Journal of Advanced Research in Data Structures Innovations and Computer Science
Volume 1, Issue 2 - 2025, Pg. No. 8-13
Peer Reviewed Journal

Research Article

Effective Machine Learning Techniques for Handling missing Data

Kapil Prashar[1], Ankush Rahuvanshi[2]

[1]Professor, Department of Computer Science & Engineering, PCTE Institute of Engineering & Technology, Ludhiana, India
[2]Student, Department of Computer Applications (MCA), PCTE Group of Institutes, Ludhiana, Punjab, India

INFO

Corresponding Author:
Kapil Prashar, Department of Computer Science & Engineering, PCTE Institute of Engineering & Technology, Ludhiana, India
E-mail Id:
kapilprashar@pcte.edu.in

How to cite this article:
Prashar K, Rahuvanshi A. Effective Machine Learning Techniques for Handling missing Data. *J Adv Res Data Struct Innov Comput Sci* 2025; 1(2): 8-13.

Date of Submission: 2025-08-18
Date of Acceptance: 2025-09-26

ABSTRACT

Missing data in machine learning is a significant challenge, impacting predictive models' performance. It can be caused by errors in data collection, incomplete responses, or system failures. Incorrect handling can lead to biased or inaccurate predictions. This paper explores various imputation methods, including mean imputation, median imputation, random imputation, and end-of-distribution imputation. Each method has specific applications based on the dataset's nature and missing information. Mean imputation, which replaces missing values with the average of available data, is most effective when the data follows a normal distribution. The mean is a reliable measure of central tendency in symmetric distributions, but it may not be suitable for skewed data due to its less sensitive nature to outliers. Median imputation, the middle value in a sorted dataset, is ideal for skewed data distributions. Random imputation, a more flexible technique, replaces missing values with randomly selected values but may require more computational resources, especially in large datasets. End-of-distribution imputation fills missing values with the lowest or highest value. This paper emphasises the significance of hyperparameter tuning in machine learning models, specifically GridSearchCV. This tool systematically explores various model parameter combinations to find the best-performing set, preventing overfitting and ensuring model generalisation to unseen data. It is particularly useful for complex models requiring fine-tuning. The paper emphasises the importance of combining robust imputation techniques with hyperparameter optimisation methods for reliable machine learning models, enhancing predictive power and reliability.

Keywords: Imputation, Cross-Validation, Hyperparameter Tuning, Median Imputation, End of Distribution Imputation

Introduction

Machine learning relies on data to build predictive models, and the accuracy of these models depends on the quality and completeness of the data they are trained upon. It has been observed in the recent studies that while collectingdata, the key challenge for researchers is to compile a complete dataset. The missing values in the data collected from diverse sources like surveys, sensors, or transactional logs can lead to inaccurate predictions and unreliable results. Addressing the concern and then filling those missing values is crucial in the data preprocessing phase of any machine learning process. The machine learning process involves the steps of data collection, data preprocessing, compilation of training and testing datasets, selection of algorithms required for training the machine learning model, testing the trained model with the testing dataset to check the accuracy of prediction, and then training the model again if required.[1-5]

The researchers have compiled an original dataset of 300 rows and 5 attributes, namely Student_ID, CGPA, Placement_Status, Internship_Count and Salary_Offered. The target attribute is Salary_Offered, which is predicted by the machine learning model after training. The researchers have emphasised the importance of handling missing data in this work to ensure the robustness and accuracy of models. Different types of missing values require tailored imputation approaches, such as mean, median, random imputation, and end-of-distribution imputation. Mean imputation is suitable in the scenario if data is normally distributed, which means it results in a bell-shaped curve when plotted on a graph. The median is the best measure of central tendency to fill the missing value, as after arranging the values in either ascending order or descending order, the central value will replace the missing value. The objective of the authors is to fill missing values in the database in such a way that maximum accuracy is achieved in predicting results. The mean method replaces missing numerical values with the average of the available data in the column. This method is suitable when the column has less than 5% missing values and the values that are missing at random.[6-10]

The median method is another approach used to impute missing numerical values in a dataset. This technique is applied if the data is missing at random, the percentage of missing values is less than 5%, and the distribution of the data is skewed. Random imputation is a technique that works for both numerical and categorical data. In this method, missing values are filled with a random value selected from the existing data in the same column.

This method is used to impute numerical missing values, especially when the data is not missing at random. Specific formulas are applied depending on the data distribution, like (Mean + 3σ) or (Mean - 3σ), where σ is the standard deviation.[11,12]

GridSearchCV is a technique used in machine learning for hyperparameter tuning. Its purpose is to identify the best combination of hyperparameters for a model by systematically testing all possibilities. The heatmap shows the correlations between CGPA, Salary_Offered, and Internship_Count. Positive correlations indicate a moderate positive relationship, while negative correlations indicate a strong negative relationship. The heatmap shows a moderate positive correlation between CGPA and salary offered, suggesting higher CGPA leads to higher salary offers, but not a strong relationship. However, the heatmap also shows a weak correlation with internship count, suggesting that internship experience has limited direct influence on salary or academic performance. The boxplot displays the distribution of Salary_Offered based on the categorical variable Placement_Status. Key elements include the median, which represents the middle value of the salary distribution for each placement status group, the interquartile range, whiskers, and outliers. Placed students are likely to have a higher salary distribution, with a visible gap in median salaries compared to not-placed students. Each method has specific conditions where it works best, like data distribution and the type of missing values. Additionally, techniques like GridSearchCV help optimise model performance by selecting the best hyperparameters.[13-15]

Review of Literature

Recent studies have explored the critical issue of missing data in machine learning, highlighting its impact on model performance and decision-making. Researchers have categorised missing data mechanisms and evaluated both traditional and modern imputation strategies to address these challenges. Traditional methods such as mean, median, and mode imputation are computationally efficient but often inadequate for complex datasets. In contrast, advanced techniques, including k-Nearest Neighbours (KNN) and deep learning-based imputation methods, demonstrate superior accuracy and adaptability in handling nonlinear relationships. Hybrid approaches that integrate conventional statistical methods with machine learning models have been recommended for achieving balanced performance and computational efficiency.[14]

Investigations into anomaly detection under conditions of incomplete data have proposed several strategies for managing missing values. These include mean imputation, Maximum A Posteriori (MAP) imputation, data reduction, marginalisation, and proportional distribution. Comparative analyses across various anomaly detection algorithms—such as Isolation Forest, LODA, and EGMM—have revealed that MAP imputation and proportional distribution often yield higher accuracy, while marginalisation techniques perform well with probabilistic models like EGMM.[15]

Further research has focused on the application of machine learning-based imputation in clinical decision-making. Studies demonstrate that employing algorithms such as KNN, Bayesian networks, and neural networks to handle missing clinical variables significantly enhances predictive accuracy compared to traditional statistical methods. These findings underscore the importance of integrating machine learning-driven imputation into healthcare analytics workflows to improve the reliability and interpretability of clinical predictions.[17]

Comprehensive reviews on missing data mechanisms and patterns emphasise the necessity of addressing incomplete information to prevent biased analyses. Evaluations of imputation techniques such as KNN and iterative random forest-based methods (e.g., missForest) show their effectiveness in managing missing values across different datasets. Empirical results indicate that both methods perform well in restoring data integrity, even when missingness rates are moderate. Such approaches are particularly suitable for structured datasets and have been identified as promising directions for future research in data preprocessing and imputation.[18,19]

In the healthcare domain, several studies have analysed the influence of missing data on predictive modelling for patient outcomes. Experiments involving the exclusion of variables with high missingness and retraining of models reveal that selective removal of incomplete data can enhance model accuracy in predicting major adverse clinical events. These studies conclude that effective management of missing values—whether through imputation or strategic variable exclusion—is essential for optimising performance in clinical machine learning systems and for supporting informed, data-driven patient care decisions.[20-25]

Methodology /Techniques

Univariate Imputation

This method uses the values of only the column containing missing data to impute the missing values. Values from other columns are not considered for imputing the missing values in the target column. By using this method, we can impute two types of missing values:

Numerical Value Imputation

This technique is used to impute numerical values in columns with missing data. Several approaches can be employed for imputing numerical values.

Mean Method

The mean method replaces missing numerical values with the average of the available data in the column. This method is suitable when the column has:

- Less than 5% missing values.
- Values that are missing at random.
- A normal distribution in the column.

Mathematical Formula of Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Where ∑xi is the sum of all the values in the dataset and n is the total number of values in the dataset.

Advantage

- Easy to implement.

Disadvantages

- Changes the shape of the distribution.
- Affected by outliers.

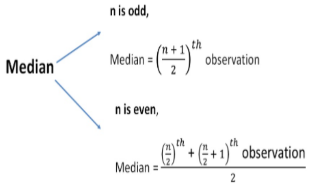
Median Method

The median method is another approach used to impute missing numerical values in a dataset. This technique is applied under the following conditions:

The data is missing at random. The percentage of missing values is less than 5%. The distribution of the data is skewed.

Mathematical Formula for Median

The median is the middle value of a dataset when the numbers are arranged in ascending order. If the dataset has an even.


Median
n is odd,
Median = $\left(\frac{n+1}{2}\right)^{th}$ observation
n is even,
Median = $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$

Advantages

- Easy to implement.

Disadvantages

- Changes the shape of the distribution.
- Affected by outliers.
- Alters covariance and correlation.

Random Imputation

Random imputation is a technique that works for both numerical and categorical data.

In this method, missing values are filled with a random value selected from the existing data in the same column.

Advantages

- Easy to use.
- Does not affect the variations of the data.

Disadvantages

- Consumes more memory for deployment.
- Best suited for linear models only.

End of Distribution Imputation

This method is used to impute numerical missing values, especially when the data is not missing at random. Specific formulas are applied depending on the data distribution: (Mean + 3σ) or (Mean - 3σ)

Where σ is the standard deviation.

When the distribution of our data is skewed, we use IQR approximations:

Q1 – 1.5 IQR

Q3 + 1.5 IQR

Where IOR is Inter Quantile Range and we are calculating it as IQR = Q3(75%) – Q1(25%).

GridSearchCV Method

GridSearchCV is a technique used in machine learning for hyperparameter tuning. Its purpose is to identify the best combination of hyperparameters for a model by systematically testing all possibilities.

Simple Steps to Use GridSearchCV

- **Define the Grid:** Create a list of hyperparameters you want to tune and specify their possible values (the “parameter grid”).
- **Try Combinations:** GridSearchCV evaluates every possible combination from the parameter grid by fitting the model on the training data.
- **Find the Best Combination:** It selects the combination that yields the best performance, such as the highest accuracy score.
- **Use Cross-Validation:** To prevent overfitting, GridSearchCV divides the training data into multiple parts and evaluates the model on each part separately.

The Titanic dataset, a crucial classification benchmark, was optimised using GridSearchCV for hyperparameters in models like logistic regression, random forest, SVM, and KNN. Preprocessing involved handling missing values, encoding categorical data, and scaling numerical features. GridSearchCV tested parameter combinations using cross-validation to prevent overfitting. The best model and parameters improved prediction accuracy, highlighting GridSearchCV’s value in automating hyperparameter tuning and optimising predictive analysis performance.

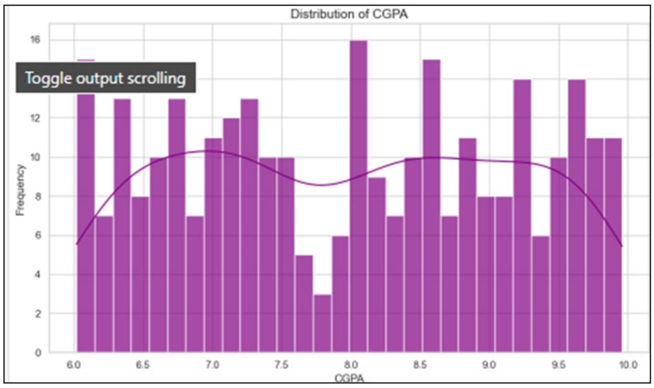


Figure 1.Comparison of Frequency & CGPA of Students

Results and Discussion

As shown in Figure 1, the distribution of the number of students according to their CGPA is displayed. It can be observed that maximum frequency of the students corresponds to 8 CGPA whereas there are 3 students with 7.7 CGPA.

As demonstrated in Figure 2, the maximum salary is offered to 20 students who achieved higher CGPA. There is a positive correlation which can be clearly seen between the number of students who achieved the highest CGPA and the high amount of salary that was offered to those students during placements.

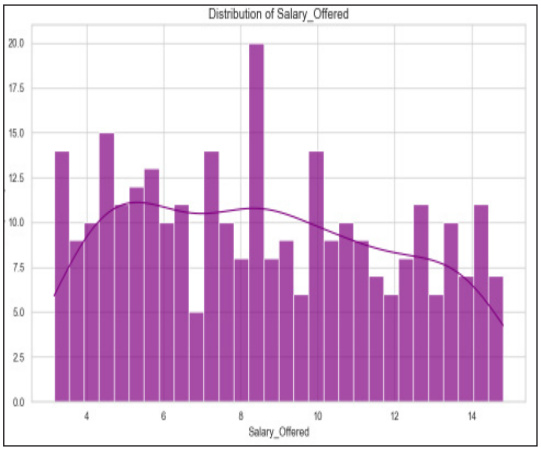


Figure 2.Graph of Salary Offered vs Number of Students

Correlation Heatmap

The heatmap in figure 3 shows the correlations between CGPA, Salary_Offered, and Internship_Count. Positive correlations indicate a moderate positive relationship, while negative correlations indicate a strong negative relationship. In this dataset, no significant negative correlations are present. Weak correlations indicate minimal or no linear relationship, as seen in Internship_Count’s weak correlation with both CGPA and Salary_Offered, suggesting minimal dependency.

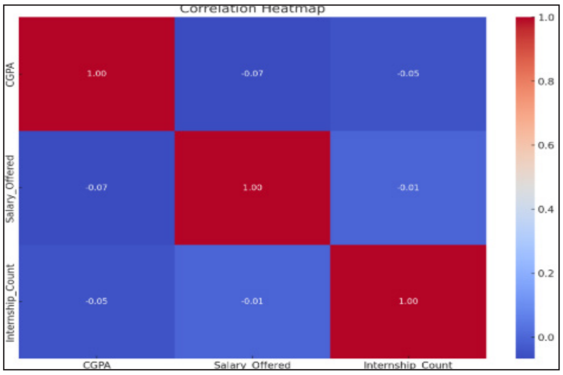


Figure 3.Correlation Heatmap of CGPA,Salary Offered & Internship Count

The boxplot displays the distribution of Salary_Offered based on the categorical variable Placement_Status. Key elements include the median, which represents the middle value of the salary distribution for each placement status group, the interquartile range, whiskers, and outliers. Placed students are likely to have a higher salary distribution, with a visible gap in median salaries compared to not-placed students. Not-placed students may have a limited or negligible salary range if recorded at all. Salary variability among placed students may indicate that while some secure high-paying jobs, others receive relatively lower offers. For not-placed students, variability could be much smaller if any salaries are recorded. Outliers in the Placed category might represent exceptionally high-paying job offers. This boxplot effectively conveys how placement status impacts salary outcomes and helps identify variability, trends, and exceptions within the dataset.

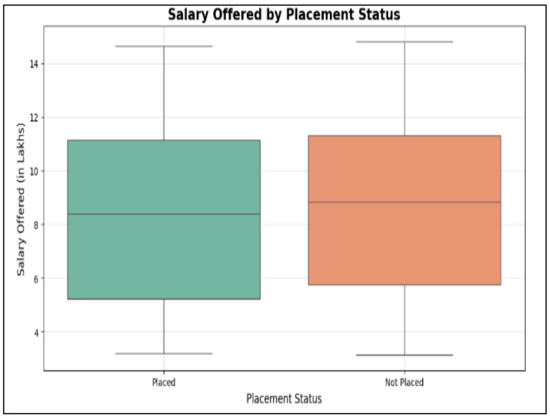


Figure 4.Comparison of Placement Status & Salary Offered

As shown in Figure 4, most students with high CGPA got placed at an average salary package of 9 LPA. The least package offered to the students was 5.5 LPA, and the maximum package offered was around 11.5 LPA. It was also observed that the predicted results of salary offered to the selected students in terms of their placements were highly accurate as compared to the actual results.

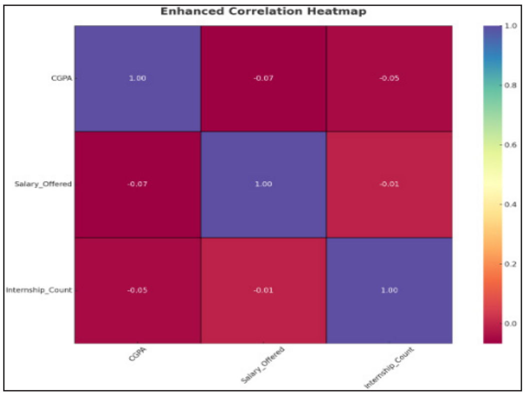


Figure 5.Enhanced Correlation Heatmap of CGPA,Salary Offered,Internship Count

Enhanced Correlation Heatmap

The heatmap in Figure 5 shows a moderate positive correlation between CGPA and salary offered, suggesting higher CGPA leads to higher salary offers, but not a strong relationship. However, the heatmap also shows a weak correlation with internship count, suggesting that internship experience has limited direct influence on salary or academic performance.

Conclusion

Handling missing data is a highly critical aspect of machine learning because missing values can negatively impact the accuracy of models. The authors have discussed various methods to address missing data, such as using the mean, median, random imputation, and end-of-distribution imputation. Each method has specific conditions where it works best, like data distribution and the type of missing values. Additionally, techniques like GridSearchCV help optimise model performance by selecting the best hyperparameters. Selecting the right imputation method and tuning the model ensures better predictions and improved accuracy.

Future Scope

The researchers can carry forward this work in the development of advanced imputation techniques, applying dynamic and context-aware imputation. Also, an insight into the integration with real-time systems can be analysed with exploration of Missing Not at Random (MNAR) Data. There is also an option for evaluation of hybrid techniques and training explainable imputation models, which may result in improvement in GridSearchCV and hyperparameter tuning methods.

References

1. Donders AR, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. Journal of clinical epidemiology. 2006 Oct 1;59(10):1087-91.
2. Graham JW. Missing data analysis: Making it work in the real world. Annual review of psychology. 2009 Jan 10;60(1):549-76.
3. Baraldi AN, Enders CK. An introduction to modern missing data analyses. Journal of school psychology. 2010 Feb 1;48(1):5-37.
4. Aydilek IB, Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences. 2013 Jun 1;233:25-35.
5. Lin J, Li N, Alam MA, Ma Y. Data-driven missing data imputation in cluster monitoring system based on deep neural network. Applied Intelligence. 2020 Mar;50(3):860-77.
6. Choudhury A, Kosorok MR. Missing data imputa-

tion for classification problems. arXiv preprint arXiv:2002.10709. 2020 Feb 25.
7. Schmitt P, Mandel J, Guedj M. A comparison of six methods for missing data imputation. Journal of biometrics & biostatistics. 2015 Jan 1;6(1):1.
8. Hameed WM, Ali NA. Enhancing imputation techniques performance utilizing uncertainty aware predictors and adversarial learning. Periodicals of Engineering and Natural Sciences (PEN). 2022 Jun 29;10(3):350-67.
9. Aljuaid T, Sasi S. Intelligent imputation technique for missing values. In2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) 2016 Sep 21 (pp. 2441-2445). IEEE.
10. Doshi B. Handling Missing Values in Data Mining. Data Cleaning and Preparation Term Paper. 2011.
11. Gupta S, Gupta MK. A survey on different techniques for handling missing values in dataset. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2018;4(1):2456-3307.
12. Grzymala-Busse JW, Goodwin LK, Grzymala-Busse WJ, Zheng X. Handling missing attribute values in preterm birth data sets. InInternational Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing 2005 Aug 31 (pp. 342-351). Berlin, Heidelberg: Springer Berlin Heidelberg.
13. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011 Dec 12;45:1-67.
14. Puri A, Gupta M. Review on missing value imputation techniques in data mining. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2017 Aug;2(7):35-40.
15. Holman R, Glas CA. Modelling non-ignorable missing-data mechanisms with item response theory models. British Journal of Mathematical and Statistical Psychology. 2005 May;58(1):1-7.
16. Pratama I, Permanasari AE, Ardiyanto I, Indrayani R. A review of missing values handling methods on time-series data. In2016 international conference on information technology systems and innovation (ICITSI) 2016 Oct 24 (pp. 1-6). IEEE.
17. Singh S, Prasad J. Estimation of missing values in the data mining and comparison of imputation methods. Mathematical journal of interdisciplinary sciences. 2013 Mar 4;1(2):75-90.
18. Hao G. Efficient training and feature induction in sequential supervised learning. Oregon State University; 2009.
19. Asuncion A, Newman D. UCI machine learning repository [Internet]. 2007 Nov
20. Liu FT, Ting KM, Zhou ZH. Isolation forest. In2008 eighth ieee international conference on data mining 2008 Dec 15 (pp. 413-422). IEEE.
21. Pevný T. Loda: Lightweight on-line detector of anomalies. Machine Learning. 2016 Feb;102(2):275-304.
22. Quinlan JR. Unknown attribute values in induction. InProceedings of the sixth international workshop on machine learning 1989 Jan 1 (pp. 164-168). Morgan Kaufmann.
23. Quinlan JR. C4. 5: programs for machine learning. Elsevier; 2014 Jun 28.
24. Rubin DB. Multiple imputation for nonresponse in surveys John Wiley & Sons.[Google Scholar].
25. Saar-Tsechansky M, Provost F. Handling missing values when applying classification models.