

SEMINAR: BIG DATA AND MACHINE LEARNING APPLICATIONS FOR ENTERPRISES

Explainable AI (XAI) for Trustworthy Business Decisions

Focus: Transparent, sustainable AutoML with **SustainaML** + Oil & Gas Demo

PRESENTED BY

Mehak Mushtaq Malik
SustainaML Lead Developer

Assoc. Prof. Radwa El Shawi
Head of Data Systems Group

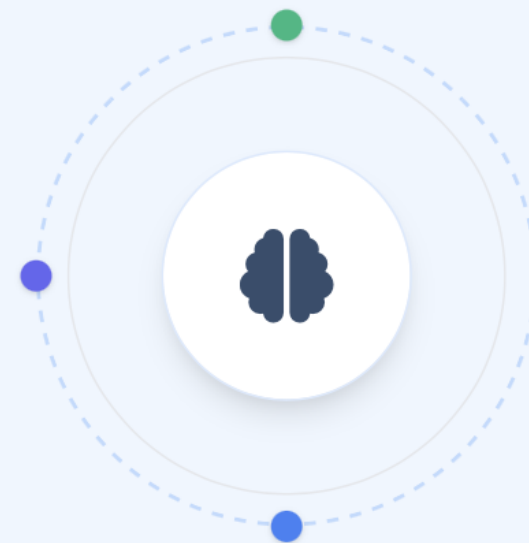
 Institute of Computer Science, University of Tartu



Co-funded by
the European Union



Investing
in your future



Resources & Code

[Github repository](#)

[ECML PKDD 2025 Demos](#)



What is AutoML?

Core Definition

AutoML (Automated Machine Learning) automates the end-to-end process of applying machine learning to real-world problems.

It handles tedious tasks like data preparation, feature engineering, model selection, and hyperparameter tuning—allowing teams to build reliable models faster with less manual effort.

Data Prep → Training → Deployment

Key Concepts



Full Automation

Covers the complete pipeline from raw data ingestion to a deployable model, not just the training phase.



Smart Optimization

Uses advanced algorithms to explore thousands of model architectures and hyperparameter combinations efficiently.



Democratization

Lowers the barrier to entry, enabling domain experts and analysts to build ML solutions without deep coding expertise.






Trust & Guardrails

Includes built-in explainability, best practices, and governance features to ensure models are transparent and reliable.




The Challenge: Opaque AutoML \neq Trust

Common Pain Points

-  **Black-box pipelines**
Unclear hyperparameter choices and model logic.
-  **Domain constraints**
Difficult to inject business rules or physics-based limits.
-  **Hidden costs**
No visibility into compute resources or energy consumption.



Business Risks

-  **Compliance & Audit**
Inability to explain decisions to regulators or auditors.
-  **Low Adoption**
Stakeholders reject models they don't understand or trust.
-  **Environmental Impact**
Unchecked emissions from inefficient model training.



The Goal



Explainable
Understand the 'why'



Steerable
Human-in-the-loop control



Sustainable
Energy-efficient AI

Explainable AI (XAI): What and How

Applicable to Oil & Gas Models

Global Explanations

Understand the model as a whole. Which features drive predictions generally?

Feature Importance Partial Dependence

Local Explanations

Explain individual predictions. Why was this specific well flagged for maintenance?

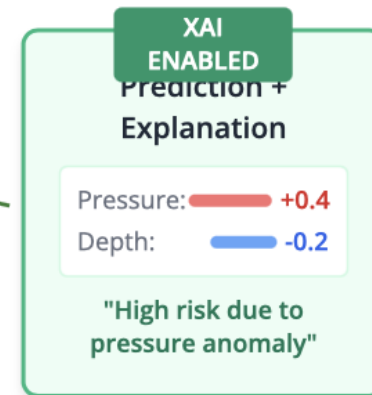
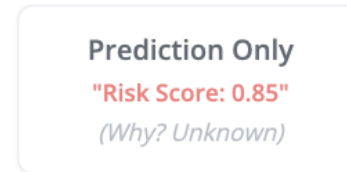
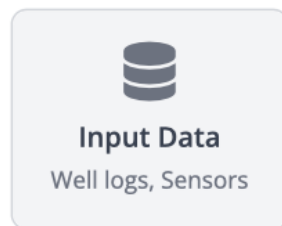
SHAP Values LIME

Counterfactuals

"What-if" analysis. How would changing pressure or temperature alter the forecast?

Actionable Insights Sensitivity Analysis

BLACK BOX VS. XAI WORKFLOW



Why this matters for Oil & Gas:

Production forecasts and equipment failure predictions (regression/classification) rely on physical parameters. XAI validates that the model respects physical laws (e.g., pressure vs. flow rate relationships), building trust with engineers.

Why XAI Matters for Business Decisions



Transparency

Demystify "black-box" models for non-technical stakeholders. Clear explanations bridge the gap between data science teams and business executives.



Faster Adoption

Build trust in automated systems. Users are more likely to adopt and rely on AI tools when they understand the rationale behind predictions.



Governance & Risk

Ensure compliance with regulations. Robust model lifecycle control enables easier auditing and better risk management.

OIL & GAS INDUSTRY APPLICATIONS



Production Planning & Allocation

Context: Explaining why specific wells are prioritized for extraction.

Benefit: Optimizes resource allocation with clear justification for field engineers.



Predictive Maintenance

Context: Clarifying why a pump or pipeline segment is flagged for repair.

Benefit: Reduces false alarms and justifies downtime costs to management.

Introducing SustainaML

 Green AI Initiative

Unified Interface

LIGHTWEIGHT VISUALIZATION

Interactive layer built atop leading AutoML frameworks (FLAML, H2O, MLJAR). Enables seamless search space refinement and result analysis in one place.



Sustainability Metrics

ENERGY & CARBON TRACKING

Go beyond accuracy. Measure **Energy Consumption** (μWh) and **CO2 Emissions** (μg) for every model training run via CodeCarbon integration.

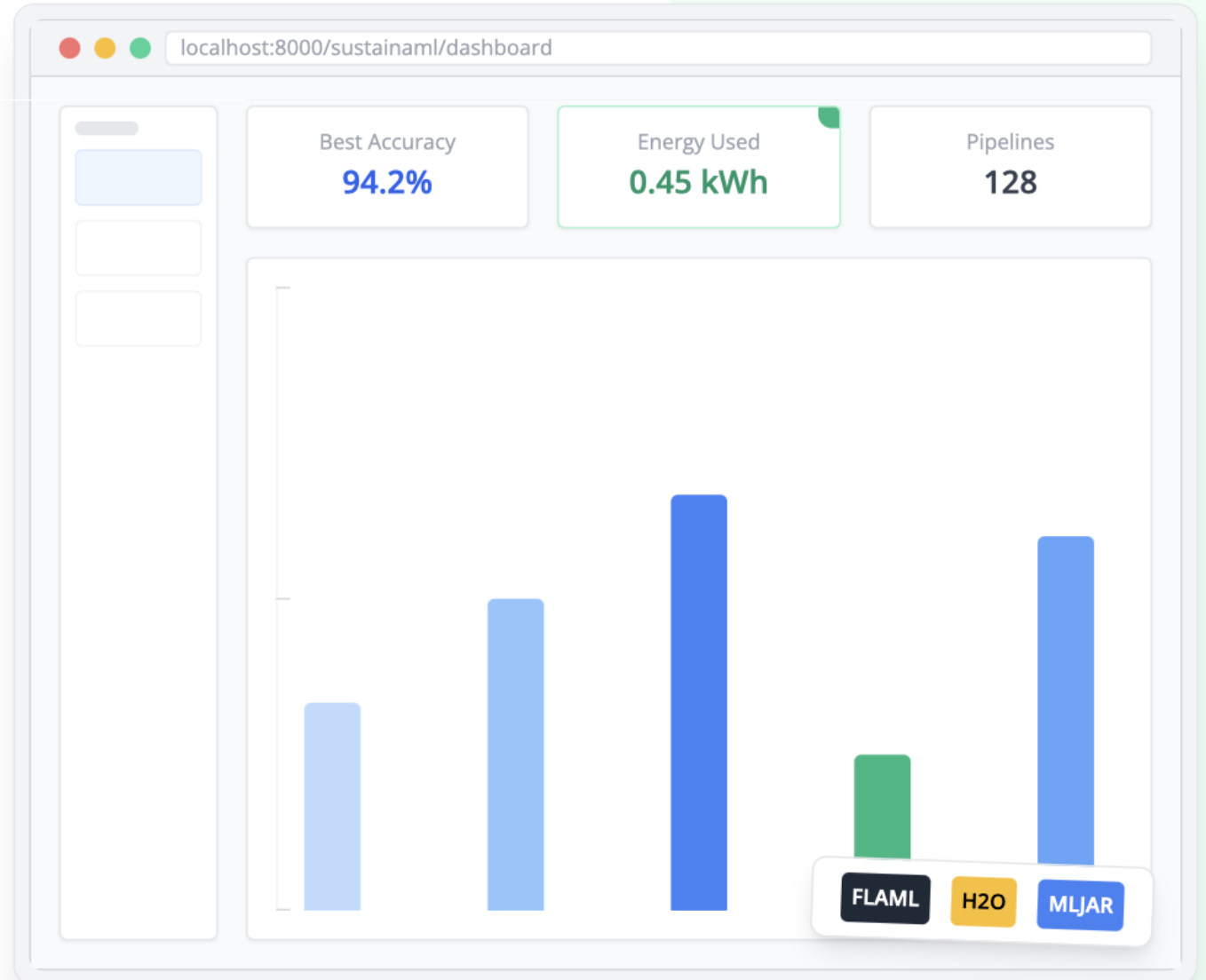
NEW INNOVATION



Actionable Feedback

TRADE-OFF DECISIONS

Visual dashboards highlighting the balance between model performance and environmental cost, empowering steerable and responsible AI.



SustainaML: Key Capabilities



Transparency

Visualize the entire AutoML pipeline process. Move beyond the "black box" by inspecting hyperparameter choices, model evolution, and performance metrics in an intuitive dashboard.



Steerability

Enable Human-in-the-Loop (HITL) control. Users can refine search spaces, select algorithms, and adjust time budgets in real-time based on intermediate results.



Sustainability

Integrate Green AI metrics directly into the evaluation. Track Energy Consumption (μWh) and CO2 Emissions (μg) alongside accuracy to make environmentally conscious decisions.

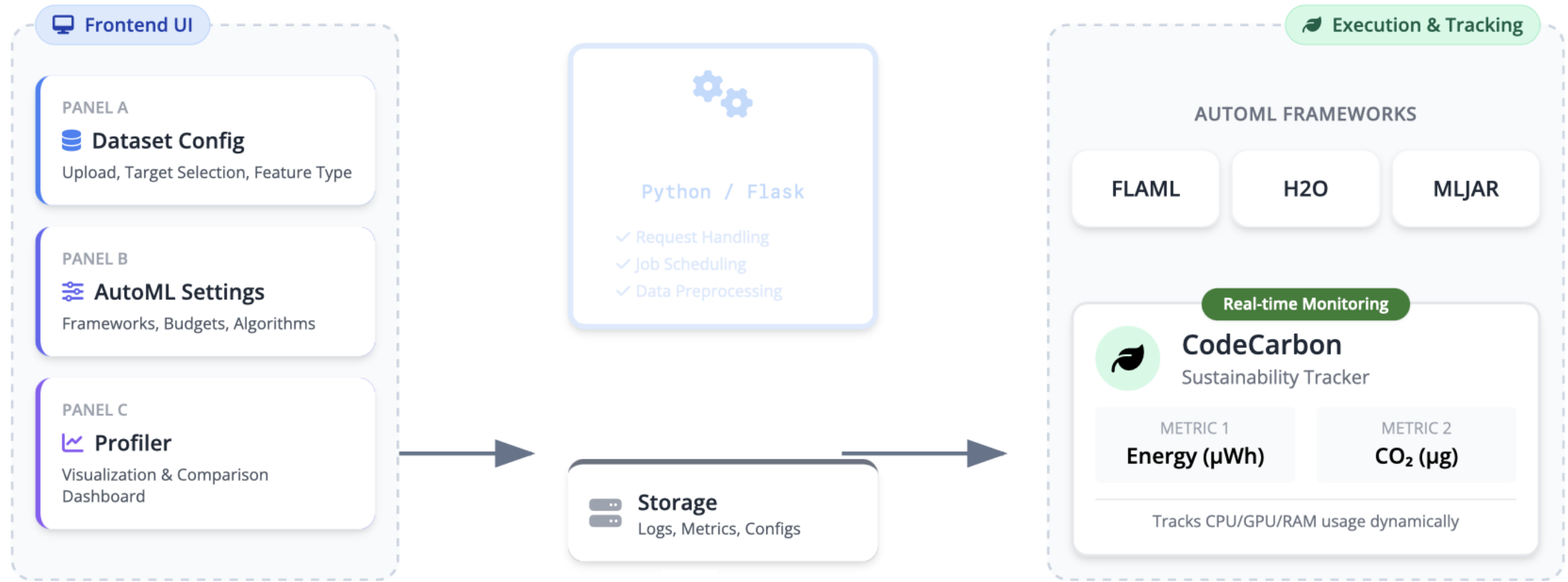


Multi-framework

Unify diverse AutoML backends under one roof. Seamlessly run, compare, and analyze results from FLAML, H2O, and MLJAR within a single consistent interface.

SustainaML Architecture

Integrating interactive transparency with environmental tracking



Measuring Sustainability in AutoML

We track real-time environmental impact using CodeCarbon to quantify the cost of model training.

ENERGY CONSUMPTION



μWh (Micro-Watt-hours)

Measures total electricity usage of CPU/GPU during AutoML search and training.

CO₂ EMISSIONS

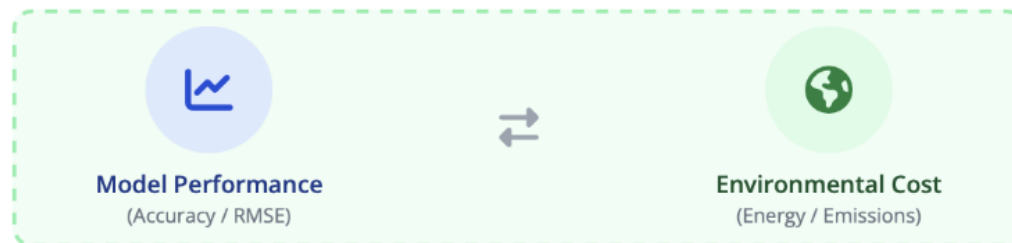
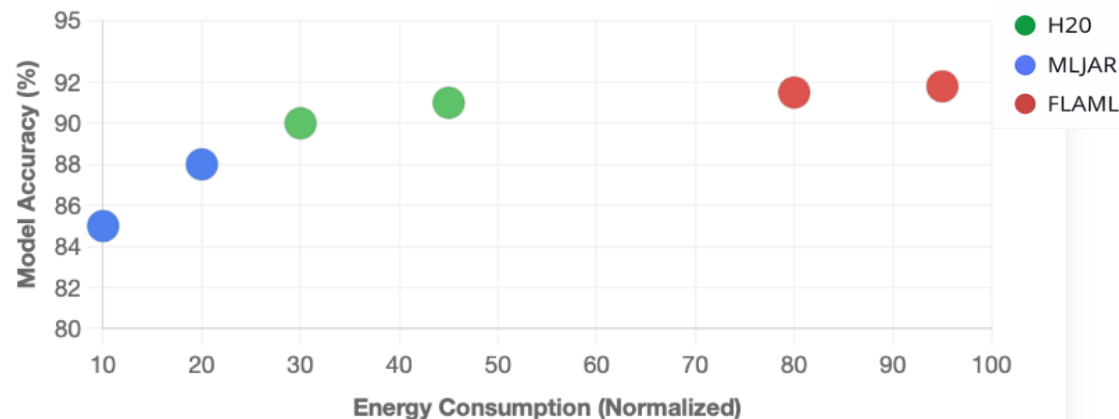


μg (Micrograms of CO₂)

Calculated based on energy usage × regional carbon intensity of the power grid.

⚖️ The Decision Trade-off

High accuracy often comes at a high environmental cost. SustainaML helps find the "Green Sweet Spot".



Demo: Oil & Gas Production Prediction



SustainaML Case Study



Business Objective

Forecast near-term production volumes to support downstream allocation, logistics planning, and revenue estimation.



ML Task

Supervised Regression: Predict next month's oil/gas volume based on historical well attributes and time-series lags.



Dataset

Public Kaggle Dataset: *US Oil & Gas Production & Disposition*. Contains region, product type, and volume history.



Production Facilities

Optimizing efficiency through transparent AI

THE CHALLENGE

Balancing Accuracy & Sustainability

High-accuracy models (like Ensembles) often require excessive compute resources. For an industry focused on efficiency, we need to ask:

⚡ High Energy Cost?

🌫️ Hidden Emissions?



Oil & Gas Production Dataset

Source: Kaggle (US Oil & Gas Production)

Raw Data Structure

~10k rows

Date	Region	Product	Volume (bbl)
2020-01-01	Permian	Crude Oil	4,520,100
2020-01-01	Bakken	Natural Gas	2,890,500
2020-02-01	Permian	Crude Oil	4,610,300
2020-02-01	Eagle Ford	Condensate	1,250,000

*Simplified view of input features

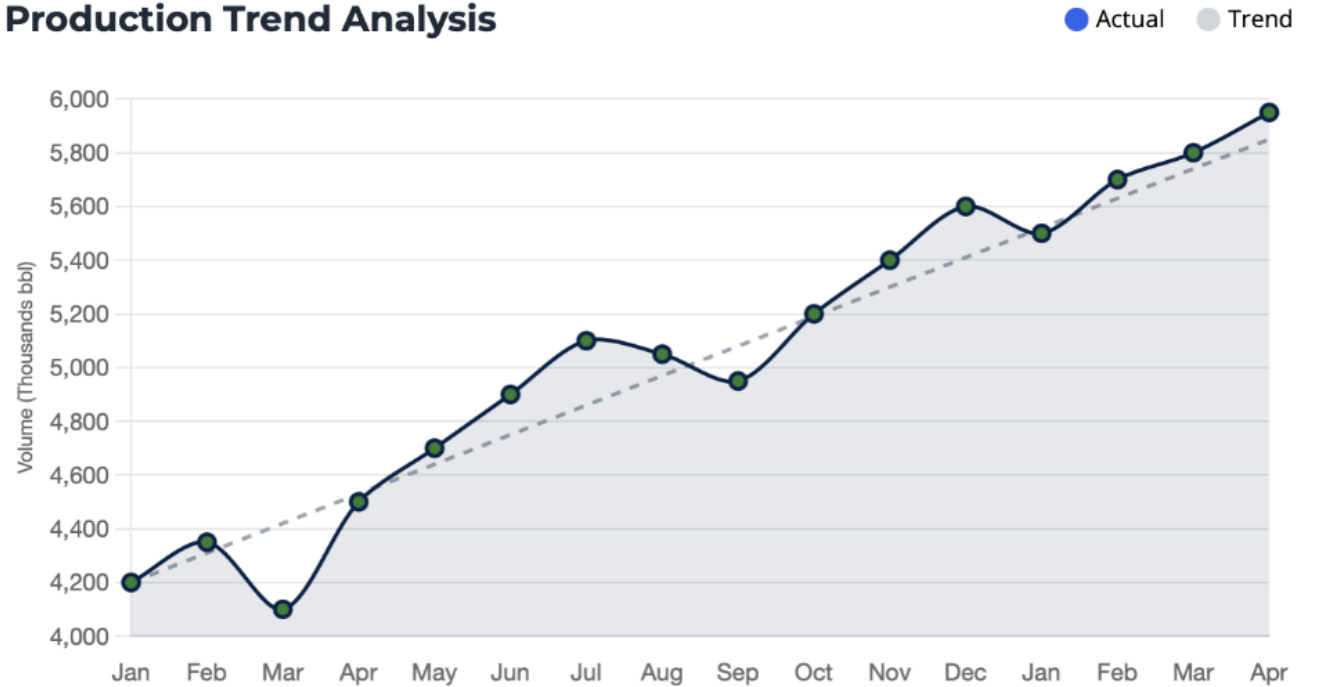
Feature Engineering

Temporal Lag_1M Lag_3M Lag_6M Rolling_Mean_3M

Contextual Region_OneHot Seasonality_Idx Price_Index

Target Next Month Production Volume (Continuous)

Production Trend Analysis



Data shows clear seasonal patterns and upward trend | Split Strategy: Time-based (Train: 2015-2023, Test: 2024)

TASK TYPE
Regression

PRIMARY METRIC
RMSE / MAE

SUSTAINAML GOALS
Min Energy & CO₂

Dataset Configuration (Panel A)

▶ Step 1: Data Setup

http://localhost:8000/sustainaml/dataset
Upload New

Dataset Configuration

CURRENT DATASET

oil_production_v2.csv

TARGET VARIABLE

production_vol_bbl
▾

TASK TYPE

Regression

Classification

TIME COLUMN (OPTIONAL)

report_date
▾

Feature Correlation Matrix i Pearson Correlation

💡 Why this step matters

Proper dataset configuration is crucial for AutoML. Incorrectly setting the target or ignoring time dependencies in oil & gas data leads to "leakage" and invalid models.

1 Data Upload & Preview

Upload CSV/Excel files directly. The system automatically detects data types (numerical, categorical, datetime).

2 Target Variable Selection

Select the column to predict (e.g., production_vol_bbl). SustainaML auto-suggests regression or classification tasks based on the target's cardinality.

3 Time Series Handling

Critical for O&G: Designate a time column to ensure train/test splits respect chronological order, preventing future data from leaking into training.

4 Correlation Heatmap

Identify highly correlated features (redundancy) or strong predictors early. Helps in deciding if feature selection is

AutoML Settings (Panel B)

http://localhost:8000/sustainaml/configure

Configure Experiments

Mode: Sustainability Focused

1. SELECT FRAMEWORKS

FLAML ✓

Fast & Low Resource

H2O ✓

Distributed ML

MLJAR ○

Auto-Exploration

2. INCLUDED ALGORITHMS

XGBoost ×
LightGBM ×
Random Forest ×
+ Add

Excluding "Neural Networks" to reduce energy footprint.

3. TIME BUDGET (PER FRAMEWORK)

Current: 300 sec

10s
1hr

4. HYPERPARAMETERS

⚙️ Edit Search Space

● Using Default Bounds

💡 **Why this step matters**

SustainaML allows you to "steer" the AutoML process. By constraining frameworks and algorithms, you can avoid wasting energy on models that don't meet business requirements (e.g., interpretability).

1 Multi-Framework Selection

Selectively enable FLAML, H2O, or MLJAR. For quick prototyping, use FLAML. For distributed power, add H2O.

2 Algorithm Constraints

Oil & Gas Context: Exclude "black-box" Deep Learning models if regulatory compliance requires explainable decision trees (e.g., XGBoost, RF).

3 Time & Energy Budget

Define how long each framework runs. Lower budgets (e.g., 60s) save energy but might miss complex patterns. Higher budgets (1hr+) maximize accuracy.

4 Hyperparameter Tuning

Customize search spaces (e.g., limit tree depth to prevent overfitting on small datasets). Or stick to defaults for ease

Profiler Dashboard (Panel C)

http://localhost:8000/sustainaml/profiler

AutoML Profiler Results

Dataset: Oil_Prod_v2 Frameworks: H2O, FLAML, MLJAR Total Time: 15m 30s

Dataset Insights (Feature Importance)

Pre-run Analysis

Top Pipelines Comparison

Sorted by CO2

Framework	Algorithm	RMSE ↓	Energy (μWh)	CO2 (μg)
H2O	GBM_Grid_1	124.5	450.2	1.2
FLAML	LGBM_3	126.8	510.5	1.4
MLJAR	Xgboost_2	122.1	890.0	2.5

Business Value

Compare not just accuracy, but the environmental cost. Identify models that are "good enough" but significantly greener.

1 Dataset Insights

Visualizes feature importance before deeper analysis. Confirm that key production drivers (e.g., pressure, hours online) are being used correctly.

2 Unified Pipeline Table

A single view comparing H2O, FLAML, and MLJAR. Key Feature: Energy and CO2 columns allow direct sustainability comparisons alongside RMSE.

3 Sustainability Trade-offs

Scatter plot reveals the "Pareto frontier." Spot models that achieve high accuracy with low emissions (bottom-left quadrant targets).

4 Multi-Budget Analysis

Analyze if longer training times yield better results. Often, a 60s run is 99% as accurate as a 300s run but uses 5x less energy.

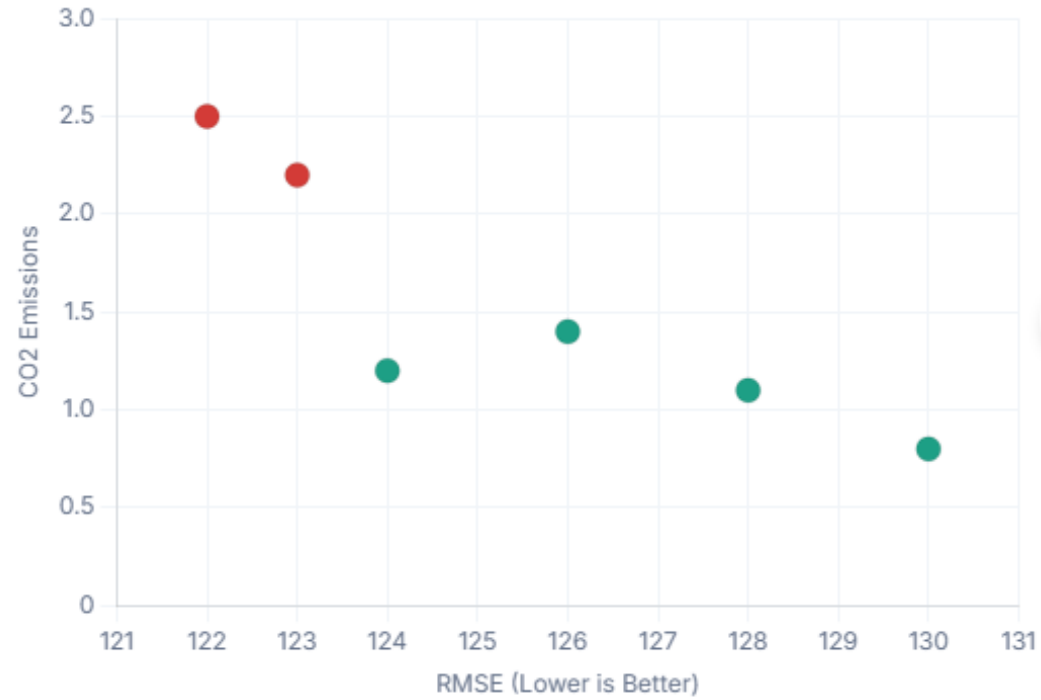
Performance Analysis

Balancing model accuracy with sustainability and compute budget

Profiler Results

Performance vs. CO2 Trade-off

H2O FLAML



RMSE (Lower is Better) vs. Emissions

Budget Impact (RMSE over Time)

H2O FLAML



Pipeline Performance Comparison

⚖️ Accuracy vs. Sustainability Trade-off

💡 Key Observation

H2O GBM achieves the lowest RMSE (best accuracy) but consumes significantly more energy due to its extensive grid search strategy.

FLAML XGBoost offers a "Pareto Efficient" alternative: only 1.2% higher error but **45% lower CO₂ emissions**.

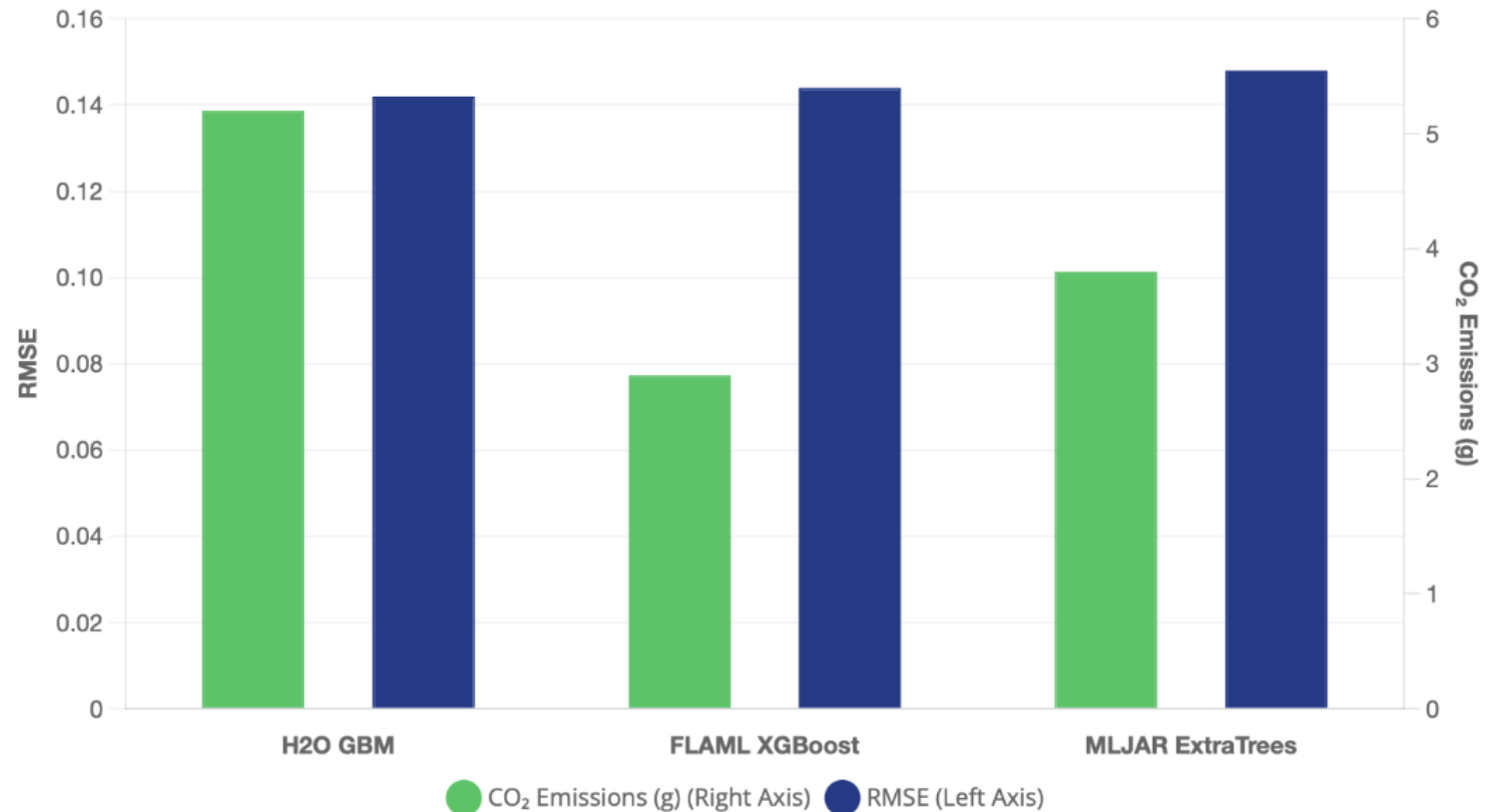
Top Pipeline Details

PIPELINE	RMSE ↓	ENERGY (WH)	CO ₂ (G)
H2O GBM	0.142	12.8	5.2
FLAML XGBoost	0.144	7.1	2.9
MLJAR ExtraTrees	0.148	9.4	3.8

📄 Metrics averaged over 5-fold cross-validation

Multi-Metric Comparison

Comparing predictive error against environmental cost per training run.

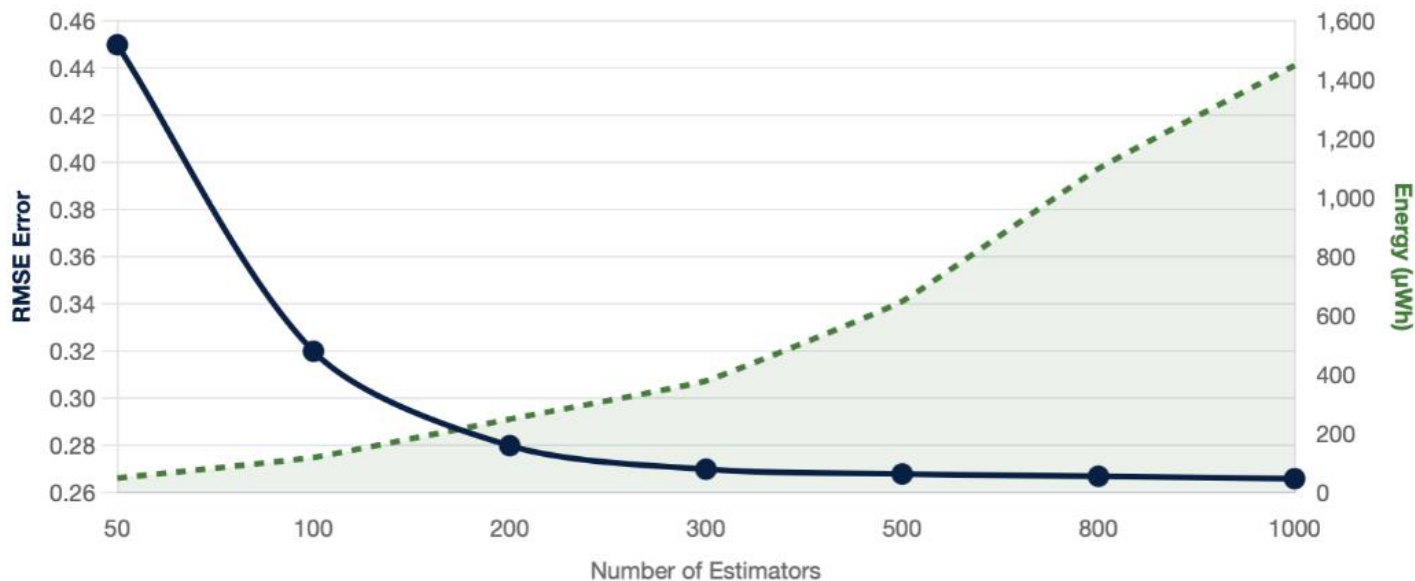


Hyperparameter Effects Analysis

Impact of Tree Count (n_estimators)

Correlation between model complexity, error rate, and energy

● RMSE (Left) ● Energy (Right)

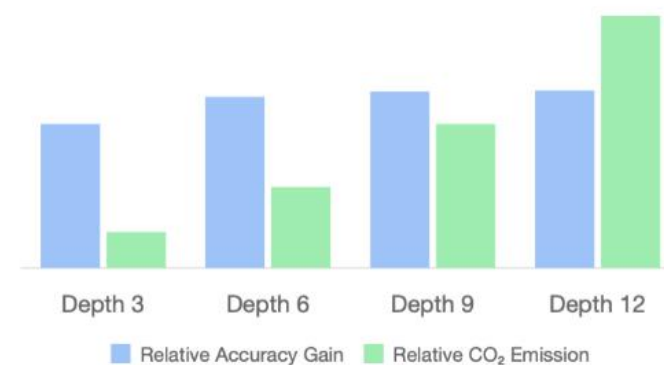


💡 Analysis Finding

Increasing estimators beyond 300 yields negligible RMSE improvement (<0.5%) but causes a linear spike in energy consumption. The optimal "Green Point" is identified at n_estimators=200-300.

Tree Depth Impact

max_depth vs. Training Cost



Deeper trees exponentially increase CO₂ emissions



SustainaML Recommendation

Based on the profile of this dataset (Oil & Gas Production):

- ✓ **Limit Search Space:** Cap n_estimators at 400.
- ✓ **Constrain Depth:** Set max_depth between 4 and 8.
- ➔ **Result:** Saves ~40% energy with identical accuracy.

Multi-Budget Evaluation & Trade-offs

CO₂ Emissions vs. Energy Consumption

Data aggregated across FLAML, H2O, MLJAR



The "Early Wins"

10S - 30S BUDGET

Simple models (Decision Trees, GLM) train rapidly with minimal carbon footprint (~5g CO₂). They achieve 85% of potential accuracy in seconds.



Resource Inflation

300S BUDGET

Extending search to 300s consumes 30x more energy but improves RMSE by only ~1.2%.

Recommendation

For routine daily production forecasts, cap AutoML budgets at 120s.

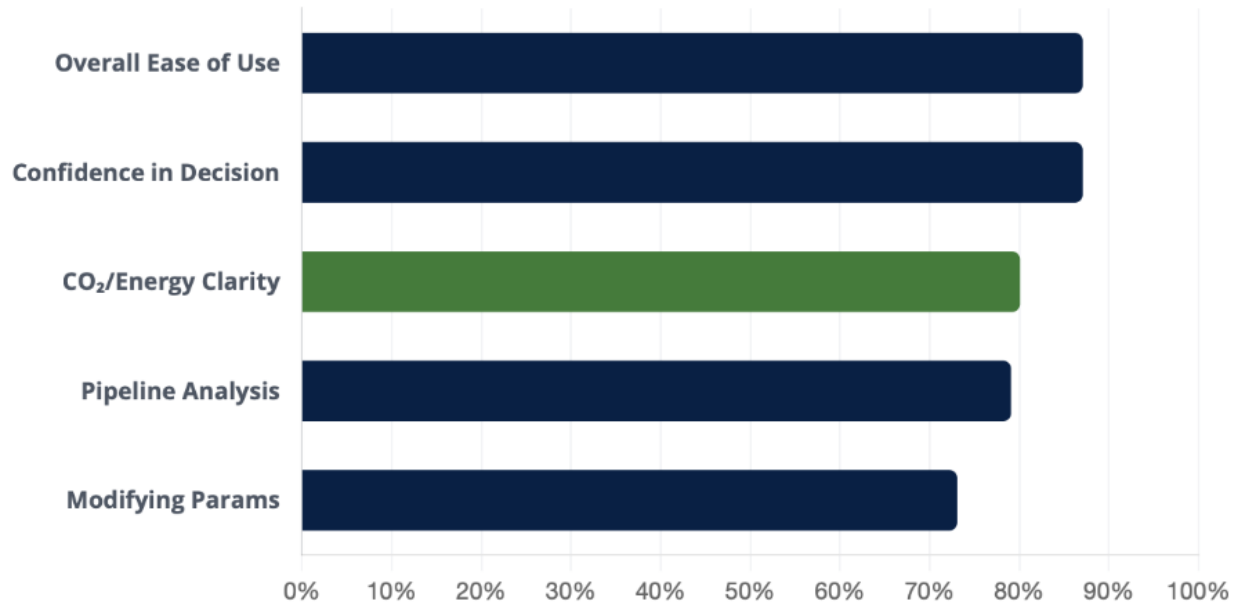
SAVE: 150 kWh/year per model

User Study Results

PARTICIPANTS
N = 15 (ML/DS Experts)

👍 System Usability (Q1-Q9)

Percentage of participants rating "Agree" or "Strongly Agree"



“Confidence in decision-making was high (87%), and participants specifically noted the clarity of energy & CO₂ metrics.”

✅ Decision Support (Q10-Q16)

Task completion success rates using SustainaML

PERFORMANCE
 Identify Best Algorithm **100%**
 (15/15)

SUSTAINABILITY
 Identify Lowest CO₂/Energy **100%**
 (15/15)

TRADE-OFFS
 Assess Budget Impact **100%**
 (15/15)

EXPLAINABILITY
 Identify Most Influential Feature **93%**
 (14/15)

SELECTION
 Identify Suitable Framework **80%**
 (12/15)

Practical Takeaways

🏠 For Oil & Gas Teams



Redefine Success Metrics

Move beyond just RMSE/MAE. Mandate **Dual Optimization** where energy consumption (kWh) and CO₂ emissions are primary KPIs alongside accuracy.



Apply Domain Constraints

Restrict AutoML search spaces using physics-based knowledge. Exclude algorithms or features that violate known reservoir engineering principles.



Mandatory XAI Audits

Before deployment, validate top models with **SHAP/LIME**. Ensure feature importance aligns with operational reality (e.g., pressure vs. flow rate).



Adopt Pareto Selection

Select models at the "knee" of the trade-off curve. Accept marginal accuracy drops (e.g., 0.1%) for significant sustainability gains (e.g., -40% Energy).



Track MLOps Emissions

Integrate tools like **CodeCarbon** into CI/CD pipelines. Log the environmental cost of every training run

Why It Matters

Trustworthy AI is not just about accuracy; it's about transparency and responsibility.

By adopting these practices, Oil & Gas firms can reduce compute costs, meet ESG targets, and build models that engineers actually trust.

🚀 Quick Win

Start by standardizing time budgets.

Setting a hard cap of 120s for daily runs immediately cuts wasted energy by up to 60% with minimal performance loss.

Hands-on: How to Run the Demo



PREREQUISITES

Python 3.8+ & Git Required

1



Get Started

Clone the repository and install dependencies via pip.


```
$ git clone ...  
$ pip install -r requirements.txt
```

2



Load Data

Launch UI server and upload your CSV dataset.

 oil_production.csv
Select Target: "Volume"

3



Configure

Select AutoML frameworks and set time budgets.

FLAML H2O MLJAR

4



Analyze

Review pipelines and sustainability metrics.

RMSE CO2

Resources & Links

Scan QR codes to access the SustainaML repository, read the full paper, or watch the tutorial video.

✓ Open Source (MIT License)



GitHub Repo
Source Code



Research Paper
ECML PKDD 2025



Tutorial Video
Walkthrough

Q&A and Next Steps

Thank you for attending. We welcome your questions on XAI, SustainaML, and potential industry applications.

How We Can Collaborate



Tailored Industry Demo

Customize the SustainaML workshop for specific segments (Upstream Exploration, Midstream Logistics, or Downstream Refining) to address your team's unique challenges.



NDA-Based Pilot Project

Run a confidential pilot using your proprietary data. We'll help benchmark your current models against sustainable AutoML alternatives.

Get in Touch



Mehak Mushtaq Malik
SustainaML Lead Developer

✉ mehak.mushtaq.malik@ut.ee



Assoc. Prof. Radwa El Shawi

Head of Data Systems Group

✉ radwa.elshawi@ut.ee

