Missing data handling is an important part of data preparation because missing values may degrade the performance of machine learning models.

Dealing with missing data requires careful thought, and there are numerous options to take.

Here are some popular techniques for dealing with missing data:

Deletion of Missing Data:

- Listwise Deletion: Also known as complete-case analysis, this involves removing entire rows with missing values. While simple, this method can lead to a significant loss of data, especially if the missing values are spread across many rows.
- Pairwise Deletion: This method retains observations with complete data for specific analyses. It keeps the available data for each pairwise comparison in the analysis. However, it may lead to different sample sizes for different analyses, which can impact the results.

Imputation Techniques:

- Mean, Median, or Mode Imputation: Replace missing values with the mean, median, or mode of the non-missing values in the same feature. This approach is straightforward but assumes that the data is missing at random (MAR) and can introduce bias if the missingness is not MAR.
- Forward Fill (or Backward Fill): Propagate the last (forward fill) or next (backward fill) valid observation to fill missing values. This method is useful when missing values occur in sequences or time series data.
- Interpolation: Use interpolation methods (e.g., linear, polynomial) to estimate missing values based on the values of neighboring data points. This approach is suitable for

time-series or sequential data.

- K-Nearest Neighbors (KNN) Imputation: Replace missing values with the average of k-nearest neighboring data points. KNN imputation takes into account the similarity between samples to estimate missing values.
- Regression Imputation: Predict missing values using regression models based on other available features. This approach assumes a linear relationship between the missing feature and the other features.

Special Values or Flags:

- Assign a special value or a specific flag to represent missing data, such as “NaN” or “-1.”

Multiple Imputation:

- Generate multiple plausible imputed datasets, each with different estimated values for missing data. Analyze each dataset separately and combine results to obtain more robust estimates and account for uncertainty.

Related Posts:

1. What is Machine Learning ?
2. Types of Machine Learning ?
3. Applications of Machine Learning
4. Data Preprocessing
5. Data Cleaning
6. Feature Scaling
7. Labeled data in Machine learning
8. Difference between Supervised vs Unsupervised vs Reinforcement learning

9. Machine learning algorithms for Big data
10. Difference between Supervised vs Unsupervised vs Reinforcement learning
11. What is training data in Machine learning
12. What is Ordinary Least Squares (OLS) estimation
13. Scalar in Machine Learning
14. Scalars in Loss Functions | Machine Learning
15. Linear Algebra for Machine Learning Practitioners
16. Supervised Learning
17. Top Interview Questions and Answers for Supervised Learning
18. Define machine learning and explain its importance in real-world applications.
19. Differences Between Machine Learning and Artificial Intelligence
20. Machine Learning works on which type of data ?
21. What is target variable and independent variable in machine learning
22. Machine Learning Scope and Limitations
23. What is Regression in Machine learning
24. Statistics and linear algebra for machine learning
25. Finding Machine Learning Datasets
26. What is hypothesis function and testing
27. Explain computer vision with an appropriate example
28. Explain Reinformcement learning with an appropriate exaple
29. Reinforcement Learning Framework
30. Data augmentation
31. Normalizing Data Sets in Machine Learning
32. Machine learning models
33. Unsupervised machine learning
34. Neural Network in Machine Learning
35. Recurrent neural network

36. Support Vector Machines
37. Long short-term memory (LSTM) networks
38. Convolutional neural network
39. How to implement Convolutional neural network in Python
40. What is MNIST ?
41. What does it mean to train a model on a dataset ?
42. Can a textual dataset be used with an openCV?
43. Name some popular machine learning libraries.
44. Introduction to Machine Learning
45. Some real time examples of machine learning