OpenML. dataset-autoHorse_fixed," Retrieved: https://www.openml.org/d/42224. [Accessed 2024.
[57] M. Rafiei, "Residential building data set," *UCI Machine Learning Repository*, 2018. https://doi.org/10.24432/C5S896
[58] M. Redmond, "Communities and crime," UCI Mach. Learn. Repos. https://doi.org/10.24432/C53W3X, 2009.
[59] J. C. Fisher, "Homicide in detroit the role of firearms," *Criminology*, vol. 14, no. 3, pp. 387-400, 1976. https://doi.org/10.1111/j.1745-9125.1976.tb00030.x
[60] J. Damborsky, M. Lynam, and M. Kuty, "Structure-biodegradability relationships for chlorinated dibenzo-p-dioxins and dibenzofurans," Springer, 1998, pp. 165-228.