

Explainable Global Error Weighted on Feature Importance: The xGEWFI metric to evaluate the error of data imputation and data augmentation

Jean-Sébastien Dessureault^{1*} and Daniel Massicotte¹

^{1*}Laboratoire des signaux et des systèmes intégrés, Département of Electrical and Computer Engineering, Université du Québec à Trois-Rivières, 3351 Bd des Forges, Trois-Rivières, G9A 5H7, Québec, Canada.

*Corresponding author(s). E-mail(s):

sebastien.dessureault@uqtr.ca;

Contributing authors: daniel.massicotte@uqtr.ca;

Abstract

Evaluating data imputation and augmentation performance is a critical issue in data science. In statistics, methods like Kolmogorov–Smirnov K-S test, Cramér–von Mises \mathbf{W}^2 , Anderson–Darling \mathbf{A}^2 , Pearson’s χ^2 and Watson’s \mathbf{U}^2 exists for decades to compare the distribution of two datasets. In the context of data generation, typical evaluation metrics have the same flaw: They calculate the feature’s error and the global error on the generated data without weighting the error with the feature’s importance. In most cases, the importance of the features is imbalanced, and it can induce a bias on the features and global errors. This paper proposes a novel metric named "Explainable Global Error Weighted on Feature Importance" (*xGEWFI*). This new metric is tested in a whole preprocessing method that 1. Process the outliers, 2. impute the missing data, and 3. augments the data. At the end of the process, the *xGEWFI* error is calculated. The distribution error between the original and generated data is calculated using a *Kolmogorov-Smirnov test* (K-S test) for each feature. Those results are multiplied by the importance of the respective features and calculated using a *Random Forest* (RF) algorithm. The metric result is expressed in an explainable format,

aiming for an ethical AI. This novel method provides a more precise evaluation of a data generation process than if only a K-S test were used.

Keywords: xGEWFI; Data imputation; Data augmentation; Random forest; SMOTE; KNNImputer

1 Introduction

Traditionally statistics methods are used to analyze the performance of data generation. Those methods were sufficient when each variable was taken alone. In a more complex dataset context, those methods do not consider that every feature has different importance. Although, errors on each feature are weighted equally. This flaw can be improved with machine learning techniques that evaluate the significance of the features in a dataset. Providing a solution for this flaw is precisely the motivation of this paper. In a context where missing data imputation and data augmentation are evolving rapidly, developing new explainable tools to evaluate those machine learning algorithms is crucial.

Data augmentation or imputation results are typically evaluated at the end of the process, resulting in an error or accuracy level. This method is straightforward, although it could be more explainable. There is a process followed by a metric of evaluation (error or accuracy). Hence, another critical flaw of the traditional method is that nothing comes to justify the "black box" between the input and the output. Traditional methods need to be explained more.

To calculate the error for each feature by comparing the original distribution and the generated (augmented or imputed) data, a "Goodness-of-fit" (GOF) must be used. The work of [1] discusses about how GOF methods will test statistics for ordinal data that uses and empirical distribution function. It compares the comparison tests of Kolmogorov–Smirnov (K-S), Cramér–von Mises (W^2), Anderson–Darling (A^2), Pearson's (χ^2) and Watson's (U^2).

The work of [2] is also a good reference on this GOF topic, comparing χ^2 with W^2 and A^2 . The conclusion of those work is that each one has it pros and cons, according to the distribution of the features and its cardinality.

To compare the generated data to the original data, the (*xGEWFI*) metric uses a *Kolmogorov-Smirnov test* [3]. This test has been widely used for decades in mathematics. It compares the properties of the distribution of two series of data. In a data imputation context, the imputed data for a feature are compared with the original data for this same feature. In a data augmentation context, the added data are compared with the original data. Comparing the generated and the original data, two metrics (statistical and value) are returned to describe the covariance between them. It shows the quality of the data generation process. This test is well-known and has been used since decades [4].

At the heart of the $xGEWFI$ metric is the evaluation of the feature's importance. Normally, all measures of feature error are equally weighted. A better representation would take into account the importance of each feature [5]. This is why a *Random Forest* (RF) algorithm is used in this novel metric. Based on [6], an RF algorithm is used as a regressor, as a classifier, but also as a tool to evaluate the importance of the features. It works in a supervised learning context, based on *bayesians networks* and on *decisions trees*. It has been widely used in a variety of applications [7].

The next algorithms (*SMOTE*, *KNNImputer*, and "Interquartile Range" (*IQR*)) are useful to implement the whole method that test the $xGEWFI$ algorithm. This method is a whole preprocessing pipeline that makes detection of outliers, data imputation and data augmentation.

The process of outliers detection is made using the *IQR* method. Outliers are feature's values that fall outside of the normal distribution. The *IQR* method defines outliers in a mathematical and formal matter, using the interquartile range rule. It is commonly used as in the papers of [8] [9]. Another good choice of outliers detection method would have been RANSAC (*RANdom SAMple Consensus*) algorithm [10]. RANSAC is an iterative algorithm that fits a model to a subset of the data and identifies outliers based on their deviation from the model. It is a more sophisticated method that can handle more complex models and is effective when there are a large number of outliers, or the data is contaminated with noise.

The data imputation process is done using the *K-Nearest-Neighbor Imputer* (*KNNImputer*). It finds a missing value using *k-Nearest Neighbors* algorithm. Each sample's missing values are imputed using the mean value from n nearest neighbors availables for a given feature. It is widely used, like in the works of [11] [12]. There are some more advanced methods to impute some missing data. Some specialized *Generative Adversial Networks* (GAN), like *Generative Adversial Imputation Networks* (GAIN) algorithm has been recently created [13] [14].

The data augmentation is implemented using the *Synthetic Minority Over-sampling Technique* (SMOTE) algorithm [15]. This algorithm help to solve the imbalanced datasets problem by over-sampling the minority classes. It exists some more recent and state-of-the-art methods to perform data augmentation. The GANs, for instance [16] [17] uses two deep neural networks to create new data. Although, the SMOTE algorithm is still widely used [18] [19] and sufficient for this method that validate the proposed $xGEWFI$ metric.

The datasets used to validate this method has been generated using the *Scikit-Learn* framework. Especially the *datasets.make_regression()* and *datasets.make_classification()* functions using different parameters. Both were used to produce a different datasets based on regression and classification. Both functions are widely used in various paper such as [20] [21]. The generated datasets had the advantage of being fully reproducible when being called with the same parameters.

This novel method weights the error calculated after generating data (missing or augmented data). The weighting is done on the importance of the features, giving an error a weight related to its feature importance. It is also important to create new tools with the ability to explain the results. For the ethical purpose, from now on, efforts have to be made to produce algorithms with explainable results. This novel "Explainable Global Error Weighted on Feature Importance" $xGEWFI$ metric includes itself in this paradigm. The explainability layer evaluates 1. the distribution error of the augmented/imputed features related to the original data distribution. 2. the importance of the features. 3. An index of each feature distribution error weighted by the each feature importance. 4. A global weighted error. Those 4 points can be presented using tables and graphics, helping to the lack of explainability in the traditional metrics of machine learning (error and accuracy)

The main contribution of this paper is to propose a more precise and less biased metric that into account the importance of the feature and, consequently, the importance of the feature's error. It provides a more accurate metric to evaluate the performances of the data imputation and augmentation process. It also contributes to AI ethics by proposing an explainable metric, especially for data missing and data augmentation.

The next sections of this paper are organized with the following structure: Section 2 describes the proposed methodology. Section 3 presents the results. Section 4 discusses about the results and their meaning and Section 5 concludes this research.

2 Methodology

2.1 Architecture

Figure 1 shows the proposed method's architecture that includes and validates the $xGEWFI$ metric.

This figure shows that the first step consists of sending the raw data to the outliers detection module. This part does the following: 1. detects outliers 2. replaces the outliers with a null value. Those null values will be replaced by correct values at the next step, using a KNNImputer Algorithm. The next part does the data generation. The first one imputes the data missing using a KNNImputer algorithm. The second one does the data augmentation process using a SMOTE algorithm. The results are the imputed data and the augmented data generated. Both are returned to the data scientist and sent to the $xGEWFI$ algorithm. The $xGEWFI$ algorithm compares the generated data distribution to the original data's distribution using a K-S test (1). It evaluates the error of the data generation process, both globally (3) and for each feature (2). An RF algorithm is executed afterward to find the feature importances. Finally, the $xGEWFI$ is calculated based on each feature's error and importance. The global weighted error is defined in (5) and in (4) for each feature. All the information for the $xGEWFI$ metric (error and explainability) is sent to the data scientist.

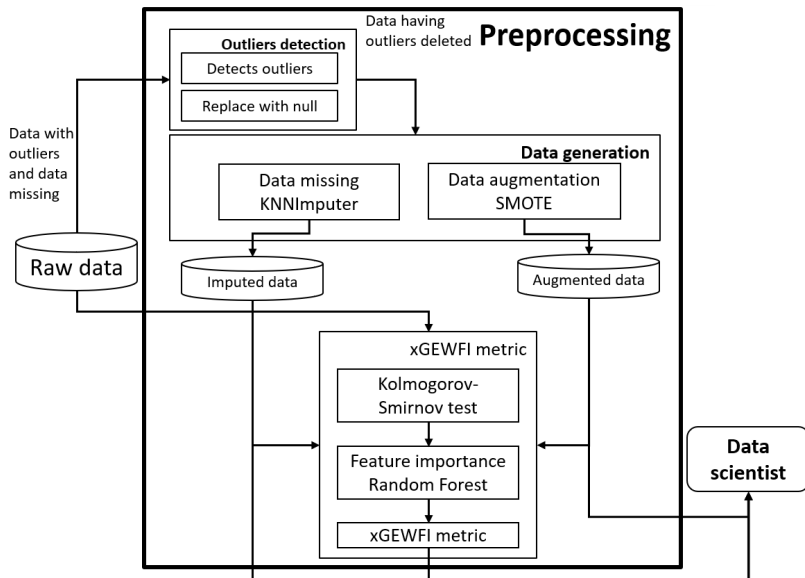


Fig. 1: Architecture of the preprocessing method that includes the $xGEWFI$ metric

2.2 $xGEWFI$ metric - Global Error Weighted on Feature Importance

Metrics usually measure errors or performances of a process without considering the feature's importance. The premise of $xGEWFI$ metric is that the feature importance should weigh the error amplitude. A high-value error should represent nothing if the feature importance is null or extremely low. Conversely, the amplitude of an error should be higher with high feature importance. We must first calculate the feature importance using an RF algorithm to calculate the $xGEWFI$ metric. RF algorithms are based on Bayesian networks and multiple decision trees. A single decision tree may lead to some bias, so having tenths of decision trees (as proposed by the RF algorithm) solve this problem, as it pools all the outcomes to return the most frequent answer. It exists two applications of RF algorithms: 1. RF regressors and 2. classifiers. Both are also used to evaluate the importance of the features. The solution can be displayed in a tree graph where the nodes represent the decision. The branches represent the possible outcomes. The leaves represent all the possible answers, combined with their probability. With this ability to compute a tree of bayesian probability, the RF algorithms can also calculate the feature importances in the regression of the classification process. The importance of each feature is given in a normalized form.

Then we must calculate the performance of the data generation algorithms (data imputation and data augmentation). The K-S test is used to do so. This test is used to compare two distributions. In our case, the original data

distribution is compared with the generated (imputed or augmented) data distribution. The formula of the K-S test is shown in (1).

$$D_f = \sup |F_{o,f}(x) - F_{g,f}(x)| \quad (1)$$

Where D_f is the result, the D-statistic, for feature f . $F_{o,f}$ is the distribution function of the feature f of o , the original x data. Similarly, $F_{g,f}$ is the distribution function of the same feature f of g , the generated x data. The case where there is no difference between the two distributions is called the null hypothesis. In (1) the result D (the D-statistic) is the evaluation of the error. A value of $D = 0.0$ is the null hypothesis.

Without calculating the feature importance, we can conclude that, since the D-statistics indicates the error between two distributions, it can be considered as the error of the generated distribution as in (2).

$$E_f = D_f \quad (2)$$

The global error is the sum of all the feature's errors, as in (3)

$$GlobalError_f = \sum_{f=1}^{featureNb.} E_f \quad (3)$$

Even though those numbers are not the finality, they are returned by the $xGEWFI$ algorithm as part of the final answer for explainability matters. Those numbers are an important part of the final answer and must be available to the data scientist to help him to have a better comprehension of the process.

Based on both the RF algorithm and the K-S test, the $xGEWFI$ metric relies on the D-Statistics and on the importance of the features. (4) and (5) shows the calculations for each feature error and for global error, respectively.

$$WeightedError_f = E_f * W_f \quad (4)$$

f is the feature index. E_f is the error of feature f using the K-S test. W_f is the weight of feature f according to the RF algorithm.

$$xGEWFI = \sum_{f=1}^{featureNb.} WeightedError_f \quad (5)$$

The method used the default parameters of the `scipy.stats.ks_2samp` for the K-S test. The `alternative` parameter is set to 'two-sided'. The null hypothesis is that the two distributions are identical, $F(x)=G(x)$ for all x . The alternative is that they are not similar. The `alternative = 'two-sided'` parameter means that the statistic is the maximum absolute difference between the empirical distribution functions of the samples. The `method` parameter can be set to 3 different values: 1. 'exact' uses the exact distribution of test statistic. 2. 'asympt' uses the asymptotic distribution of test statistics, and 3. 'auto' uses 'exact' for small-size arrays, 'asympt' for large. This last one is the default parameter used.

The class `sklearn.ensemble.RandomForestClassifier` is used by the RF classifier. All the default parameters have been used. The two most important are `n_estimators` is set to 100 and `criterion` is set to 'gini'. The `n_estimators` is the number of decision trees pooled by the RF algorithm. The Gini impurity is a metric used in decision trees to evaluate the impurity or homogeneity of a split in the data. It measures the probability of misclassifying a randomly chosen element from the dataset if it is incorrectly labeled according to the distribution of the classes in the split.

2.3 Outliers data

Outliers are data points that deviate significantly from the normal distribution of the data and can substantially impact the results of statistical analyses. Detecting outliers is critical in many research studies and can pose several challenges. One of the primary challenges is deciding on an appropriate method for detecting outliers sensitive to the data's specific characteristics and the research question. Some methods may be more suitable for detecting extreme outliers, while others may be better suited for identifying subtle deviations from the normal distribution. Additionally, the definition of an outlier can be subjective and depend on the researcher's interpretation of the data.

The first part of the process consists of identifying the data outliers. The need is to identify the outliers and replace them with some null values that will be processed later. To do that part, we use the IQR method. It consists of dividing each feature of the dataset into quartiles. Fig. 2 shows a box plot representation of the IQR method.

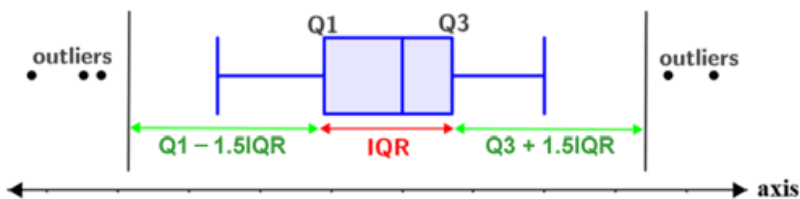


Fig. 2: IQR method to detect outliers [22]

The first quartile point Q1 indicates that 25% of the data points are below that value. The second quartile Q2 is the median point of the feature. The third quartile Q3 is at the point where 75% of the data points are below that value. Eq. (6) defines the IQR, (7) defines the lower limit, and (8) defines the upper limit.

$$IQR = Q3 - Q1 \quad (6)$$

$$LowerLimit = Q1 - 1.5 * IQR \quad (7)$$

$$UpperLimit = Q3 + 1.5 * IQR \quad (8)$$

When a value outside the lower and upper is found, it is replaced by a null value. This value will be processed once again in the next step while processing the missing data.

2.4 Data imputation

Data imputation is a common approach for handling missing data in research studies, particularly in large datasets. However, the process of imputing missing data poses several challenges that can affect the accuracy and validity of the results. One of the primary challenges is ensuring that the imputed values are accurate and representative of the missing data. Data imputation can introduce bias into the analysis if the imputed values are not independent of the missing data or if the imputation method does not account for the underlying data distribution. Therefore, it should be carefully considered before usage. Nonetheless, when used correctly, it can be advantageous to gain the ability to improve a dataset by this means.

The system trained a KNNImputer to impute the values of the missing data. This algorithm relies on the KNN algorithm aiming to find the k nearest neighbour of the missing data. This algorithm is based on the computation of the distance between the data. In this case, the euclidian distance is used as in (9).

$$D = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (9)$$

D is the euclidian distance, k is the number of neighbours to find. Finally, x and y are the origin and destination data, respectively.

2.5 Data augmentation

Data augmentation is the process of artificially increasing the size of a dataset by generating new examples from existing ones. While data augmentation can be a powerful tool for improving the performance of machine learning models, it can also be challenging for several reasons. For instance, the distribution of augmented data might be different from the original data. There is also a need for domain-specific knowledge, quality control, computational resources, risk of overfitting, and additional labeling effort. However, with careful consideration and appropriate techniques, data augmentation can be a powerful tool for improving the performance of machine learning models.

A SMOTE architecture is used to oversample data of the original dataset. It helps to solve the imbalanced data problem. It balances class distribution by randomly increasing minority instances. Virtual training records are generated by linear interpolation for the minority class. The algorithm selects the k -nearest neighbours for each data added in the minority class. To compute the distance between the neighbours, the euclidian distance is used as defined in (9).

3 Results

Let us study two cases to explain the advantages of the $xGEWFI$ metric that weights the error with the feature's importance. In a case where all the features have the same importance, the weighting specific to the $xGEWFI$ process would not affect the result much. Although, in the majority of the cases, the difference is significant. The following subsections present two typical and common cases where the $xGEWFI$ metric changes drastically our evaluation of the data imputation and data augmentation process. Both cases use a different dataset. Both datasets have five features and 25000 data. Before presenting the cases, the section 3.1 presents the datasets.

3.1 Preprocessing of dataset

As mention in section 1, the datasets are generated by the `datasets.make_regression()` and the `datasets.make_classification()` of the *Scikit-learn* framework. When called, the parameters allow selecting different dataset characteristics according to the test. It is possible to customize the number of data and features. Here is a description of the parameters that generate the synthetic data. *n_sample*: The number of rows or generated data. *n_features*: The number of columns or features. *n_target* and *n_classes*: The number of field that targets the regression and the classification data respectively. In all our cases, this one is always equal to 1. *shuffle*: When equal to True, it randomly changes the order of the data. This parameter is always True in our cases. *random_state*: It is the seed used to randomly generated the data. If the same seed (an integer number) is used, then the generated

data will always be the same. In our case, a value of 1 is always used for better reproducibility. For this study, 5% of outliers and 30% of data missing have been randomly generated.

3.2 Case 1 - Similar features errors and different features importances

This first case evaluates the results of a data imputation on a regression problem. We have a similar feature error as presented in Fig. 3. We can see in this figure that every feature shows an error of near 0.35 on the K-S test (axe Y) for each feature on axe X.

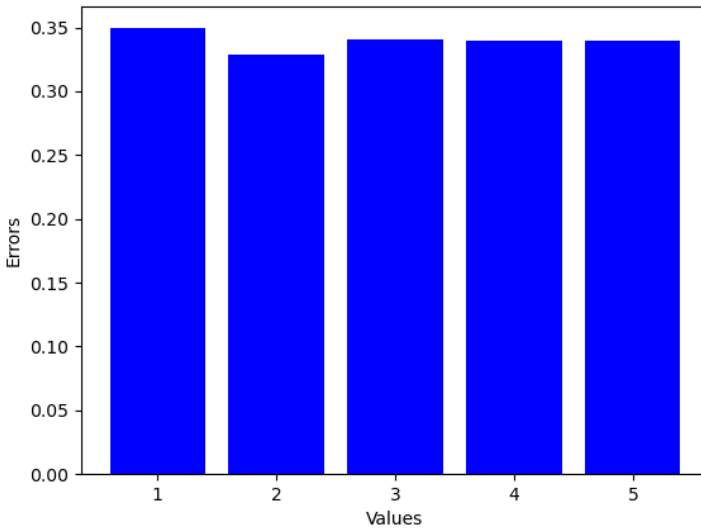


Fig. 3: Standard K-S test error.

The sum of the errors is 1.68, representing a final metric on the quality of the imputation process. Without the xGEWFI metric, evaluating the data imputation would stop here. Is this metric alone the best method to represent the quality of this data imputation process? The *xGEWFI* method proposes a method that considers each feature's weight. The final result (the *xGEWFI* error) will be more representative of the quality of the augmentation/imputation process, having weighted the K-S test error on the feature importance. For this case, Fig. 4 presents the importance of the features. We can see that every feature on axe X has a different level of normalized importance on axe Y.

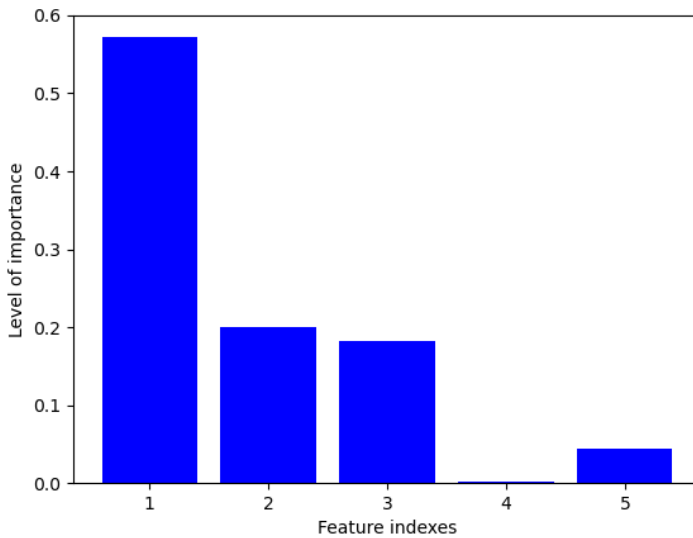


Fig. 4: Importance of the features.

Fig. 5 shows the result of the K-S test weighted by the feature importance. Since every feature has about the same error as shown in Fig. 3, the result of the $xGEWFI$ error is also about the same as the feature importance. It can be seen clearly by comparing Fig. 4 and Fig. 5.

We can conclude for this test that the error using a K-S test is very different than the $xGEWFI$ error (Fig. 3 and Fig. 5). The global error of the K-S test was 1.68, and the $xGEWFI$ global error was 33.75. The magnitude of the two metrics cannot be compared. Each one must be compared with other errors of the same kind.

3.3 Case 2 - Important differences between features importances and errors

This second case evaluates the results of data augmentation on a classification problem. Fig. 6 shows the result of the K-S test representing the error (ax Y) on each feature (ax X). This figure shows an important variation between each feature's error at the opposite of the K-S test error presented in case 1.

The feature's importance varies significantly, as shown in Fig. 7. Without going further using $xGEWFI$, we would conclude that the errors on features 1 and 4 are higher. The errors on features 3 and 5 are insignificant compared to the other feature's errors. A better way to evaluate the error would be to consider the features' importance. Fig. 7 presents the importance of the features for case 2.

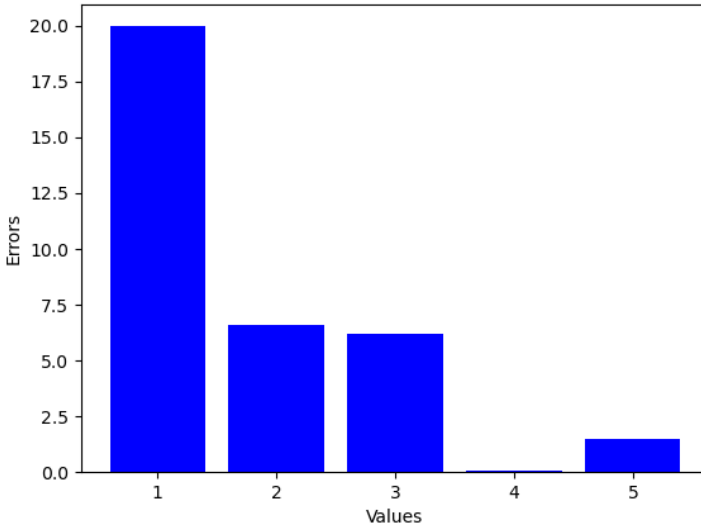


Fig. 5: Weighted $xGEWFI$ error.

In this figure, the X axe is the dataset's features, and the Y axe is the normalized level of importance of the features. We can conclude, for instance, that feature 3 is by far the most important, and feature 4 is negligible. Let us see the impact of weighting the K-S test error with the feature's importance. This is shown in the $xGEWFI$ graphic in Fig.8.

We can note that every feature's error has been weighted according to the importance of the features. Let us study specifically feature 3 and feature 4. Feature 3 K-S test error (Fig.6) is very low. It is also by far the most important feature (Fig. 7). Ultimately, the relative importance of the feature compared to the other features has been raised. As for feature 4, the K-S test error was quite high (Fig.6). Since the importance of this feature is very low (Fig. 7), the $xGEWFI$ error for this feature is lower than its K-S test error evaluation.

Like in case 1, the second case allows us to conclude that the results are different, whether they are weighted or not using the feature importance using the $xGEWFI$ error (Fig. 3 and Fig. 5). In this case, the global error of the K-S test was 0.60, and the $xGEWFI$ global error was 0.54. As mentioned in 3.2, the magnitude of the two metrics cannot be compared. They must be compared with their respective kind.

This case showed that $xGEWFI$ metric as the advantage of being weighted using the feature's importance. Also, the results are explainable, as shown in Section 3.4.

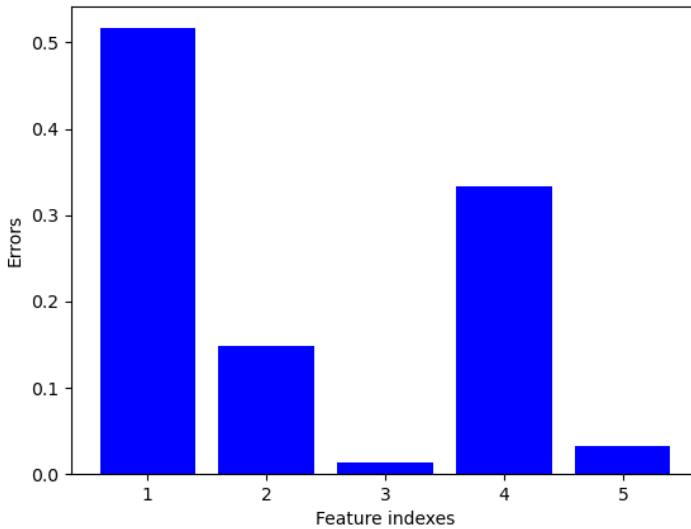


Fig. 6: Standard K-S test error.

3.4 Explainability

The "x" at the beginning of $xGEWFI$ stands for "explainable". This metric has not only been developed to give a quantitative evaluation of a data imputation or a data augmentation. It has also been developed to explain its process. That is why graphics and tables were encapsulated in the $xGEWFI$ method. Fig. 6 7), 8, for instance, are fully part of the method, aiming to give the users a better comprehension of the process. There are also 3 other types of graphics created to help the user better comprehend the data and the process. Fig. 9 is an example (the dataset used for case 1) of a graphic that shows the distribution of each feature using a box plot graphic.

The median marks the mid-point of the feature's values and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value, and half are lower. The middle box represents the middle 50% of scores for the feature. The upper and lower whiskers represent scores outside the middle 50%. Dots outside of the whiskers are outliers.

Fig.10 is another visual tool used to explain the data and the process. The $xGEWFI$ method generates one graphic of this type by feature. It is a histogram of the original and generated (imputed or augmented) data.

Blue bars represent the original data, and yellow bars represent the generated data. X axe is the standard deviation of the feature, and Y axe is the number of occurrences of those standard deviation ranges. Gaussian distribution is visible for both generated and original data.

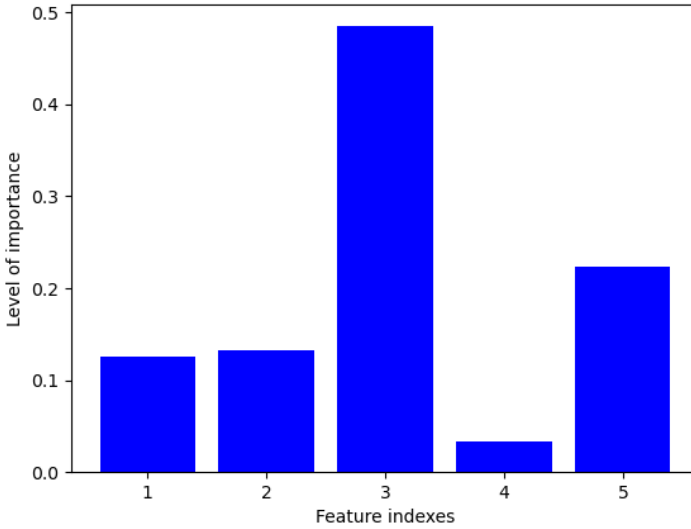


Fig. 7: Importance of the features.

The third type of graphic combines the 3 important parameters used in the $xGEWFI$ equation ((4)). It shows the magnitude of the 3 parameters.

X axis shows the features, and Y axis shows the magnitude of the values. Red bars are the feature importances, blue bars are the K-S test feature errors, and the magenta bar (combines red and blue) are the resulting $xGEWFI$ error.

Ultimately, the method offers the user the numerical data required to calculate the $xGEWFI$ metric. The data are displayed on the screen, and two LaTeX tables are generated and ready to be included in a document. For instance, tables 1 and 2 has been generated by the $xGEWFI$ method in case 1.

Table 1: Results of the $xGEWFI$ metric.

Metrics	Values
$xGEWFI$ mean error	33.76
K-S mean error	1.68

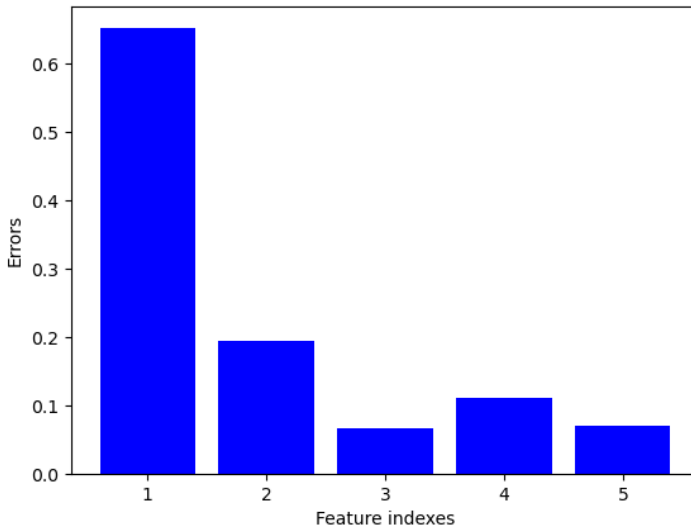


Fig. 8: Weighted $xGEWFI$ error.

Table 2: Explainability of the $xGEWFI$ metric.

Features	Imp.	K-S error	$xGEWFI$ error
Feature 1	0.57	0.34	19.23
Feature 2	0.2	0.34	6.84
Feature 3	0.18	0.34	6.16
Feature 4	0.0	0.34	0.07
Feature 5	0.04	0.33	1.45

3.5 Comparison between this novel method and original methods

Section 1 introduces the K-S, W^2 , A^2 , χ^2 and U^2 statistical tests. They are used to assess how well a sample of data fits a particular distribution. However, their assumptions, test statistics, and sensitivity to deviations from the assumed distribution differ.

The K-S test is a non-parametric test used to compare the empirical distribution of a sample with a theoretical distribution. The maximum difference between the empirical and theoretical cumulative distribution functions is the test statistic. The K-S test is sensitive to differences in both location and shape of the distributions.

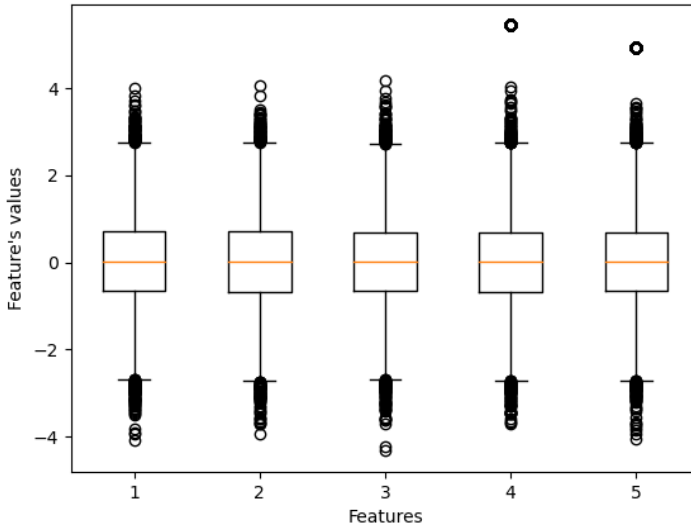


Fig. 9: Example of a box plot graphic encapsulated in the $xGEWFI$ method for better explainability.

W^2 test is also a non-parametric test used to compare the empirical distribution of a sample with a theoretical distribution. The test statistic is based on the squared differences between the empirical and theoretical cumulative distribution functions. It is more sensitive to differences in the tail regions of the distribution than the K-S test.

A^2 test is a parametric test used to compare the empirical distribution of a sample with a specific parametric distribution. The test statistic is based on the squared differences between the observed and expected values of the cumulative distribution function. It is particularly sensitive to deviations in the tails of the distribution.

χ^2 test is a parametric test used to compare the observed frequency distribution of a sample with an expected frequency distribution derived from a theoretical distribution. The test statistic is based on the squared differences between the observed and expected frequencies. It assumes that the sample data is discrete and independent.

U^2 test is a non-parametric test used to compare the empirical distribution of a sample with a theoretical distribution. The test statistic is based on the differences between the empirical and theoretical characteristic functions. It is particularly sensitive to differences in the shape of the distribution.

In this paper, K-S test is used to compare the original data to the generated (imputed or augmented) data. This evaluation method has been used in

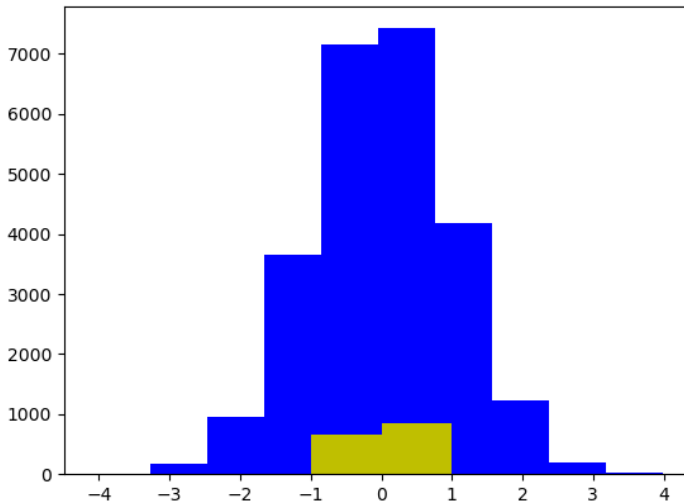


Fig. 10: Distribution of the feature 1 of the dataset used in case 1.

this work since it is known as a reference in statistics to compare two data distributions. This proposed $xGEWFI$ method principle could also apply to any other metric.

Some quantified comparisons between the K-S test and the $xGEWFI$ method are shown in last section, especially in Fig. 11, Table 1, and 2.

The results show that the evaluation method used (the K-S test) will give an evaluation of the similarity between the original data and the generated data for each feature. The $xGEWFI$ method weights the results of the K-S test using the importance of each feature given by an RF algorithm. We have better feature errors and global error evaluations when those are adjusted according to their importance. In other words, a feature error should be minimized when this feature is less valuable to differentiate the data. Conversely, a feature error should be maximized when this feature is handy to differentiate the data. Having highly differentiable data makes the better classification and regression processes.

The values returned by this $xGEWFI$ metric cannot be directly compared with the other metrics like the K-S test. They do not have the same magnitude. It cannot be seen as a disadvantage since most metric magnitudes (like the K-S test) are neither compatible with others.

Finally, a traditional method like the K-S test is explainable since it is a statistical formula. Although, when we use it combined with a machine learning algorithm like RF, the results are less explainable. We must use some

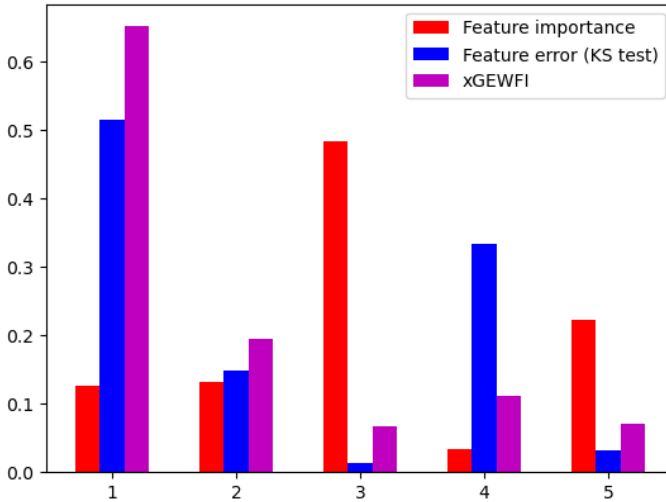


Fig. 11: Presentation of the 3 important parameters in the $xGEWFI$ formula.

strategies to explain why this metric gives those results. The $xGEWFI$ method is designed for explainability.

4 Discussions

A novel explainable metric for data imputation and data augmentation has been developed in this paper. To validate this metric, a complete method to detect and correct outliers, impute data and augment data has been made. It has been tested on two datasets of 25000 rows and 5 features. One dataset is a classification problem, and the other is a regression problem. Both datasets are fully reproducible using the `make_regression()` and `make_classification()` functions available in the `scikit-learn` framework.

The two significant advantages of the $xGEWFI$ metric are 1. The feature importance weights each feature error (composing the global error). 2. The result is explainable, as mentioned in section 3.4.

The results are presented in two different cases. The first case was a data imputation problem for a regression-oriented dataset. The K-S test gives the same error to all the features (Fig. 3). The importance of the features is unequally distributed, as shown in Fig. 4. It is obvious that feature 1 is very significant, and feature 4 is insignificant. Since the process gives similar K-S test errors for all features, the results of the $xGEWFI$ algorithm (5) will be very similar to the feature importance (Fig. 3). Although, the $xGEWFI$ made a huge improvement, since it is the Fig. 3 and Fig. 5 that must be compared.

The second case was a problem of data augmentation in a classification-oriented dataset. The K-S test errors differed for all features (Fig. 6). The importance of the features was also different, as presented in Fig. 7. The results are shown in Fig. 8. It represents the K-S test values weighted by the feature importances.

There is only one situation where the $xGEWFI$ method would not significantly improve the metric. That is when the feature importances are equally distributed between them. Nonetheless, this situation is the exception and not the rule. Hence, the improvement of the $xGEWFI$ method is significant in most situations.

This novel $xGEWFI$ method exploits the K-S test, which has some limitations. It assumes that the distributions being compared are independent and continuous and that their cumulative distribution functions fully specify them. In practice, however, the independence assumption may not always hold, and the distributions may be discrete or have unknown or incomplete specifications. K-S test can also be sensitive to sample size, particularly for small samples. As the sample size increases, the test becomes more potent in detecting differences between the distributions but may also become more susceptible to Type I errors. In statistical hypothesis testing, a Type I error refers to the rejection of a true null hypothesis, while a Type II error occurs when a false null hypothesis is not rejected.

Limitations of imputation techniques lead to underestimating standard errors and, thus, overestimating test statistics. $xGEWFI$'s users must be aware of that inevitable bias. Nevertheless, $xGEWFI$ aims to significantly reduce error by weighting the importance of each feature. In the end, the user has a better error estimation than if the error would not be weighted by the feature importance.

This novel method helps evaluate the performance of a data imputation or augmentation process. A data scientist could try different meta-parameters on a data imputation algorithm (like KNNImpute or GAIN), on a data augmentation algorithm (like SMOTE or GAN), or on an outlier detection algorithm (IQR or RANSAC). The $xGEWFI$ metric will give him an explainable evaluation, weighted on the importance of the feature specific to his dataset. In the end, this research succeeds in improving data imputation and data augmentation metrics developed in an explainable manner.

5 Conclusion

This paper proposed a novel explainable metric to evaluate the performance of any data imputation and data augmentation method. Using a classification-oriented dataset, or a regression-oriented dataset, this research implements a whole process to 1. Detect outliers and replace them with null values. 2. Impute missing data, and 3. Augment the data. At the end of this process, the proposed $xGEWFI$ algorithm computes the error based on the K-S test, a well-known statistical reference aiming to compare two feature's distribution

(the original one and the generated one). The results of this K-S Test are then multiplied by the importance of the respective features generated by an RF algorithm, resulting in a weighted error more representative than the non-weighted error. This result, the $xGEWFI$ metric, allows the data scientist to evaluate better the impact of the meta-parameters used in the imputation and augmentation process and on his datasets.

This method can be improved in the future by replacing the classic K-S test with another metric. K-S test is a well-known classic method to compare the distribution of two variables. In this context, it has been utilized to test the validity of the imputation and the augmentation process. A more practical test may be found for $xGEWFI$ in the future. In machine learning, the concept of explainability is relatively recent. This new field is proliferating and will keep growing in the following years. New concepts and methods will flourish, and it is very likely that $xGEWFI$ explainability could be improved shortly.

6 Acknowledgement

This work has been supported by the "Cellule d'expertise en robotique et intelligence artificielle" of the Cégep de Trois-Rivières and the Natural Sciences and Engineering Research Council.

7 Statements and Declarations

Fundings This work has been supported by the Natural Sciences and Engineering Research Council.

Conflict of interest The authors confirm there are no conflicts of interest.

Ethical approval The work uses publicly available and non-identifiable information. No ethical approval was needed.

Consent to participate Not applicable, since no human participant was involved in the evaluation of our study.

Consent for publication Not applicable, since all datasets used in this study are released by third parties.

Availability of data and material We used only datasets that are publicly available.

Code availability The code is not published yet. It can be provided on demand.

Authors' contributions JSD : Conceptualization, Methodology, Software, Writing - Original Draft, Software. D.M.: Conceptualization, Methodology, Validation, Resources, Writing - Review & Editing, Supervision, Project

administration, Funding acquisition.

References

- [1] Steele, M., Chaseling, J.: Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions **35**(4), 1067–1075 (2006). <https://doi.org/10.1080/03610910600880666>. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/03610910600880666>. Accessed 2022-06-30
- [2] Elmore, K.L.: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts **20**(5), 789–795 (2005). <https://doi.org/10.1175/WAF884.1>. Publisher: American Meteorological Society Section: Weather and Forecasting. Accessed 2022-06-30
- [3] Massey, F.J.: The kolmogorov-smirnov test for goodness of fit **46**(253), 68–78 (1951). <https://doi.org/10.1080/01621459.1951.10500769>. Publisher: Taylor & Francis
- [4] Berger, V.W., Zhou, Y.: Kolmogorov–smirnov test: Overview. In: Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd, ??? (2014). <https://doi.org/10.1002/9781118445112.stat06558>
- [5] Pfeifer, B., Holzinger, A., Schimek, M.G.: Robust random forest-based all-relevant feature ranks for trustworthy ai. *Studies in Health Technology and Informatics* **294**, 137–138 (2022)
- [6] Biau, G., Scornet, E.: A random forest guided tour **25**(2), 197–227 (2016). <https://doi.org/10.1007/s11749-016-0481-7>. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 2 Publisher: Springer Berlin Heidelberg. Accessed 2021-03-23
- [7] Lv, J., Wang, Y., Liang, X., Yao, Y., Ma, T., Guan, Q.: Simulating urban expansion by incorporating an integrated gravitational field model into a demand-driven random forest-cellular automata model **109**, 103044 (2021). <https://doi.org/10.1016/j.cities.2020.103044>. Accessed 2021-03-29
- [8] Vinutha, H.P., Poornima, B., Sagar, B.M.: Detection of outliers using interquartile range technique from intrusion dataset, 511–518 (2018). https://doi.org/10.1007/978-981-10-7563-6_53
- [9] Sánchez-González, J.-M., Rocha-de-Lossada, C., Flikier, D.: Median absolute error and interquartile range as criteria of success against the percentage of eyes within a refractive target in IOL surgery **46**(10),

- 1441 (2020). <https://doi.org/10.1097/j.jcrs.0000000000000248>. Accessed 2022-01-04
- [10] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* **24**(6), 381–395 (1981)
- [11] Tutz, G., Ramzan, S.: Improved methods for the imputation of missing data by nearest neighbor methods **90**, 84–99 (2015). <https://doi.org/10.1016/j.csda.2015.04.009>. Accessed 2022-03-11
- [12] de Silva, H., Perera, A.S.: Missing data imputation using evolutionary k- nearest neighbor algorithm for gene expression data. In: 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 141–146 (2016). <https://doi.org/10.1109/ICTER.2016.7829911>. ISSN: 2472-7598
- [13] Wang, Y., Li, D., Li, X., Yang, M.: PC-GAIN: Pseudo-label conditional generative adversarial imputation networks for incomplete data **141**, 395–403 (2021). <https://doi.org/10.1016/j.neunet.2021.05.033>. Accessed 2022-01-05
- [14] Popolizio, M., Amato, A., Politi, T., Calienno, R., Di Lecce, V.: Missing data imputation in meteorological datasets with the GAIN method. In: 2021 IEEE International Workshop on Metrology for Industry 4.0 IoT (MetroInd4.0 IoT), pp. 556–560 (2021). <https://doi.org/10.1109/MetroInd4.0IoT51437.2021.9488451>
- [15] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique **16**(1), 321–357 (2002)
- [16] Han, B., Jia, S., Liu, G., Wang, J.: Imbalanced fault classification of bearing via wasserstein generative adversarial networks with gradient penalty. *Shock and Vibration*, 1–14 (2020)
- [17] Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning **6**(1), 60 (2019). <https://doi.org/10.1186/s40537-019-0197-0>. Accessed 2022-01-02
- [18] Hasanin, T., Khoshgoftaar, T.M., Leevy, J.L., Bauder, R.A.: Severely imbalanced big data challenges: investigating data sampling approaches **6**(1), 107 (2019). <https://doi.org/10.1186/s40537-019-0274-4>. Accessed 2022-03-11
- [19] Guo, S., Liu, Y., Chen, R., Sun, X., Wang, X.: Improved SMOTE algorithm to deal with imbalanced activity classes in smart homes **50**(2), 1503–1526 (2019). <https://doi.org/10.1007/s11063-018-9940-3>. Accessed

2022-03-11

- [20] Veugen, T., Kamphorst, B., van de L’Isle, N., van Egmond, M.B.: Privacy-preserving coupling of vertically-partitioned databases and subsequent training with gradient descent, 38–51 (2021). https://doi.org/10.1007/978-3-030-78086-9_3
- [21] Guedj, B., Srinivasa Desikan, B.: Kernel-based ensemble learning in python **11**(2), 63 (2020). <https://doi.org/10.3390/info11020063>. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Accessed 2022-01-03
- [22] III, D.L.W.: The interquartile range: Theory and estimation - ProQuest (2005). <https://www.proquest.com/openview/8449e263bd9f96a22e0348e6abdeb5a9/1?pq-origsite=gscholar&cbl=18750&diss=y>