Saifullah Khan / Int.J.Data.Sci. & Big Data Anal. 3(2) (2023) 51-58
https://doi.org/10.51483/IJDSBDA.3.2.2023.51-58

ISSN: 2710-2599

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Data Science and Big Data Analytics

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper **Open Access**

Seasonal Mean Imputation Algorithm

Saifullah Khan[1*]

[1]Ph.D. Scholar, COMSATS University Islamabad, Attock Campus, Pakistan. E-mail: saifullahedu0@gmail.com

Article Info

Volume 3, Issue 2, November 2023
Received : 20 July 2023
Accepted : 22 October 2023
Published : 05 November 2023
doi: 10.51483/IJDSBDA.3.2.2023.51-58

Abstract

This invention introduces a novel data imputation algorithm named as Seasonal Mean Imputation (SMI), designed to address the challenge of dealing with missing values in data preprocessing stage for tasks related to data science or Machine Learning (ML) implementation. In contrast to conventional Mean Imputation (MI) technique that involves filling of all the missing values in a dataset with only mean of the whole data, this novel approach improves upon by seasonally imputing or filling the missing values pertinent to the seasonality of the data such that original data's seasonality pattern is considered. The seasonality is a mandatory part of all the ML implementations because ML algorithms' core purpose is to learn the patterns in data and predict future data according to the past data patterns. The goal of this algorithm is to improve the overall predictive accuracy of the ML models with a similar complexity cost as incurred by the traditional MI technique.

Keywords: *Seasonal mean imputation, Mean imputation, Seasonal mean imputation versus mean imputation, Data imputation techniques, Seasonal mean, Data filling, SMI*

1. Introduction

The imputation or filling of data is one of the most important as part of the preprocessing stage of a Machine Learning project (Padgett *et al.*, 2014). The importance of preprocessing stage can be realized by the fact that 80% time of such a project is taken by the preprocessing stage (Press, 2016). The imputation of missing values has significance due to the fact that most Machine Learning models do not support a data with missing values in it (Kumar, 2020). Furthermore, time-series data's missing values affects the continuity of series since date and time will also be missing (Pratama *et al.*, 2016). There are various techniques used for the filling of missing values of data. Such as Mean imputation (Aljuaid and Sasi, 2016), which takes the whole data's mean value and imputes the same single value at every missing place found in the data. A similar, Mode imputation method is also employed which considers mode of the data for said purpose. In cases, where the less than 5% of data has missing values in it, imputation may not be required and the missing parts can just be deleted from the data (Montelpare, 2020). However, it all depends upon the resultant accuracy, application

* *Corresponding author: Saifullah Khan, Ph.D. Scholar, COMSATS University Islamabad, Attock Campus, Pakistan. E-mail: saifullahedu0@gmail.com*

and objectives. For a very high missing percentage of data, for example a 25% of missing data may require the implementation of machine learning algorithms themselves to predict and impute the missing values. K-Nearest Neighbors (KNN) has been implemented in such scenarios (Sessa and Syed, 2016). Algorithms like Decision Tree and others are also applied for such purpose (Gavankar and Sawarkar, 2015). However, the process of such imputation by machine learning algorithms may increase the complexity, consequently the time spent at preprocessing stage and computational processing time is also increased compared to using techniques that use statistics as there foundation, such as the mean/mode imputation. It is to be noted that such techniques do not consider the patterns in the data set, which the machine learning imputation methods do.

This paper introduces a novel algorithm known as Seasonal Mean Imputation, designed to address the challenge of incorporating seasonal patterns into data analysis without relying on complex machine learning techniques. Unlike conventional mean imputation methods, Seasonal Mean Imputation not only considers seasonal patterns but also maintains a comparable level of complexity to traditional statistical techniques. This algorithm represents an enhancement over the standard Mean Imputation approach, offering data scientists a flexible option to achieve improved results without the need for intricate machine learning pipelines. However, it is important to note that Seasonal Mean Imputation is not intended as a substitute for machine learning algorithms. Instead, its purpose is to provide data scientists with a streamlined alternative for enhancing results without the direct implementation of complex machine learning work flow. This allows for the utilization of the simplicity associated with traditional mean/mode imputations while still capturing and addressing seasonal variations within the data. The Table 1 shows information of the subject novel algorithm regarding patent oriented specification details.

Table 1: Specifications Table

Subject Code	**1702 Artificial Intelligence**
Specific subject area	Data imputation of tabular datasets in data preprocessing stage for Machine Learning applications.
Industry code	Y: General tagging of new technological developments; general tagging of cross-sectional technologies spanning over several sections of the IPC; technical subjects covered by former USPC cross-reference art collections [XRACs] and digests.
Details of inventors	Saifullah Khan (Email address: saifullahedu0@gmail.com, PhD scholar from COMSATS University Islamabad, Attock Campus, Pakistan)
Dates of invention	Conception: 08 August, 2023 Disclosure: 05 September, 2023
Patent details *(only if patented)*	N/A
Intended use	The creation of this invention is intended to be used by researchers, students and general public. The author has no intention to gain monetarily from this.
Related research article	N/A
Related other sources: (datasets, software, diagrams, plans, etc.)	N/A

2. Value of the Invention

- This algorithm is useful as because it helps in filling the missing values of data since various ML algorithms

has mandatory requirement that the data is complete and not missing any data points in between. The invention also employs method that keeps its computational complexity low.

- This algorithm is advantageous for individuals working in the fields of data science, machine learning engineering, research, and students who engage in data science and analytics or utilize machine learning algorithms for data analysis.
- Instead of employing Mean equation as in Seasonal Mean Imputation, the Mode or Median of the data can also be used with this invention's seasonality part for imputation purposes.

3. Invention Description

The novel approach, SMI, which is a data imputation technique (Donders *et al.*, 2006), aims to filling up the missing values in a dataset at data preprocessing stage of tabular data for onwards machine learning or data analytics applications.

The SMI algorithms follow a number of set of rules for some scenarios that may take place in a given tabular dataset. To understand the implementation of this algorithm, calculated examples are necessary to be shown.

In Figure 1, a practical working example is shown. SMI is used on the data with one missing point, and there's one non-missing data point both above and below it. This creates a symmetrical scenario where the number of non-missing points above and below considered separately, are the same as the number of non-missing points in between. The data in the figure is just a simple column of five sequential data points to illustrate this situation.

In such a case, an average of the above and below data points are taken. It is to be noted that the above example can be extended to be applied for any *X* number of missing data points. For example, if ten missing rows were there, the algorithm will search for ten non-missing data points above and below these missing data points and take average and impute the same average in all the ten missing data points.

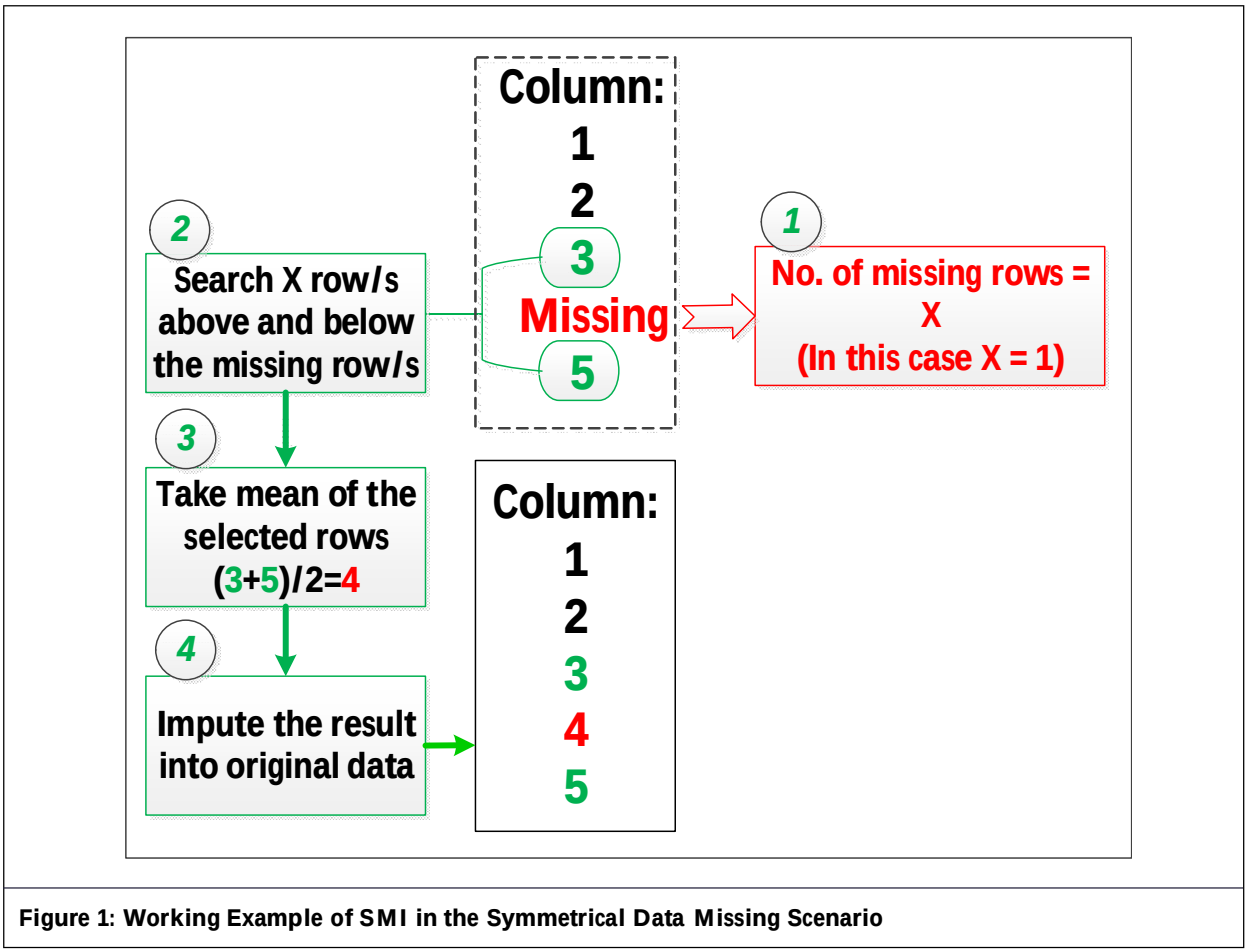

Figure 1: Working Example of SMI in the Symmetrical Data Missing Scenario

Figure 2, shows the scenario for the algorithm where it encounters an un-symmetrical case. Data set of seven data points present from *1* to *7* in sequence for convenience is considered as an example in this case. The two missing values of *5* and *6* has two non-missing data points above it but below them is only one data point of *7*. The way to impute missing values is that we suppose the above data points from these missing data points to be *Xa* and below as *Xb.* The weighted average of *Xa* data points is taken. In this example, total *Xa* and *Xb* data points are three and *Xa* has two data points or (2 / 3 of the total contribution, so 0.67 is multiplied with the average of *Xa*), similarly if *Xb* would have more or less number of data points then accordingly the weighted average would change based on contribution. In this case, 33% of contribution is being made by *Xb* (1 / 3 with 1 being the number of data points of *Xb* and 3 being the total data points belonging to both *Xa* and *Xb*). These two weighted averages are then added to impute the first missing data point.

To fill the next data point, the algorithm notices that the symmetrical scenario is applicable since only one data point is missing surrounded by non-missing data point above and below it. Remember, we filled the first missing data point and are considering its value in this next step. So, similar to Figure 1, the simple average of the symmetrical case is taken and imputed accordingly.

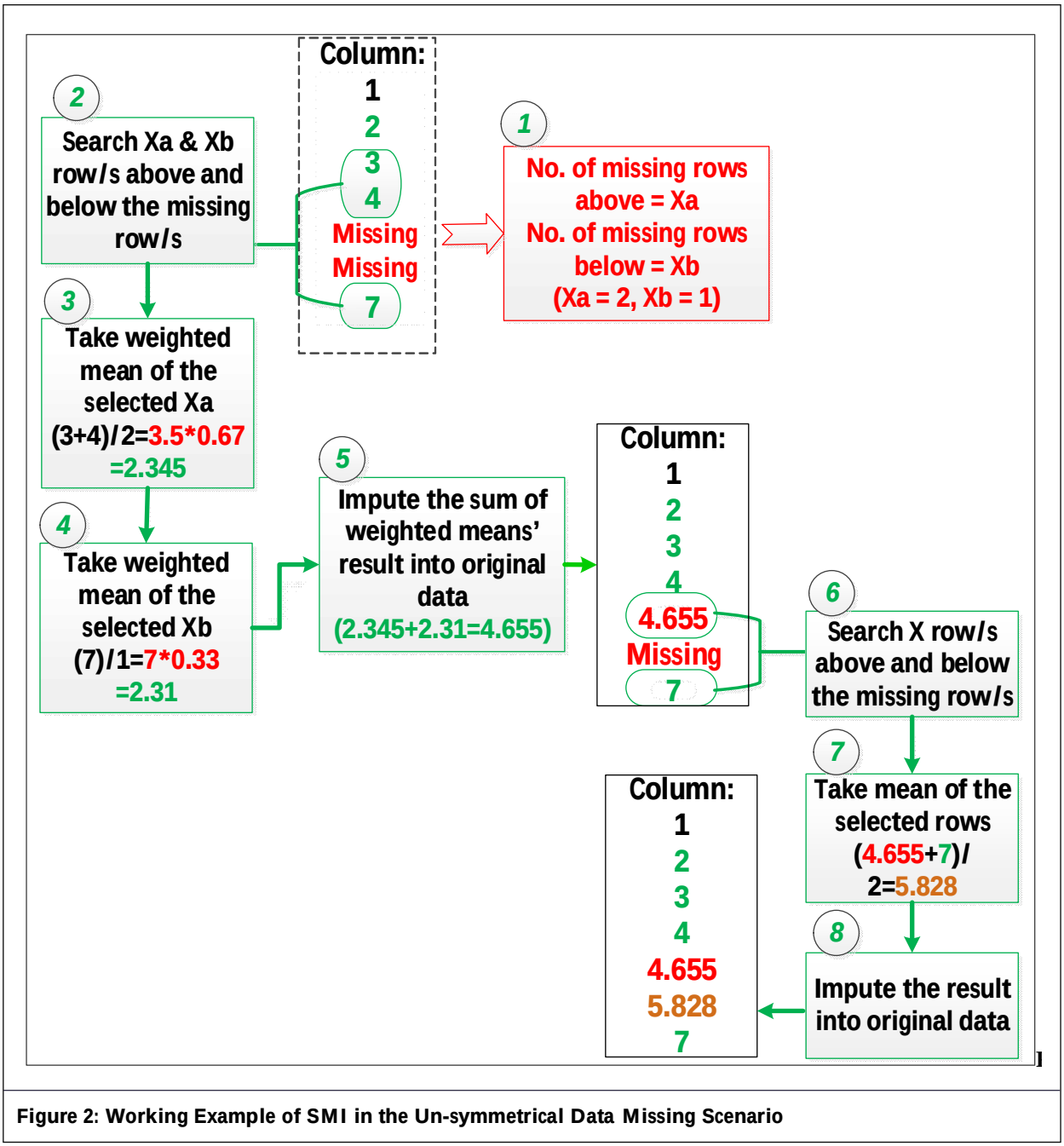

Figure 2: Working Example of SMI in the Un-symmetrical Data Missing Scenario

If Figure 3 is observed, there is a scenario given where missing data points named as *Missing 1*, *Missing 2*, and *Missing 3* are present with three non-missing data points above it and two missing data points below it. Firstly, the *Missing 1* is imputed by the method of taking weighted average of three non-missing data points immediately above the group of missing data points (from *Missing 1* to *3*), and two non-missing data points immediately below the group of missing values. Above ones are supposed to be equal to *Xa* while below ones are considered as *Xb*. Weighted average with weights of 0.6 and 0.4 is multiplied with *Xa* and *Xb*, respectively. Because *Xa* is contributing three data points out of total 5 data points when *Xa* and *Xb* are combined (3/5 or 60%) while *Xb* contributes two data points (2/5 or 40%). Then, according to Figure 3, the calculation for *Missing 1* is taken place and computed value imputed. Next, the calculation is made for *Missing 2* as shown in the figure. Notice that now when we have *Missing 1* filled, two non-missing rows above and below the two missing rows remains, meaning we are dealing with a symmetrical case. So, simple average is taken with *0.5* weights for both *Xa* and *Xb*. Do note that the answer to the *Missing 2* data point is same if a simple average of the four data points was considered collectively (*(4+5.2+8+9)/4)*), but the weighted method is written in

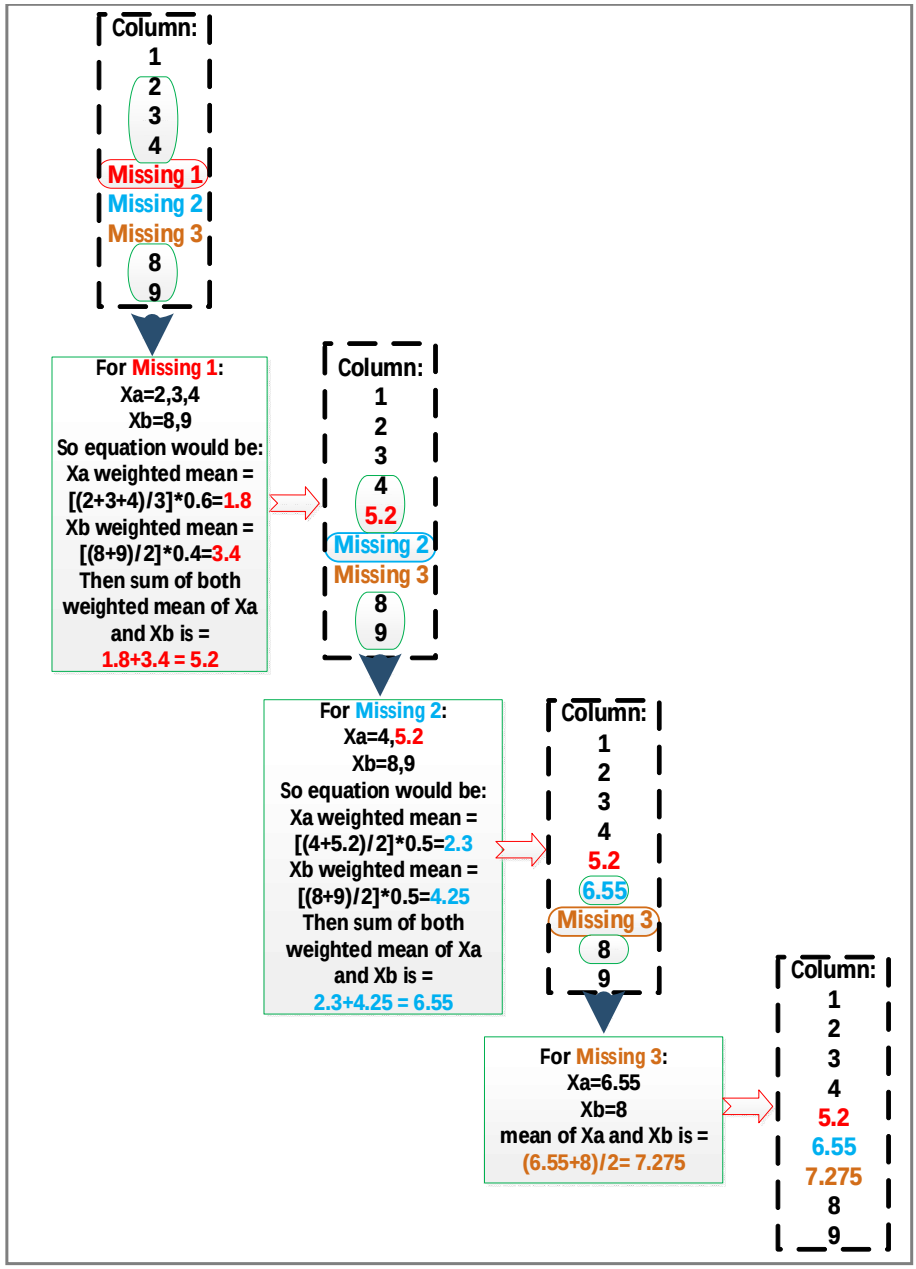

Figure 3: Another Example of SMI in the Un-symmetrical Data Missing Scenario

Figure 3 just for convenience. Similarly, for *Missing 3* data point, a symmetrical case is observed and simple average is taken to impute the value.

4. Background

The state of the art, and the method upon which this invention improves, is the Mean Imputation (MI) method. The MI technique only takes the mean of all data point values' present in the data set and impute or fills the missing values with this mean value. By inheritance, this method does not consider any seasonal pattern of the data, since every empirical data (recorded and not randomly formulated) has some patterns. For example, the power generated data for a wind turbine will have highs and lows of generation with respect to wind velocity and probably higher at certain time of year or day depending upon the area. This implies that power generation data has some patterns. Now, since ML algorithms make future predictions based on the past data available by learning the patterns involved, this type of pattern in the data won't be available in those data points where only the static mean of the entire dataset is imputed for the missing values, neglecting the actual patterns. Simply put, mean values as is in the whole data can actually increase bias in the data (Emmanuel *et al.*, 2021).

In Table 2, if a basic dataset with a single column containing numbers from 1 to 10 in sequence is supposed and the imputed values using both SMI and MI are evaluated, an intuitive grasp of the concept can be formed in terms of the difference between the two techniques used with reference of how close each imputation method represents the original data's pattern by looking at the Root Mean Square Error (RMSE). The lower the RMSE is to 0, the better the missing values of the original data reproduced.

$$RMSE = \sqrt{\frac{\sum_1^N (\boldsymbol{Y} - \overline{\boldsymbol{Y}})^2}{\boldsymbol{N}}} \quad \text{...(1)}$$

where $\boldsymbol{Y}$ is actual data, $\overline{\boldsymbol{Y}}$ is predicted data by MI or SMI and $\boldsymbol{N}$ is total number of data points in dataset.

Table 2 displays the outcomes of the imputation process, with missing values visually highlighted in red shaded cells. When examining the imputed results for the SMI column, it becomes apparent that this method reproduces the original data's values more faithfully in comparison to MI's imputation approach, which is evident by the evaluation results of Root Mean Square Error (RMSE), as depicted in the Table 2 for both SMI and MI. RMSE value is closer to zero for SMI by a large margin highlighted by the improvement factor of 14.91, which emphasizes the superior accuracy of SMI in representing the underlying data structure.

Table 2: Intuitive Example of Effect of SMI and MI on a Data's Imputation

Original Data	Supposed Data with Missing Values	Imputed by SMI	Imputed by MI
1	1	1	1
2	2	2	2
3		3	4.83
4	4	4	4
5	5	5	5
6		6	4.83
7	7	7	7
8		7.66	4.83
9		8.83	4.83
10	10	10	10
RMSE		*0.12*	*1.79*

5. Application Potential

The application potential for subject invention is its massive utility for beginner students in Artificial Intelligence to amateur researchers and professionals, since data imputation is one of the most, and often times, a mandatory step in the data preprocessing stage for ML applications involving tabular data (Padgett *et al.*, 2014; Emmanuel *et al.*, 2021).

The limitation of the subject algorithm to be noted is that it uses the Mean equation as a core functionality, which is a very simple and computationally inexpensive operation relative to other regression based imputation methods which utilizes ML algorithms, though with an obvious compromise for computational expense (Emmanuel *et al.*, 2021). SMI is applicable in situations where a balance between achieving satisfactory accuracy and maintaining reasonable computational time and resource costs is required.

6. Conclusion

In summary, the Seasonal Mean Imputation (SMI) algorithm proposed in this paper offers a novel solution for addressing missing values in data preprocessing for machine learning. By incorporating seasonal patterns, SMI aims to enhance predictive accuracy while maintaining computational efficiency similar to traditional MI techniques.

SMI's value is two-fold: it efficiently fills missing data, a critical requirement for various machine learning algorithms, and it is applicable across a spectrum of educational, research, and professional ML projects. The algorithm's flexibility, exemplified by its compatibility with alternative statistical measures, amplifies its resourcefulness. Its broad application potential, catering to beginners and experts in artificial intelligence, underscores its utility in tabular data preprocessing. Despite relying on the Mean equation for simplicity, SMI achieves a delicate equilibrium between accuracy and computational efficiency, rendering it valuable in resource-sensitive scenarios.

Detailed instances showcase how SMI can be applied in diverse scenarios, enabling researchers to implement the method in the programming languages or Graphical User Interface (GUI) based data science software of choice. An intuitive comparative analysis in Table 1 underscores SMI's performance in replicating original data patterns, as evidenced by a lower Root Mean Square Error (RMSE). However, more studies on the effectiveness of SMI is encouraged.

In conclusion, Seasonal Mean Imputation stands out as an efficient and practical solution for handling missing values in diverse datasets, providing a valuable tool for data scientists and machine learning practitioners.

Acknowledgment

This research is dedicated to the author's parents and siblings for their unwavering support and encouragement.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

Aljuaid, T. and Sasi, S. (2016). Proper Imputation Techniques for Missing Values in Data Sets. In2016 International Conference on Data Science and Engineering (ICDSE), August 23, 1-5, IEEE.

Donders, A.R., Van Der Heijden, G.J., Stijnen, T. and Moons, K.G. (2006). A Gentle Introduction to Imputation of Missing Values. *Journal of Clinical Epidemiology*, 59(10), 1087-1091.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O. (2021). A Survey on Missing Data in Machine Learning. *Journal of Big Data*, 8(1), 1-37.

Gavankar, S. and Sawarkar, S. (2015). Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility. In2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS), December 2, 122-126, IEEE.

Kumar, S. (2020). 7 Ways to Handle Missing Values in Machine Learning. *Towards Data Science*. [Online]. Available: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e#. [Accessed 2023].

Montelpare, W.J. (2020). Working with Missing Data. Pressbooks Library. [Online]. Available: https://pressbooks.library.upei.ca/montelpare/chapter/working-with-missing-data/. [Accessed 2023].

Padgett, C.R., Skilbeck, C.E. and Summers, M.J. (2014). Missing Data: The Importance and Impact of Missing Data from Clinical Research. *Brain Impairment*, 15(1), 1-9.

Pratama, I., Permanasari, A.E., Ardiyanto, I. and Indrayani. R. (2016). A Review of Missing Values Handling Methods on Time-series Data. In 2016 International Conference on Information Technology Systems and Innovation (ICITSI), October 24, 1-6, IEEE.

Press, G. (2016). Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. [Online]. Available: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/. [Accessed 2023].

Sessa, J. and Syed, D. (2016). Techniques to Deal with Missing Data. In 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), December 6, 1-4, IEEE.

Cite this article as: Saifullah Khan (2023). Seasonal Mean Imputation Algorithm. *International Journal of Data Science and Big Data Analytics*, 3(2), 51-58. doi: 10.51483/IJDSBDA.3.2.2023.51-58.