Download and Process the Thyroid Dataset

Important! Be sure to read About this liveProject before beginning. It contains crucial information for your work.

Objective

- The objective of this project is to build an automated diagnostic system for thyroid disorders, while also learning anomaly detection. Anomaly-detection techniques are very useful when we are dealing with class-imbalance problems. These are problems in which one class is much more frequent than the other. Many problems in medicine fit this description. It is easy to find many instances of healthy patients, but only a few of them will develop a disorder.
- In this milestone, you will download the thyroid disease dataset and apply some of the novelty- and outlier-detection techniques available in scikit-learn. The dataset contains a decent number of instances (close to 4000), has a small dimensionality, and a high percentage of anomalies, which makes it ideal as a starting point for anomaly algorithm comparison. These techniques don't need any special `pip` installs, unlike more advanced techniques using the imbalanced-learn and PyOD libraries, which we will encounter in later projects in this series.

Importance to project

- This dataset will be used for further analysis in the subsequent milestones.

Workflow

1. Use the starter template provided in the GitHub repository.
2. Download the data from Thyroid Disease dataset.

3. Store it in a directory two levels above where the notebook is.
4. Import the relevant packages, including SciPy, one of the most popular packages in Python for scientific computing.
5. Edit the metadata section in the notebook to define the relevant parameters.
6. Use the function `scipy.io.loadmat()` to load the .mat file into a Python dictionary.
7. Extract X (the input data) and y (the target data) from the dictionary and concatenate to form a pandas DataFrame.
8. Inspect the DataFrame and apply the `.info()` and `.describe()` methods to see summary stats.
9. Check to see if there are any columns that need to be cleaned up.
10. Export the DataFrame as a csv file.

Deliverable

The deliverable for this milestone is a Jupyter Notebook that performs the operations in the Workflow on the loaded thyroid disease dataset.

Upload a link to your deliverable in the Submit Your Work section and click submit. After submitting, the author's solution and peer solutions will appear on the page for you to examine.

Help

Feeling stuck? Use as little or as much help as you need to reach the solution!

resources

Pandas in Action by Boris Paskhaver

Chapter 1, "Introducing Pandas," provides a good refresher overview of pandas.

Pandas in Action by Boris Paskhaver

Chapter 4, " The DataFrame object," explains pandas and working with a dataframe.

Pandas in Action by Boris Paskhaver

Chapter 5, “Filtering a DataFrame,” elaborates on reducing, extracting, and filtering DataFrames.

Pandas in Action by Boris Paskhaver

Chapter 10, “Merging, joining, and concatenating,” further discusses these ideas applied to DataFrames.