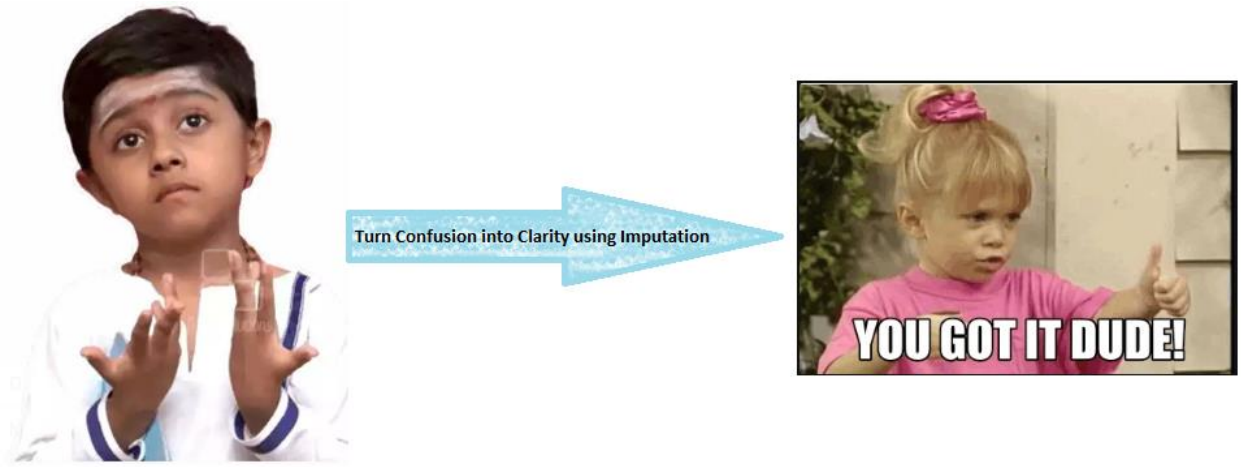
The Hidden Impact of Missing Data

How Imputation Transformed My Machine Learning Models

Contents

Introduction

Missing data is one of the biggest challenges in data science, not doing proper null value cleanup and choosing the wrong imputation method can lead to biased results and inaccurate predictions. In this blog, I am explaining how I learned imputation is an unavoidable step in data preprocessing.



Discussion

In the initial days of my machine learning coding, I preferred to get into the prediction part of the coding and did not have much patience to perform preprocessing. One time I got into a scenario in which my model was giving 95% accuracy with my data. But when I passed the new data, it was giving 65% accuracy. Then I started looking into optimization, PCA, Scaling, and other techniques and finally found that there were some NaN values in the new dataset, so I had to use Null value imputation to fix it. As I was using the Decision Tree Classifier, the model didn't throw any errors.

Here, I am simulating the same with a *sklearn* dataset. Additionally, I am performing different imputation strategies and comparing the results. Also, I am using histogram for visualizing the difference between the updated values.

Steps followed:

<u>Scenario 1:</u>

For testing null value impact on accuracy with simulated data

1. Load *sklearn* library, cancer dataset
2. Create a Decision Tree Classifier Model

3. Look for the accuracy

4. Manually make 100 rows of the column of interest to Null

'mean radius', 'mean texture', 'mean perimeter', 'mean area', 'mean smoothness', 'mean compactness', 'mean concavity', 'mean concave points', 'mean symmetry', and 'mean fractal dimension'

5. Perform the model prediction with this nullified data

6. Look for the accuracy

Scenario 2:

For testing - Comparison of calculated mean radius using different imputation techniques

1. Load sklearn library, cancer dataset

2. Manually make 5 rows of the column of interest (mean radius) to NaN

3. Select one row (# 70) and collect values other related column values

4. These values are used to compare the calculated mean radius value with these values

5. Implement multiple imputations and find the updated mean radius value

6. Check if there is any difference in the calculated value and expected value (based on the similar observed data.

Scenario 3:

Comparison of calculated mean radius using Simple Imputer strategy mean/median using visualization.

1. Load sklearn library, breast cancer dataset

2. Manually make 2 rows of the column of interest (mean radius) to NaN

3. Calculate mean radius values with different imputation methods (Simple Imputer (mean/median)

4. Collect data frames with the updated values

5. Plot the graph of the mean radius with these updated values for comparison

Scenario 4:

Comparison of calculated mean radius using different imputation methods includes Simple Imputer (mean and median imputation) strategy, KNN Imputer, Iterative Imputer/MICE, and Linear regression and perform visualization.

1. Load sklearn library, breast cancer dataset

2. Manually make 2 rows of the column of interest (mean radius) to NaN

3. Calculate mean radius values with different imputation methods (Simple Imputer (mean/median), KNN Imputer, Iterative Imputer, and Linear Regression model

4. Collect data frames with the values from these imputation techniques

5. Plot the graph of the mean radius with these updated values for comparison

Results

Scenario 1:

Null value impact on accuracy with simulated data -

There is 25% difference in accuracy while testing null value impact on accuracy with simulated NaN data. The 25% accuracy decline shows that Null values make a bigger impact on the model performance.

```
Accuracy: 0.9473684210526315

Confusion Matrix:
 [[40  3]
 [ 3 68]]

Updated Accuracy: 0.6929824561403509

Updated Confusion Matrix:
 [[11 32]
 [ 3 68]]
```

Scenario 2:

Comparison of calculated mean radius using different imputation techniques

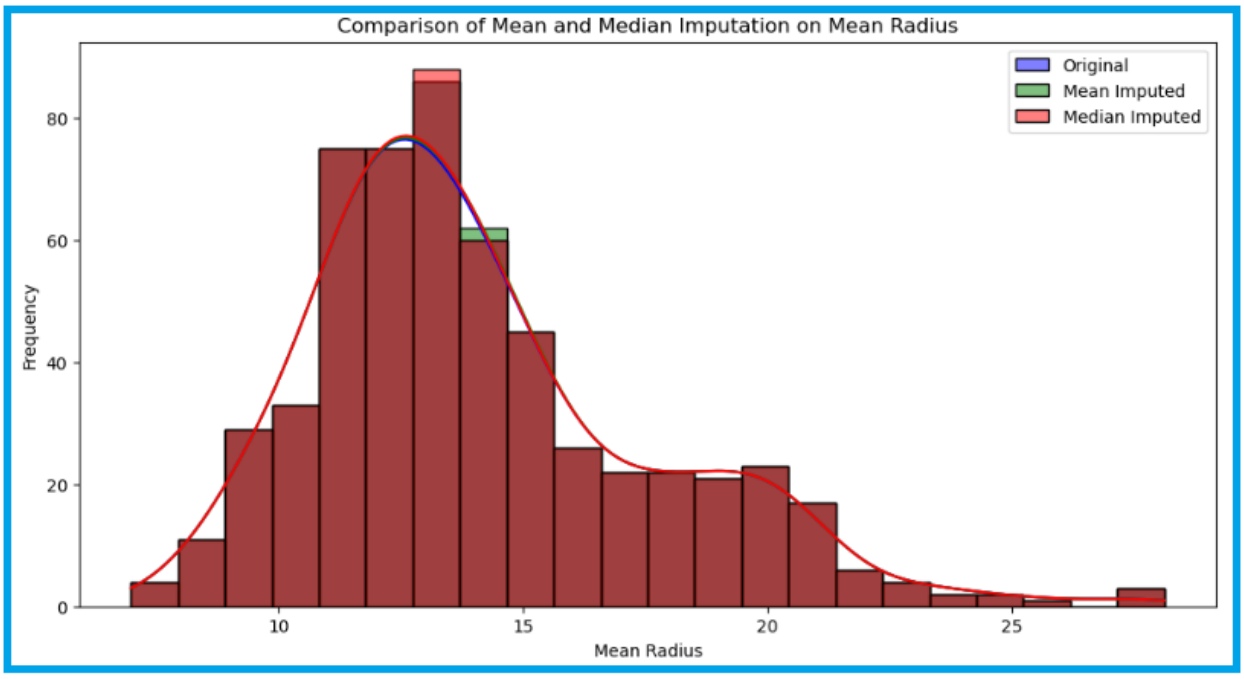
Based on the updated values of the mean radius column for the row index 70, comparing all the 4 imputation strategies, there is only minute change in the median radius value.

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean radius simpleImputer median	mean radius simpleImputer mean	mean radius KNN Imputer	mean radius Iterative Imputer	mean radius Linear Reg
70	NaN	21.31	123.6	1130.0	0.09009	18.82	18.60	18.60	18.60	18.949927
42	19.07	24.81	128.3	1104.0	0.09081	19.07	19.07	19.07	19.07	19.070000
400	17.91	21.02	124.4	994.0	0.12300	17.91	17.91	17.91	17.91	17.910000
433	18.82	21.97	123.7	1110.0	0.10180	18.82	18.82	18.82	18.82	18.820000

Scenario 3:

Comparison of calculated mean radius using Simple Imputer strategy mean/median using visualization (Histogram).

Comparison of Mean Median Imputation Plot



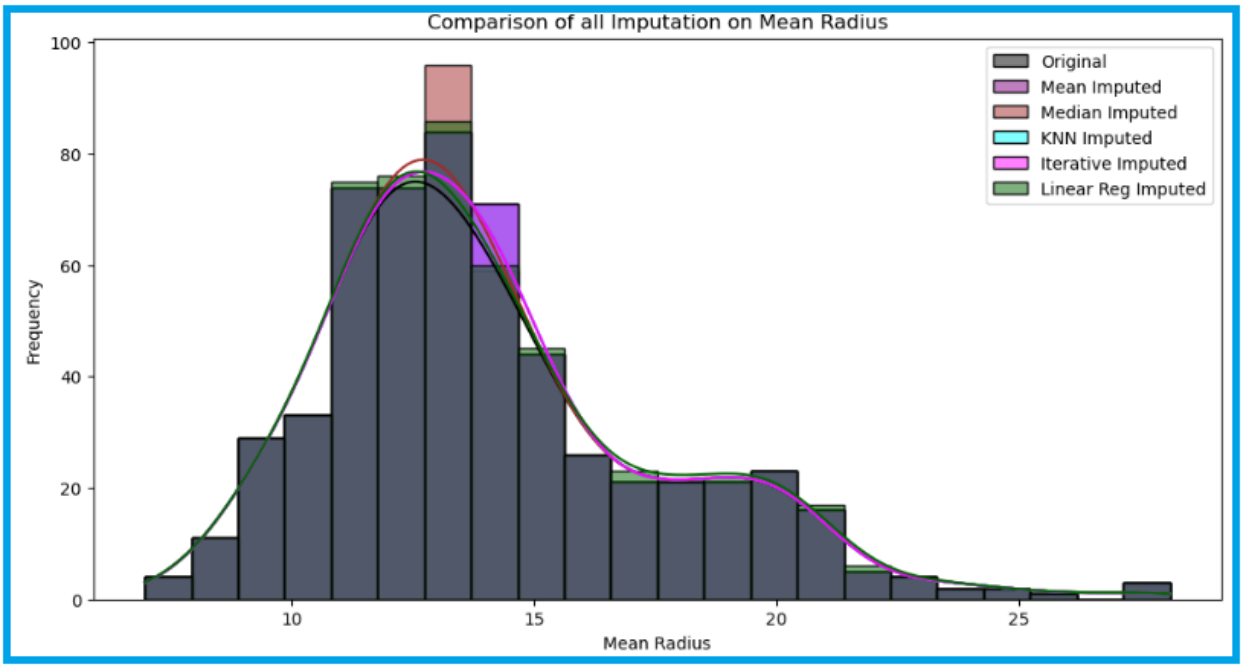
Based on the plot Mean Imputation - green bars - are concentrated near the peak of the original distribution, aligning closely with the mean value. This results in a slight overrepresentation at the center of the distribution, little affecting the natural spread and reducing variability. The shape of the distribution is more preserved than median imputation, but it may introduce bias if the data is skewed.

Median Imputation - red bars - are also concentrated near the peak but appear more localized compared to mean imputation. It retains the distribution's central tendency but leads to a slight distortion of the original shape.

Scenario 4:

Comparison of calculated mean radius using different imputation methods includes Simple Imputer (mean and median imputation) strategy, KNN Imputer, Iterative Imputer/MICE, and Linear regression and perform visualization (Histogram).

Comparison of all imputations



Based on the plot, I think Iterative Imputation is preserving the original distribution's shape and variability, making it suitable for predictive modeling. KNN Imputation is also maintaining local patterns and variability. I feel linear Regression Imputation performs well. Mean and Median Imputation are simpler but distort the distribution, especially in skewed data.

Implementation

I used python, Visual Studio Code for development and testing. The git hub location of the source code is https://github.com/rajasangeetha/MachineLearning/tree/main.

Scenario 1:

null value impact on accuracy with simulated data - brst_cancer_accuracy_change.ipynb

Scenario 2:

Comparison of calculated mean radius using different imputation techniques - imputation_value_comparison.ipynb

Scenario 3:

Comparison of calculated mean radius using Simple Imputer strategy mean/median using

Visualization -

imputation_mean_median_plot.ipynb

Scenario 4:

Comparison of calculated mean radius using different imputation strategies includes Simple Imputer (mean and median imputation) strategy, KNN Imputer, Iterative Imputer/MICE, and Linear regression and perform visualization -

imputation_comparison_plot.ipynb

References:

https://scikit-learn.org/stable/

https://python.pages.doc.ic.ac.uk/2021/

https://giphy.com/gifs/

Rithu Rocks - BM e-Solutions