Streamline Your Scientific Work with AutoML

Aleksandra Płońska, Piotr Płoński

https://mljar.com   piotr@mljar.com   aleksandra@mljar.com

How Automated Machine Learning helps scientists?

Building a machine learning pipeline generates a lot of questions and tasks to handle. The whole analysis process can be automated with the **mljar-supervised** Python package with just a few lines of code. The machine learning task detection is automatically based on target values.

The MLJAR AutoML can work with:

- binary classification
- multi-class classification
- regression

01 Prepare data for training

AutoML allows you automatically to make:

- data preprocessing
- feature engineering
- feature selection

02 Choose algorithm

Let the AutoML choose the best one for you. It tries a variety of algorithms and creates leaderboard with the scores.

03 What are the best hyperparameters?

AutoML will answer your questions:

- How many training iterations?
- Which learning rate?
- What max depth trees should have?
- and many others ...

04 Document and explain models

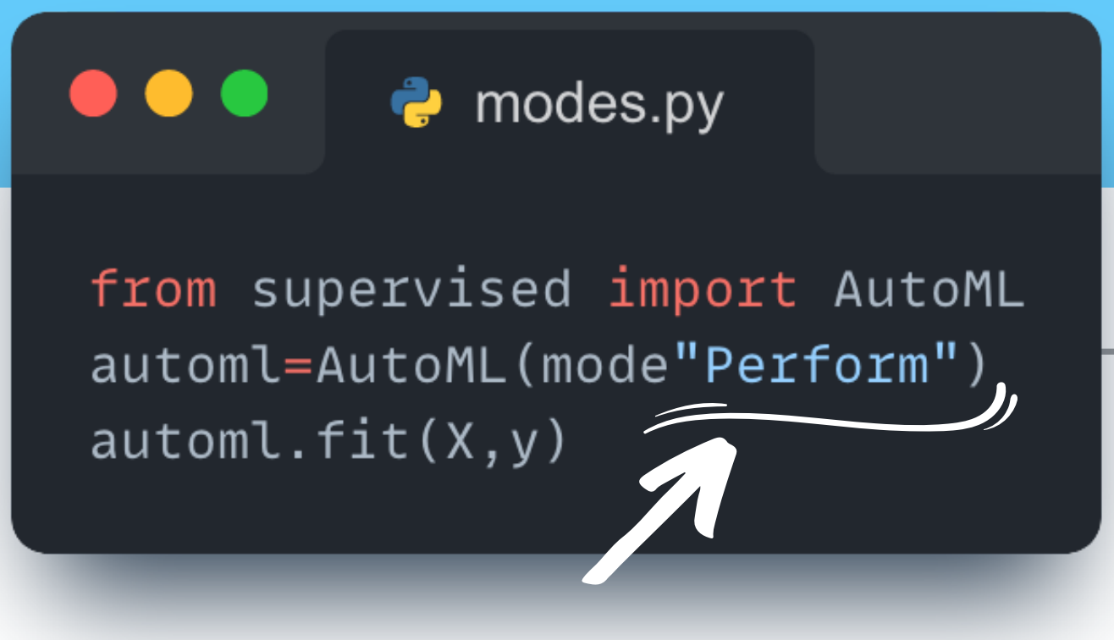
Get the full report of trained models with SHAP plots, learning curves, importance plot...

education-num
capital-gain
relationship
age

Features

- complete pipeline
- golden feature
- model leaderboard
- feature selection
- auto-saving models
- hyper-parameters tuning
- automatic documentation
- variety of algorithms
- automated reports

Modes



modes.py

```python
from supervised import AutoML
automl=AutoML(mode"Perform")
automl.fit(X,y)
```

Perform

- Production-ready ML pipeline
- 5 fold cross-validation
- Feature engineering
- Search for a model under constraint for prediction time on a single sample

Compete

- ML competitions under time budget
- Adjusted validation
- Train/test on 5 or 10 fold cross-validation
- Feature engineering
- Try many models

Explain

- Ideal for initial data analysis
- 75% 25% train/test split
- explanations

Optuna

- 10 fold cross-validation
- Tune algorithm with Optuna framework

MACHINE LEARNING PIPELINE - STAGES AUTOMATED WITH AUTOML

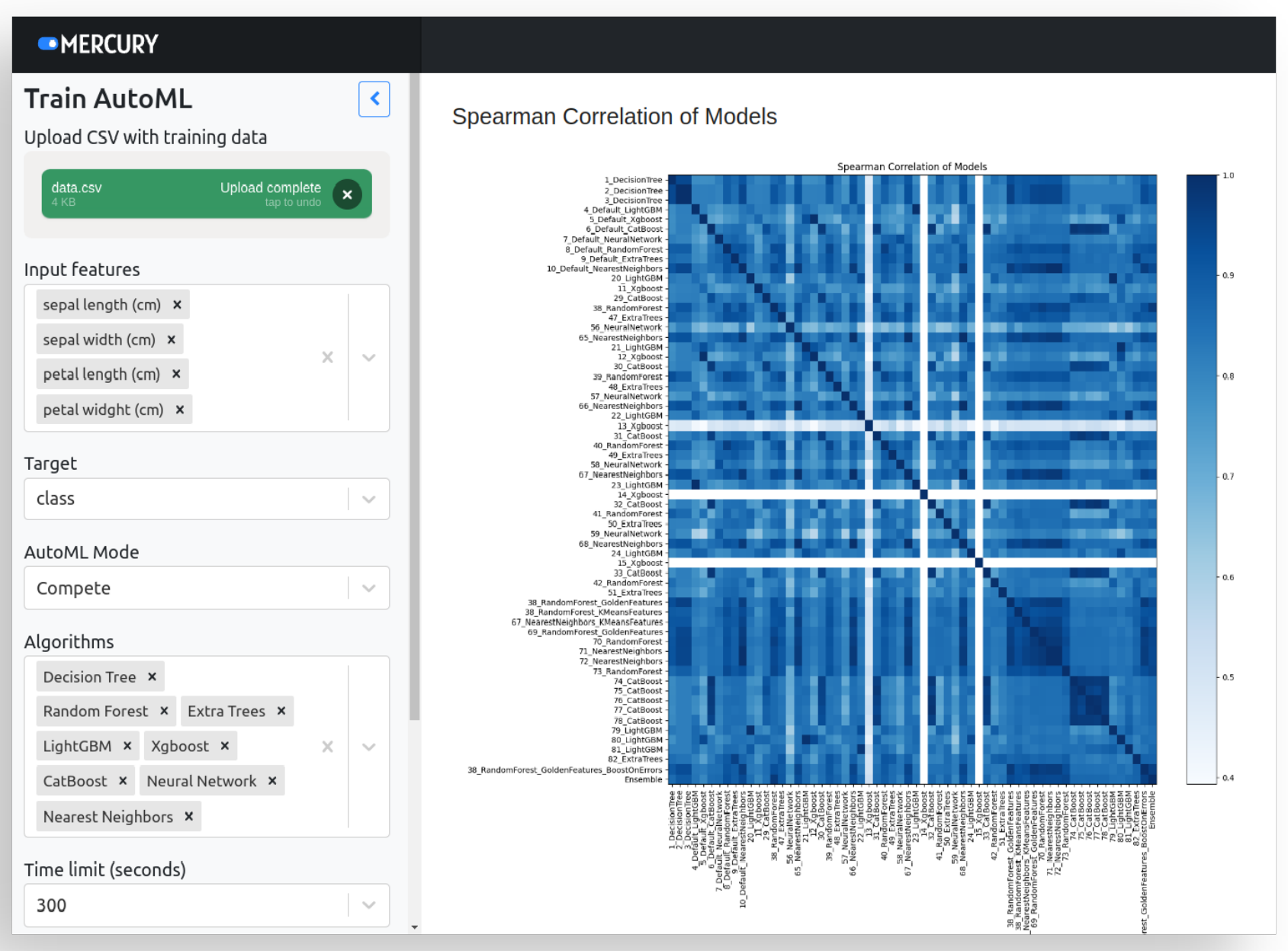
PRE-PROCESSING → FEATURES ENGINEERING → FEATURES CONSTRUCTION → TRAINING → ALGORITHM SELECTION → TUNING → DEPLOYING → EXPLAINING → DOCUMENTING

Try AutoML

UI WEB APP

https://github.com/mljar/automl-app

AutoML web app with a graphical user interface. Click and run your first project!



MERCURY

Train AutoML

Upload CSV with training data

data.csv 4 KB — Upload complete (tap to undo)

Input features: sepal length (cm), sepal width (cm), petal length (cm), petal widght (cm)

Target: class

AutoML Mode: Compete

Algorithms: Decision Tree, Random Forest, Extra Trees, LightGBM, Xgboost, CatBoost, Neural Network, Nearest Neighbors

Time limit (seconds): 300

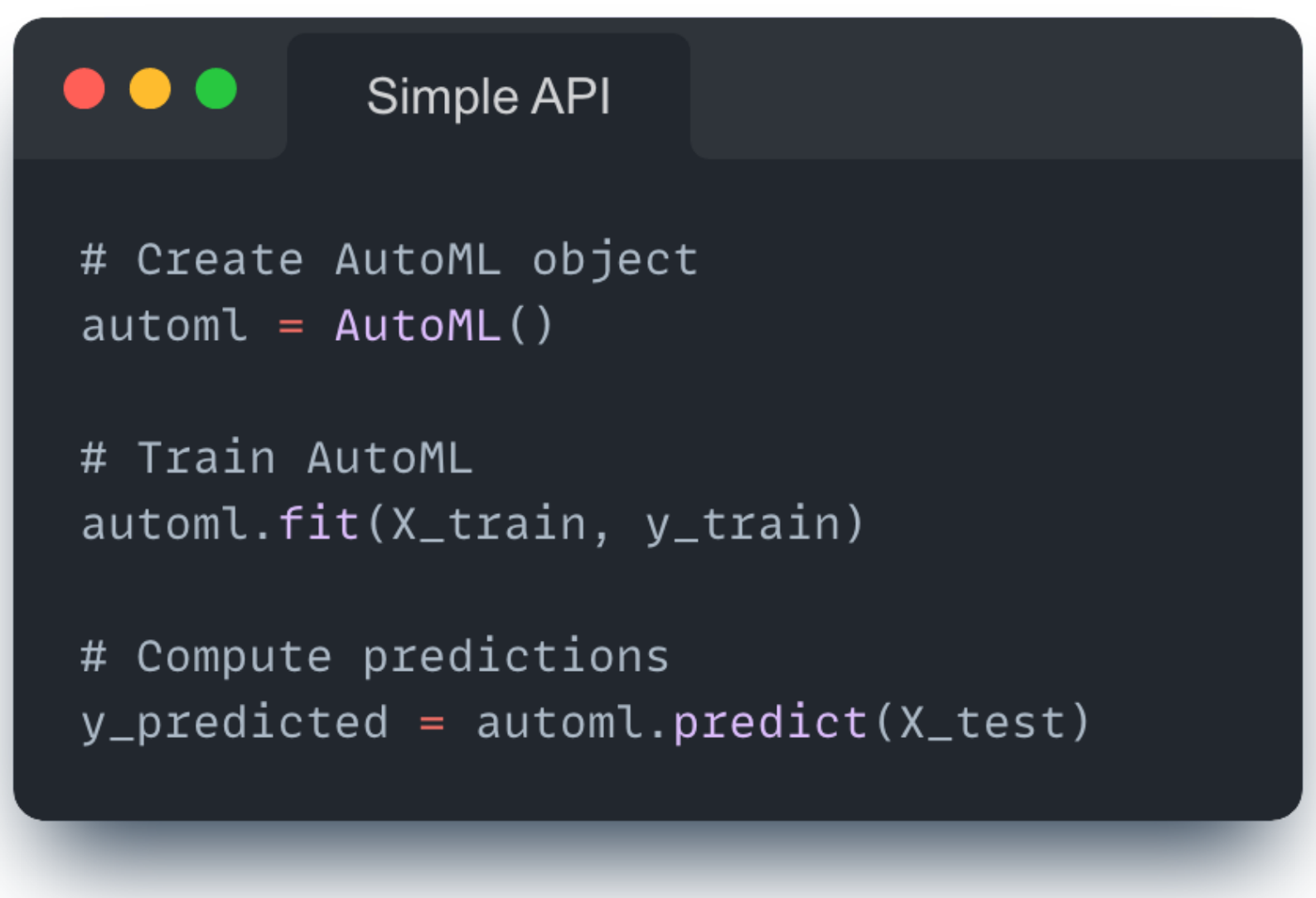
Spearman Correlation of Models

PYTHON PACKAGE

Install mljar-supervised in 2 ways:

1. PyPi repository
2. From Conda

https://github.com/mljar/mljar-supervised



Simple API

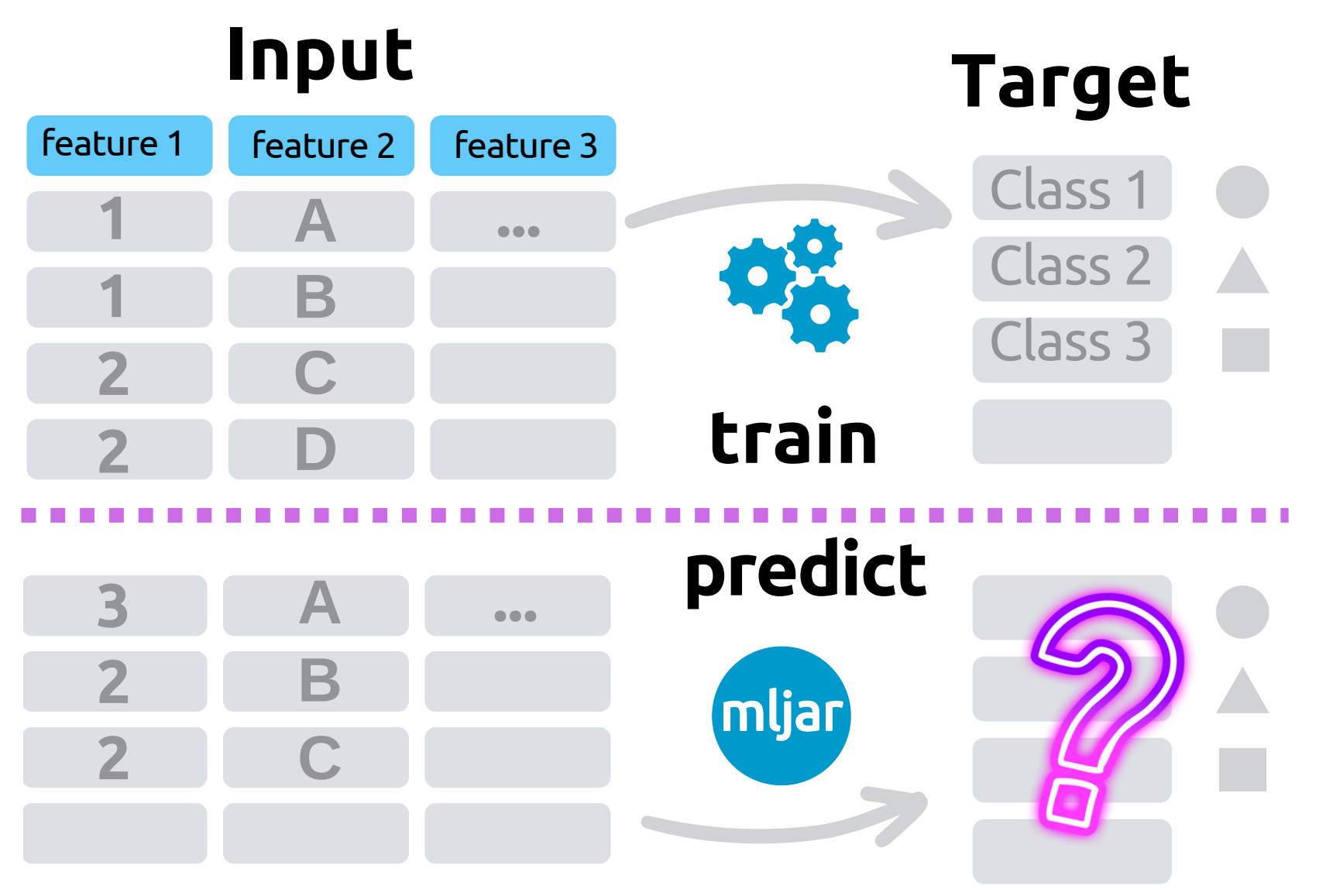
```python
# Create AutoML object
automl = AutoML()

# Train AutoML
automl.fit(X_train, y_train)

# Compute predictions
y_predicted = automl.predict(X_test)
```

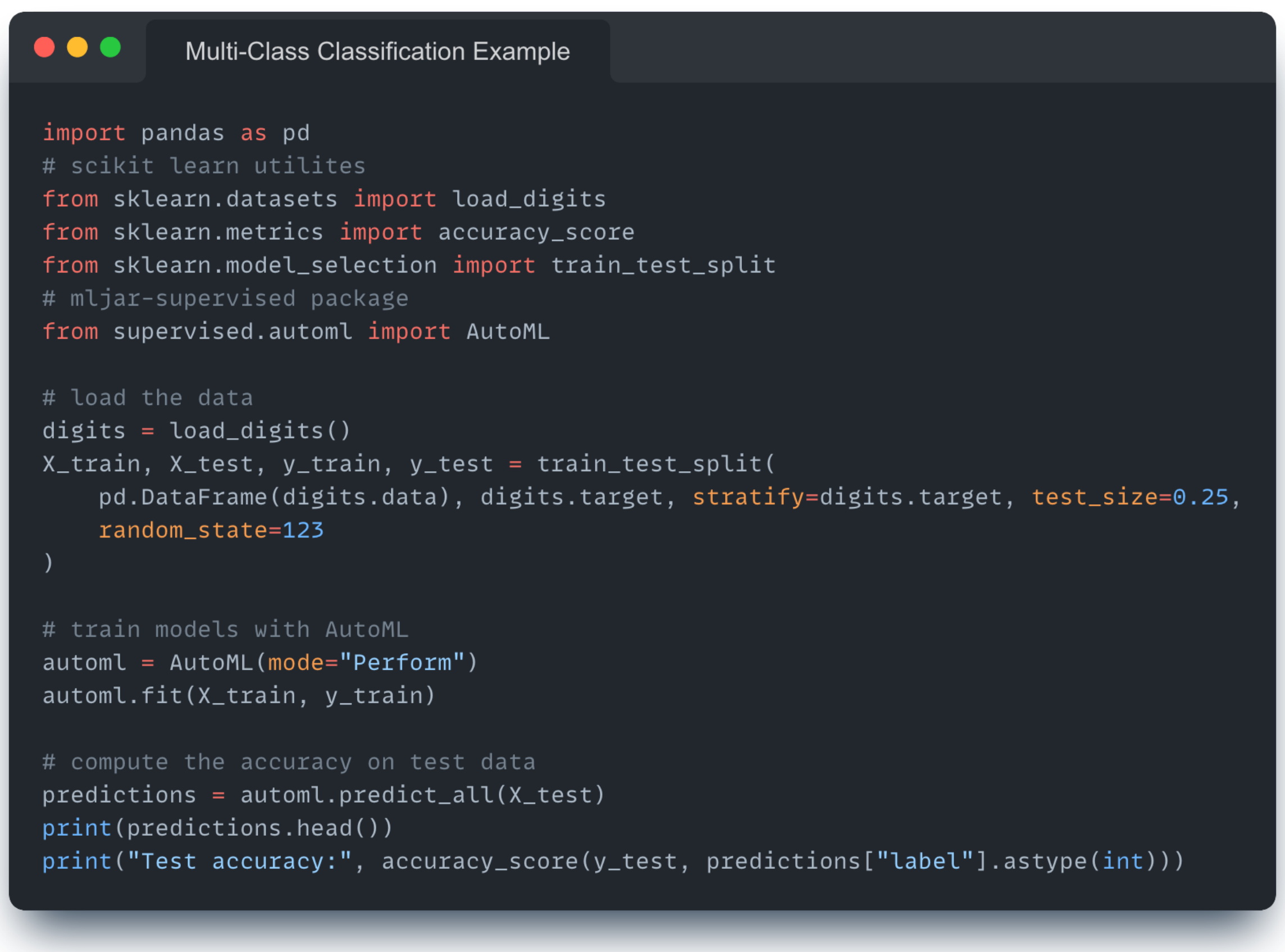
mljar-supervised

one Python package that includes

CatBoost, NumPy, learn, pandas, Python, LightGBM, M↓, OPTUNA, matplotlib, XGBoost

What is supervised learning





Multi-Class Classification Example

```python
import pandas as pd
# scikit learn utilites
from sklearn.datasets import load_digits
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
# mljar-supervised package
from supervised.automl import AutoML

# load the data
digits = load_digits()
X_train, X_test, y_train, y_test = train_test_split(
    pd.DataFrame(digits.data), digits.target, stratify=digits.target, test_size=0.25,
    random_state=123
)

# train models with AutoML
automl = AutoML(mode="Perform")
automl.fit(X_train, y_train)

# compute the accuracy on test data
predictions = automl.predict_all(X_test)
print(predictions.head())
print("Test accuracy:", accuracy_score(y_test, predictions["label"].astype(int)))
```

Microbiology

Designing and identifying β-hairpin peptide macrocycles with antibiotic potential

Justin R. Randall, Cory D. DuPai, T. Jeffrey Cole, Gillian Davidson, Kyra E. Groover, Sabrina L. Slater, Despoina A. I. Mavridou, Claus O. Wilke and Bryan W. Davies

"We are excited about the future possibilities of pairing functional cell-based peptide screening technology with machine learning strategies, especially for antibiotic discovery. We believe that as more antibacterial peptide data become available (...) machine learning may be able to predict antibacterial activity de novo, bypassing the need for human design and functional screening entirely."

Pharmacy

Artificial Intelligence-Based Quantitative Structure–Property Relationship Model for Predicting Human Intestinal Absorption of Compounds with Serotonergic Activity

Natalia Czub, Jakub Szlęk, Adam Pacławski, Klaudia Klimończyk, Matteo Puccetti, and Aleksander Mendyk

"In this work, we focused on drug permeability looking at human intestinal absorption as a marker for intestinal bioavailability. (...)The proposed system based on AI represents a promising tool useful for oral drug screening at an early stage of drug discovery and development."

Math

Machine Learning Class Numbers of Real Quadratic Fields

Malik Amir, Yang-Hui He, Kyu-Hwan Lee, Thomas Oliver, Eldar Sultanow

The article explores the application of supervised learning techniques to distinguish real quadratic fields with class numbers 1, 2, and 3. It delves into the challenges faced in separating class numbers 1 and 3 and proposes incorporating additional features inspired by the analytic class number formula to improve classification accuracy.

Medicine

Prediction of Recurrent Mutations in SARS-CoV-2 Using Artificial Neural Networks

Bryan Saldivar-Espinoza, Guillem Macip , Pol Garcia-Segura, Júlia Mestres-Truyol, Pere Puigbò,Adrià Cereto-Massagué , Gerard Pujadas and Santiago Garcia-Vallve

"Predicting SARS-CoV-2 mutations is difficult, but predicting recurrent mutations driven by the host, such as those caused by host deaminases, is feasible. We used machine learning to predict which positions from the SARS-CoV-2 genome will hold a recurrent mutation and which mutations will be the most recurrent."

Finance

Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring

Michael Bucker, Gero Szepannek, Alicja Gosiewska, Przemyslaw Biecek

A major requirement for credit scoring models is to provide a maximally accurate risk prediction. Additionally, regulators demand these models to be transparent and auditable. (...) This paper works out different dimensions that have to be considered for making credit scoring models understandable and presents a framework for making "black box" machine learning models transparent, auditable, and explainable.

Technology

Predictive Quality Modeling for Ultra-Short-Pulse Laser Structuring utilizing Machine Learning

Lars Leyendecker, Milena Zuric, Muhammad Atique Nazar, Karl Johannes, Robert H. Schmitt

"Laser structuring offers precision and versatility for material processing but holds potential for optimization due to high-energy consumption and long production-times. Based on a process parameter study, we utilize Machine Learning and multi-modal data fusion of process parameters, high-frequency monitoring data and workpiece properties."