Filling Missing Values

In this we are using interpolate to fill the missing values. We have observed that filling missing data with interpolate work well.

In this we are taking 2 dataset of similar values…

1. Prediction_data.xlsx (with missing values)
   Link: https://shorturl.at/zOqmm (download the dataset from the provided link)
2. Original_data.xlsx (No-missing values)
   Link: https://shorturl.at/na48n (download the dataset from the provided link)

Then we are using the below code to fill the missing data.

```
# importing the necessary libraries

import pandas as pd

import numpy as np

from sklearn.metrics import r2_score

# Loading the dataset in dataframes

Predicted_data = pd.read_excel("/content/prediction_data.xlsx")

Original_data = pd.read_excel("/content/prediction_data.xlsx")

# Now lets fill the missing values in predicted_data using Interpolation method = "linear"

Predicted_data_filled = predicted_data.interpolate(method="linear",limit_direction="both")

# Now lets ensure the index match between two index

Predicted_data_filled = predicted_data_filled.reindex_like(original_data)

# select the desired columns

Columns_to_compare =
["temperture_S1","temperature_S2","humidity_S1","humidity_S2","co2","VOC","airQual"]

Pred_cols = predicted_data_filled[columns_to_compare]

Orig_cols =  original_data[columns_to_compare]

# Now lets check the r2_score how much well it has filled the values

For col in columns_to_compare:

   R2 = r2_score(orig_cols[col],pred_cols[col])

    Print(f"R2_score for {col} : {r2.round(2)} ")
```

Filling Missing Values

# The below Is the result 1 indicate excellent filled 0 indicate not filled accurately…

Predicted score for temperature_S1 is: 92%

Predicted score for temperature_S2 is: 97%

Predicted score for humidity_S1 is: 89%

Predicted for humidity_S2 is: 94%

Predicted for co2 is: 77%

Predicted for VOC is: 92%

Predicted for airQual is: 94%

From the above we can see that your interpolate method is working well in filling missing values

Conclusion:

- The document describes a method for filling missing values in a dataset using the interpolate method. Two datasets are used: "Prediction_data.xlsx" which contains missing values, and "Original_data.xlsx" which does not have any missing values.

- The interpolate method is applied to the "Predicted_data" dataset, using the "linear" interpolation method. The missing values are filled based on the values present in the "Original_data" dataset, which serves as a reference.

- To evaluate the effectiveness of the interpolation, the R-squared (R2) score is calculated for several columns: "temperture_S1", "temperature_S2", "humidity_S1", "humidity_S2", "co2", "VOC", and "airQual". The R2 scores range from 0.77 (for the "co2" column) to 0.97 (for the "temperature_S2" column), indicating a good fit between the interpolated values and the original values.

- Link to Colab: https://shorturl.at/8MIzA

- Overall, the document concludes that the interpolate method with the "linear" option works well for filling missing values in the given dataset, as evidenced by the high R2 scores obtained for most of the columns.