Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

DashiBoard

Alberto Vezzoso, Beta 80 spA

Spoke 3 Progetti Bandi a Cascata, 24/09, 2024

Project Overview

Problem
Data science and research must be integrated consistently with production-grade pipelines. Often, deploying data-driven solutions is time-consuming and even requires the intervention of a dedicated development team.

Aim
Orchestrate the work of researchers, data scientists, and developers through unified software. This allows research teams to conduct swift data exploration, create robust and reproducible analysis pipelines, and generate deployable, reusable solutions.

Output
Providing a minimalist—yet fully general—interface for data analysis and visualization able to automatically building, filtering and preprocessing functions for a given dataset;

Project Overview

Graph-theoretical Approach

The declarative style characterizing the entire Dashiboard approach allows us to represent filtering, preprocessing, and processing (cards) steps as a Directed Acyclic Graph (DAG).

On the one hand, analyzing the DAG structure enables efficient data and computational management.

On the other hand, we can cluster pipelines based on graph-theoretical invariants.

Reference

Vertechi, P., and Bergomi. M. G. (2023). "Machines of finite depth: Towards a formalization of neural networks." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 8.

Bergomi, M. G., Ferri, M., Mella, A., & **Vertechi, P**. (2023). Generalized Persistence for Equivariant Operators in Machine Learning. *Machine Learning and Knowledge Extraction*, *5*(2), 346-358

Vertechi, P. (2022). Dependent optics. *Applied Category Theory.*

Bergomi, M. G., Ferri, M., **Vertechi, P.**, & Zuffi, L. (2021). Beyond topological persistence: Starting from networks. *Mathematics*, *9*(23), 3079.

Bergomi, M. G., & **Vertechi, P.** (2020). Rank-based persistence. *Theory and Applications of Categories,* Vol. 35, No. 9, pp. 228–260.

Technical Objectives, Methodologies and Solutions

Data Loading | Filtering & Preprocessing | Processing | Visualization

The data loader module is designed to facilitate the ingestion of tabular data in widely utilized formats such as CSV, Parquet, Arrow, and JSON.

Additionally, this module will feature an interactive visualization tool for the loaded table.

This visualization will allow for real-time enrichment of the data with analysis results, ensuring dynamic updates based on filtering, preprocessing, and other configurable parameters.

Technical Objectives, Methodologies and Solutions

Data Loading

The data loader module is designed to facilitate the ingestion of tabular data in widely utilized formats such as CSV, Parquet, Arrow, and JSON.

Additionally, this module will feature an interactive visualization tool for the loaded table.

This visualization will allow for real-time enrichment of the data with analysis results, ensuring dynamic updates based on filtering, preprocessing, and other configurable parameters.

Filtering & Preprocessing

Filtering widgets are automatically created according to the detected column type: continuous data are filtered through a double-range selection tool, checkboxes allow the filtering of categorical columns, and radio buttons are used for boolean values.

The table visualization is updated dynamically.

Continuous columns (possibly grouped by categorical ones) can be preprocessed according to the most popular normalisation and standardization techniques.

Processing

Visualization

Technical Objectives, Methodologies and Solutions

Data Loading

The data loader module is designed to facilitate the ingestion of tabular data in widely utilized formats such as CSV, Parquet, Arrow, and JSON.

Additionally, this module will feature an interactive visualization tool for the loaded table.

This visualization will allow for real-time enrichment of the data with analysis results, ensuring dynamic updates based on filtering, preprocessing, and other configurable parameters.

Filtering & Preprocessing

Filtering widgets are automatically created according to the detected column type: continuous data are filtered through a double-range selection tool, checkboxes allow the filtering of categorical columns, and radio buttons are used for boolean values.

The table visualization is updated dynamically.

Continuous columns (possibly grouped by categorical ones) can be preprocessed according to the most popular normalisation and standardization techniques.

Processing

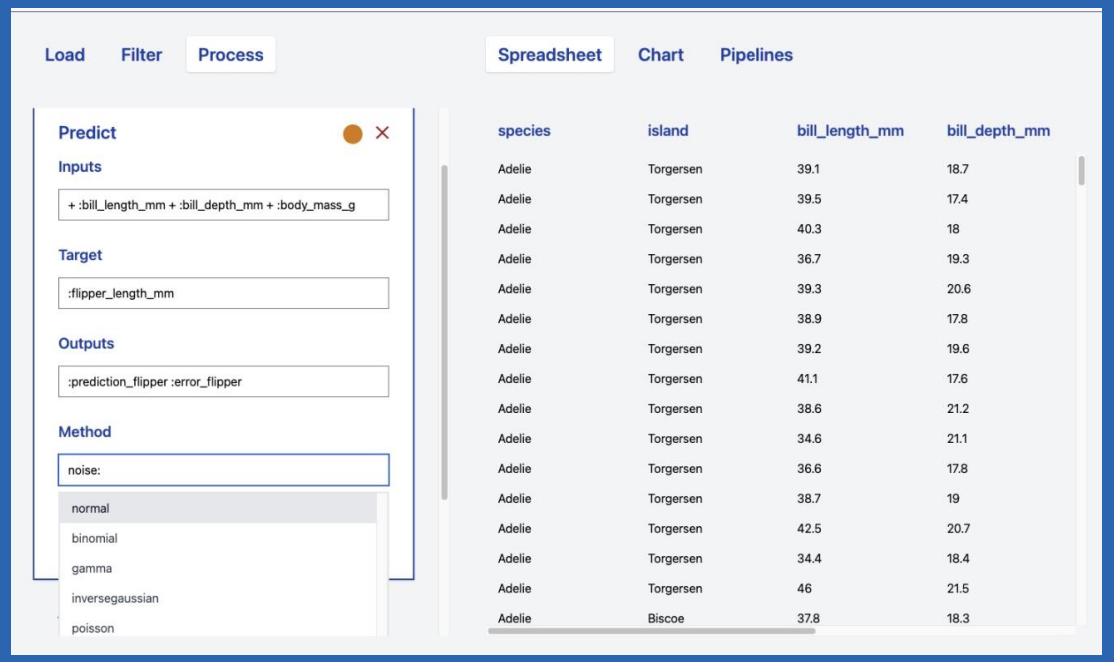
Processing and analysis are managed through cards, a simple yet parameterizable and modular interface.

We foresee the following cards:

1. Clustering
2. Interpolation & Prediction (GLM, Kernel Methods, MLP, CNN, RNN, Parametric Machines)
3. Classification (GLM, Kernel Methods, MLP, CNN, Parametric Machines)
4. Projection (Dimensionality Reduction)
5. Wild Card

Visualization

Technical Objectives, Methodologies and Solutions



Cards help users select the right parameters and algorithms for their tasks. They can be automatically composed as cells in a notebook, and an intelligent autocomplete feature allows for quick parameterization of complex algorithms.

popular normalisation and standardization techniques.

Processing

Processing and analysis are managed through cards, a simple yet parameterizable and modular interface.

We foresee the following cards:

1. Clustering
2. Interpolation & Prediction (GLM, Kernel Methods, MLP, CNN, RNN, Parametric Machines)
3. Classification (GLM, Kernel Methods, MLP, CNN, Parametric Machines)
4. Projection (Dimensionality Reduction)
5. Wild Card

Visualization

Technical Objectives, Methodologies and Solutions

Data Loading

The data loader module is designed to facilitate the ingestion of tabular data in widely utilized formats such as CSV, Parquet, Arrow, and JSON.

Additionally, this module will feature an interactive visualization tool for the loaded table.

This visualization will allow for real-time enrichment of the data with analysis results, ensuring dynamic updates based on filtering, preprocessing, and other configurable parameters.

Filtering & Preprocessing

Filtering widgets are automatically created according to the detected column type: continuous data are filtered through a double-range selection tool, checkboxes allow the filtering of categorical columns, and radio buttons are used for boolean values.

The table visualization is updated dynamically.

Continuous columns (possibly grouped by categorical ones) can be preprocessed according to the most popular normalisation and standardization techniques.

Processing

Processing and analysis are managed through cards, a simple yet parameterizable and modular interface.

We foresee the following cards:

1. Clustering
2. Interpolation & Prediction (GLM, Kernel Methods, MLP, CNN, RNN, Parametric Machines)
3. Classification (GLM, Kernel Methods, MLP, CNN, Parametric Machines)
4. Projection (Dimensionality Reduction)
5. Wild Card

Visualization

The visualization module allows for

1. Translating questions about data into relevant visualizations declaratively.

2. Removing cognitive overhead via opinionated graphical defaults and broad data format support.

Technical Objectives, Methodologies and Solutions

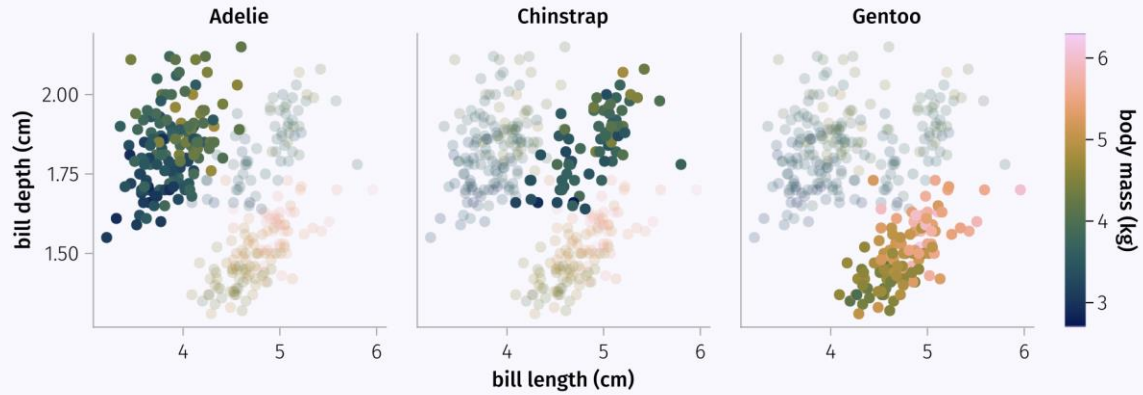
Data Loadi

The data loader module is d
facilitate the ingestion of tab
widely utilized formats such
Parquet, Arrow, and JSON.

Additionally, this module will
interactive visualization tool
loaded table.

This visualization will allow f
enrichment of the data with
results, ensuring dynamic up
on filtering, preprocessing, a
configurable parameters.

```
layers = visual(alpha = 0.2) #=semitransparent layer=# + species_as_column #=faceted layer=#
plt = dataset * bill * body_mass_as_color * layers
```



We employ a simplified version of AlgebraOfGraphics (Pietro Vertechi) to create high-quality figures using simple instructions and opinionated, high-quality defaults. The AoG approach is particularly well-suited to our card formalism.
.

g

managed

r interface.

ds:

(GLM,
NN, RNN,

nel Methods,
lachines)
y Reduction)

preprocessed according to the most popular normalisation and standardization techniques.

5. Wild Card

Visualization

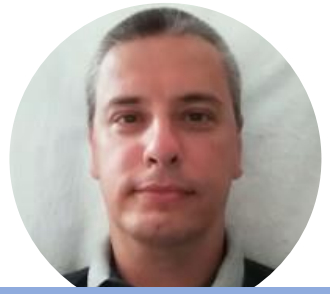
The visualization module allows for

1. Translating questions about data into relevant visualizations declaratively.
2. Removing cognitive overhead via opinionated graphical defaults and broad data format support.

Involved Staff

Francesco Silanos, Eng.

Scientific Coordinator

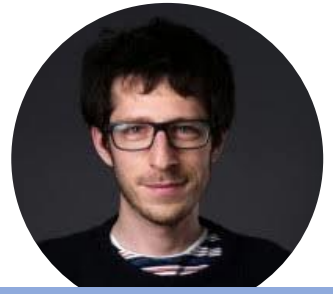


Cristiano Notargiacomo, Eng.

Senior Architect

Gabriele Aprile, Eng.

Software Engineer



Pietro Vertechi, PhD.

ML and Applied Mathematics. Domain Expert.

Author of AoG

Timescale, Milestones, SAL

1st Quarter

Blueprint

Data Loader (DuckDB)

Filtering & Preprocessing

2nd Quarter

Prediction & Interpolation Cards

Visualization (e.g., gt vs preds, loss function, metrics)

Preliminary tests on N-body problem

3rd Quarter

Classification Card

Clustering and Projection

Visualization

Automated reporting

4th Quarter

Wild Card

Efficient data orchestration

DAG formalism

Pipeline Export