MAna

**Authors:**

**Github:**

**Summary:**

MAna is a Python package that provides a collection of functions, classes, and modules for data analysis, manipulation, and modeling. It aims to simplify and streamline the data analysis and modeling process by providing commonly used tools and techniques. The package is organized into sub-packages, including data for loading and cleaning data, analysis for data visualization, and modeling for data preprocessing and model evaluation and visualization. Users can import the desired functions or classes from the relevant sub-packages to use the package's functionality. MAna provides a comprehensive set of tools for data analysis and modeling.

**Design:**

**data sub-package**

data\_io: This module provides functions for reading and writing data from/to various file formats, including CSV, Excel, and SQL databases.

data\_cleaning: This module provides functions for cleaning and preprocessing data, including dropping missing values, encoding categorical variables, and scaling numeric features.

**analysis sub-package**

visualizations: This module provides functions for visualizing data, including scatter plots, histograms, and heatmaps.

**modeling sub-package**

model\_evaluation: This module provides functions for evaluating the performance of machine learning models, including accuracy, precision, recall, and F1 score.

data\_preprocessing: This module provides functions for preprocessing data for machine learning models, including scaling numeric features, encoding categorical variables, and splitting data into training and test sets.

models: This module provides classes for building and training machine learning models, including linear regression, logistic regression, decision trees, and random forests.

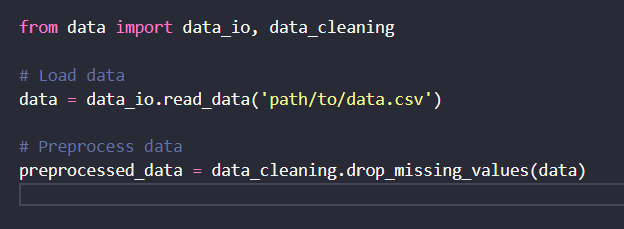
The package is organized into sub-packages, each of which contains modules that group related functions and classes. The data sub-package contains functions for loading and preprocessing data, the analysis sub-package contains functions for visualizing data, and the modeling sub-package contains functions and classes for building, training, and evaluating machine learning models. Overall, the package aims to provide a comprehensive set of tools for data manipulation, analysis, and modeling in a user-friendly and easy-to-use interface.

**Usage:**

The MAna package can be used to simplify and streamline the process of data analysis and modeling. It provides a set of commonly used tools and techniques for loading, cleaning, visualizing, and modeling data. Here are some examples of how the package can be used:

**Loading and cleaning data:**

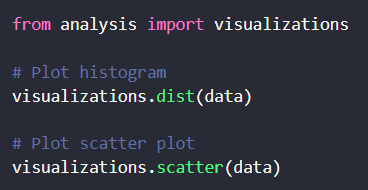
The data sub-package provides functions for loading and cleaning data. Here is an example of how to use the package to load and preprocess data:



In this example, we import the data\_io and data\_cleaning modules from the data sub-package. We use the read\_data function to load the data from a CSV file and then use the drop\_missing\_values function to preprocess the data by removing any rows with missing values.

**Data visualization:**

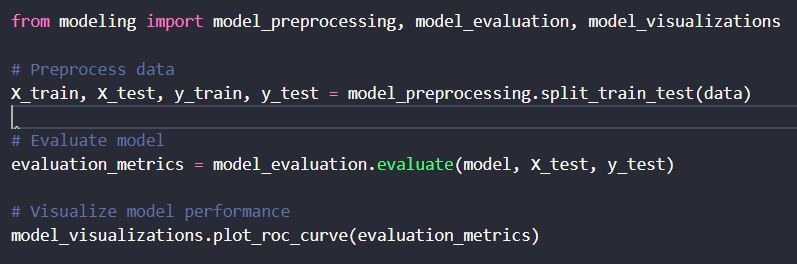
The analysis sub-package provides functions for visualizing data. Here is an example of how to use the package to plot a histogram and a scatter plot:



In this example, we import the visualizations module from the analysis sub-package. We use the dist function to plot a histogram of the data and the scatter function to plot a scatter plot of the data.

**Modeling:**

The modeling sub-package provides functions for data preprocessing and model evaluation and visualization. Here is an example of how to use the package to train and evaluate a machine learning model:



In this example, we import the model\_preprocessing, model\_evaluation, and model\_visualizations modules from the modeling sub-package. We use the split\_train\_test function to split the data into training and testing sets, the evaluate function to evaluate the performance of the model on the testing data, and the plot\_roc\_curve function to visualize the performance of the model.

Overall, the MAna package can be used to simplify and streamline the data analysis and modeling process by providing a set of commonly used tools and techniques. The package can be integrated into a data science workflow to help users load, clean, visualize, and model data more efficiently.

**Discussion:**

There are several related libraries in the Python ecosystem that provide similar functionality to the MAna package, such as Pandas, NumPy, Matplotlib, Scikit-learn, and Statsmodels. However, MAna aims to provide a more streamlined and simplified approach to data manipulation, analysis, and modeling by focusing on a set of commonly used tools and techniques. It also provides a clear and organized package structure that makes it easier for users to find and use the functions and classes they need.

One advantage of using MAna over other libraries is its ease of use and readability. The code is written in a way that is easy to understand and follow, even for users with limited experience in data analysis and modeling. The package also provides clear and concise documentation, which makes it easier for users to understand how to use the functions and classes.

Another advantage of MAna is its focus on streamlining the data analysis and modeling workflow. By providing a set of commonly used tools and techniques in a single package, users can save time and avoid the need to search for and learn new libraries for each step of the workflow.

However, MAna may not be suitable for all use cases or data analysis tasks, as it focuses on a limited set of tools and techniques. Users who require more advanced or specialized tools may need to use additional libraries.

To improve the package, future updates could include the addition of more advanced and specialized tools, as well as improvements to performance and scalability for larger datasets. Additionally, incorporating more machine learning algorithms and techniques could further expand the package's capabilities.

**Statement of contributions:**

**References:**

**External libraries used by this package:**

pandas

numpy

matplotlib

seaborn

scikit-learn

**Sources used as a reference:**

McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.", 2012.

VanderPlas, Jake. Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.", 2016.

Scikit-learn: Machine Learning in Python. Pedregosa, F. et al., JMLR 12, pp. 2825-2830, 2011.