Predicting missing values in an Environmental Dataset

Outline

The overall aim of the research is the development of a technique to predict missing environmental data for climate prediction. This can be used to improve the understanding of past, present, and future climate conditions and their potential impacts on the environment and society. By developing a more accurate and reliable technique for predicting missing environmental data, researchers can improve the accuracy and reliability of climate prediction models, which are critical tools for informing environmental management, policy, and decision-making.

Goal

The ultimate goal of this research is to support efforts to mitigate and adapt to the impacts of climate change, such as climate change, sea level rise, and ecosystem disruption. By improving our ability to predict and understand these impacts, we can better prepare for them and develop effective strategies for managing and reducing their effects on the environment and society.

Process Documentation

1. **Import and read data file:**  The data file was imported from the computer and read using pandas and then we found the first five rows of the DataFrame and the last five rows was also checked.
2. **Features of the Dataset:** This include the variables in the dataset that is necessary for the data exploration and analysis, and for the prediction of the missing values. They include: Year, Month, Day, Mean Surface Time (MST)- Hour, Atmospheric Pressure (Hpa), Temperature (° C ), Dew point ( ° C ), Relative Humidity ( % ), Wind Direction ( ° ), Wind speed ( m/s ), Rainfall Duration( min ) and Rainfall Amount ( mm ).
3. **Data Inspection:** This is the process of understanding the dataset, we use functions like shape() to check the number of rows and columns, df.columns to see the column labels of the dataset, info() to check the metadata of the dataset which includes; data types, number of null values, columns, memory used and index, we checked for missing values using isnull() which will return the Boolean form dataframe while isnull().sum will count the total missing values in a column, we checked the percentage of the missing values for each column, we found the descriptive statistical summary of each of the numerical columns in the dataset.
4. **Visualizing pattern of missing values:** we visualize missing values in the data using missingno package of the python libraries. It gives a very clear picture. Visualize the presence of missing data. The matrix method will make a black bar for each column with the missing data points marked at various locations. We check whether the missingness of values in one column affect the missing data behaviour or pattern in another.
5. **Filling the missing values:** it is understand that there are negative values in the dataset and which shouldn’t be for the parameters, so we replace the negative values and 0 as NaN(Not a Number) and called them missing values, then visualize the missingness of the dataset using the heatmap chart. The missing values in the dataset was imputed using the K-Nearest Neighbor Imputer from Scikit learn.
6. **Patterns of missigness of parameters:** The pattern of missingness data was visualize for a month for the year 2012, 2013 and 2014 while we also visualize for a particular Day in a month and year using the heatmap, bar chart, dendrogram and missingno matrix.
7. **Trends of the parameters:** The trend of parameters like Hpa, Temp, Relative humidity and Wind speed with respect to a particular day, month and year.
8. **Exploratory data exploration:** Exploratory data analysis is divided into Univariate Analysis, Bivariate Analysis and Multivariate Analysis. I performed the univariate, bivariate and multivariate analysis on Hpa, Temp, Relative humidity and Wind speed. Also, documenting observation after each analysis performed.
9. **Data Preprocessing:** Data preprocessing is a critical step in building machine learning models. It involves transforming raw data into a format that can be easily understood and processed by machine learning algorithms. The goal of data preprocessing is to improve the quality and relevance of the data, reduce noise, and improve the accuracy of the model. These are the steps carried out: data cleaning and manipulation, feature selection and data splitting into training set and testing set for machine learning model constructions and predictions.
10. **Machine Learning Models Construction:** The machine learning models is a necessary step to predicting the missing values of the selected features. The following algorithms were implored to build the machine learning models for the predictions. They are: Linear Regression, Decision Tree Regressor, Random Forest Regressor and K-Nearest Neighbor Regressor.
11. **Evaluation and Performance Metrics:** The machine learning models constructed are needed to be evaluated to know the best performing models. The following performance metrics were used to evaluate the model performance, they are: mean squared error, mean absolute error, root mean squared error and R2 score.
12. **Visualization of the residuals and predicted values:** The visual for the residuals and predicted values was done to show the distribution and correlation between them.