data_utils.py
Functions Overview

Contents

1 Dependencies

The following libraries are required for this module, providing essential functionalities such as data processing, statistical analysis, and visualization.

- pandas
- numpy
- matplotlib.pyplot
- mpl_toolkits.axes_grid1.inset_locator
- sklearn
- scipy.stats

2 Data Cleaning and Encoding

The Data Cleaning and Encoding functions ensure that the dataset is properly formatted, free of inconsistencies, and transformed into a suitable structure for analysis and modeling.

2.1 one_hot_encoding

one_hot_encoding(dataset, one_hot_col)

Performs one-hot encoding on a specified column in the dataset.

Parameters:

- dataset: A Pandas DataFrame containing the data.
- one_hot_col: The name of the column to be one-hot encoded.

Returns:

- A new Pandas DataFrame with the one-hot-encoded columns added and the original column dropped.

2.2 fill_nan_one_hot

fill_nan_one_hot(dataset, feature_cols, one_hot_cols, method='mean')

Fills NaN values in the specified feature columns with the mean (default) or median of the corresponding one-hot-encoded category.

Parameters:

- dataset: A Pandas DataFrame containing the data.
- feature_cols: A list of column names (features) to fill NaN values for.
- one_hot_cols: A list of one-hot-encoded column names defining the categories.
- method: A string specifying the method to use for filling NaN values. Options are "mean" or "median". Default is "mean".

Returns:

- A Pandas DataFrame with NaN values in 'feature_cols' filled based on the one-hot-encoded means.

3 Scaling and Outlier Handling

The Scaling and Outlier Handling functions are designed to enhance the quality of the data by normalizing values, addressing outliers, and ensuring that the dataset is optimized for more accurate and reliable analysis and model performance.

3.1 log_scaling

log_scaling(dataset)

Applies logarithmic scaling to numerical columns in the dataset where the range (max - min) exceeds 5000.

Parameters:

- dataset (pd.DataFrame): The input dataset containing numerical columns.

Returns:

- pd.DataFrame: A new DataFrame with logarithmically scaled values for selected columns.

3.2 remove_outliers

remove_outliers(dataset, columns, threshold=3.0)

Removes outliers from selected columns in a Pandas DataFrame using the Z-score method.

Parameters:

- dataset (pd.DataFrame): The input DataFrame containing numerical values.
- columns (list): List of column names to apply outlier removal.
- threshold (float): The Z-score threshold for detecting outliers (default is 3.0).

Returns:

- A DataFrame with outliers removed in the specified columns.

3.3 scale_cols

scale_cols(dataset, cols_to_scale)

Scales specified columns in the dataset using MinMaxScaler and updates the dataset in place.

Parameters:

- dataset: A Pandas DataFrame containing the data to scale.
- cols_to_scale: A list of column names to scale using MinMaxScaler.

Returns:

- A Pandas DataFrame with the specified columns scaled to a [0, 1] range.

4 Automated Statistics

The Automated Statistics functions provide streamlined methods for quickly calculating key statistical metrics, helping to summarize and analyze data efficiently without manual intervention.

4.1 calculate_correlation_and_pvalues

calculate_correlation_and_pvalues(dataset)

Calculates the correlation matrix and the corresponding p-value matrix for a given DataFrame.

Parameters:

- dataset: A Pandas DataFrame containing numerical data.

Returns:

- corr_matrix: A Pandas DataFrame containing the Pearson correlation coefficients.
- pval_matrix: A Pandas DataFrame containing the corresponding p-values.

4.2 filter_non_significant_vals

filter_non_significant_vals(corr_matrix, pval_matrix, athreshold=0.05)

Filters out non-significant correlation values based on a p-value threshold.

Parameters:

- corr_matrix: A Pandas DataFrame containing Pearson correlation coefficients.
- pval_matrix: A Pandas DataFrame containing the corresponding p-values.
- athreshold: A float representing the p-value threshold for significance.

Returns:

- filtered_corr_matrix: A Pandas DataFrame containing only significant correlations; non-significant values are replaced with NaN.

5 Visualization

The Visualization functions offer powerful tools for creating clear, informative graphs, enabling the effective presentation and interpretation of data insights.

5.1 plot_geospatial_data

plot_geospatial_data(dataset, geo_cols, feature_cols, one_hot_cols, one_hot_title, tot_rows, tot_cols, figsize, output_file="geospacial_data.png")

Plots geospatial data distributions and category-specific scatter plots.

Parameters:

- encoded_data: Pandas DataFrame containing the data to be plotted.
- feature_cols: List of feature column names to be visualized in individual scatter plots.
- one_hot_cols: List of one-hot-encoded column names for the combined scatter plot.
- one_hot_title: String that describes all the one_hot_cols.
- tot_rows: Number of rows in the subplot grid (must be a positive integer).
- tot_cols: Number of columns in the subplot grid (must be a positive integer).
- figsize: Tuple specifying the size of the figure (width, height), both must be positive numbers.
- output_file: Name of the file to save the generated plot. Defaults to 'geospacial_data.png'.

Returns:

- None. Displays the plots and saves the figure as a file.

5.2 plot_data_distribution

plot_data_distribution(dataset, feature_cols, one_hot_cols, one_hot_title, bins, tot_rows, tot_cols, figsize, output_file="data_distribution.png")

Plots histograms for selected features and a bar chart for one-hot-encoded column frequencies.

Parameters:

- dataset: Pandas DataFrame containing the data to be plotted.
- feature_cols: List of numerical feature column names to plot histograms for.
- one_hot_cols: List of one-hot-encoded column names to plot frequency distributions for.
- one_hot_title: String that describes the one_hot_cols.
- bins: Number of bins for the histograms.
- tot_rows: Number of rows in the subplot grid.
- tot_cols: Number of columns in the subplot grid.
- figsize: Tuple specifying the figure size (width, height).
- output_file: Name of the file to save the generated plot. Default is 'data_distribution.png'.

Returns:

- None. Displays the plots and saves the figure as a file.

5.3 plot_one_hot_boxplot

plot_one_hot_boxplot(dataset, feature_cols, one_hot_cols, one_hot_title, tot_rows, tot_cols, figsize, output_file="box_plot.png")

Plots histograms for selected features and a bar chart for one-hot-encoded column frequencies.

Parameters:

- dataset: Pandas DataFrame containing the data to be plotted. For better visualization the values should be normalized.
- feature_cols: List of numerical feature column names to plot histograms for.
- one_hot_cols: List of one-hot-encoded column names to plot frequency distributions for.
- one_hot_title: String that describes the one_hot_cols.
- tot_rows: Number of rows in the subplot grid.
- tot_cols: Number of columns in the subplot grid.
- figsize: Tuple specifying the figure size (width, height).
- output_file: Name of the file to save the generated plot. Default is "box_plot.png".

Returns:

- None. Displays the plots and saves the figure as a file.

5.4 plot_one_hot_corr_matrix

plot_one_hot_corr_matrix(dataset, feature_cols, one_hot_cols, one_hot_title, tot_rows, tot_cols, figsize, output_file="correlation_matrix.png")

Plots histograms for selected features and a bar chart for one-hot-encoded column frequencies.

Parameters:

- dataset: Pandas DataFrame containing the data to be plotted. For better visualization the values should be normalized.
- feature_cols: List of numerical feature column names to plot histograms for.
- one_hot_cols: List of one-hot-encoded column names to plot frequency distributions for.
- one_hot_title: String that describes the one_hot_cols.
- tot_rows: Number of rows in the subplot grid.
- tot_cols: Number of columns in the subplot grid.
- figsize: Tuple specifying the figure size (width, height).
- output_file: Name of the file to save the generated plot. Default is "correlation_matrix.png".

Returns:

- None. Displays the plots and saves the figure as a file.

5.5 plot_one_hot_significance

plot_one_hot_significance(dataset, feature_cols, one_hot_cols, one_hot_title, athreshold, tot_rows, tot_cols, figsize, output_file="sig_corr_matrix.png")

Plots histograms for selected features and a bar chart for one-hot-encoded column frequencies.

Parameters:

- dataset: Pandas DataFrame containing the data to be plotted. For better visualization the values should be normalized.
- feature_cols: List of numerical feature column names to plot histograms for.
- one_hot_cols: List of one-hot-encoded column names to plot frequency distributions for.
- one_hot_title: String that describes the one_hot_cols.
- athreshold: A float representing the p-value threshold for significance.
- tot_rows: Number of rows in the subplot grid.
- tot_cols: Number of columns in the subplot grid.
- figsize: Tuple specifying the figure size (width, height).
- output_file: Name of the file to save the generated plot. Default is "sig_corr_matrix.png".

Returns:

- None. Displays the plots and saves the figure as a file.