Industrial Data Analytics: Core Machine-Learning Concepts for the Chemical-Engineering Curriculum

Mattia Vallerio

CMIC Department "Giulio Natta", Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan 20133, Italy

E-mail: *mattia.vallerio@polimi.it*

Industry 4.0 demands that tomorrow's "Chemical Engineer 4.0" command not only thermodynamics and transport phenomena, but also machine-learning (ML) methodology. This abstract defines the industrial data analytics know-how into a concise syllabus outline, identifying the core ML concepts that a modern chemical-engineering curriculum should impart.

1. **Data stewardship and preprocessing** – Students must grasp data hierarchies (ANSI/ISA-95), historian structures, sampling rates, and contextual metadata. Skills in cleansing, filtering, alignment, and outlier treatment form the indispensable first step of any ML workflow.
2. **Supervised learning** – Regression (linear, decision-tree, random-forest, gradient-boosting, Gaussian-process) and classification (logistic, support-vector, neural-network) algorithms embody the digital evolution of empirical correlations familiar from heat-transfer and distillation design. Lectures should emphasise bias–variance trade-offs, cross-validation, and hyperparameter tuning, illustrated with soft-sensor creation for online quality prediction.
3. **Unsupervised learning** – Clustering (k-means, DBSCAN, self-organising maps) and density estimation map normal operating envelopes, enabling alarm rationalisation and early-warning systems. Students should practise using cluster centroids for actionable control-room guidance.
4. **Reinforcement learning and adaptive control** – Bridging advanced process control and operations research, reinforcement learning formalises how an agent (controller) optimises cumulative reward (profit, emissions) under plant dynamics and constraints. Simple Q-learning on a CSTR and policy-gradient tuning of an MPC layer ground the theory in familiar equipment.
5. **Hybrid physics/ML modelling and uncertainty quantification** – Combining first-principles balances with neural-network residuals or Gaussian-process modifiers preserves extrapolation safety while capturing plant non-idealities. Propagating sensor noise through Monte-Carlo simulations trains students to report confidence intervals, not point predictions.
6. **Explainability, ethics and deployment** – SHAP values, partial-dependence plots, and counterfactuals turn black-box outputs into transparent recommendations that satisfy process-safety and regulatory audits. Coverage of cyber-secure cloud/edge deployment, model lifecycle management, and ethical considerations closes the loop from classroom to control room.

All elements in the course should be taught with hands-on chemical engineering examples. The flipped-classroom projects can include: compressor surge detection, batch-fermentation optimisation, and carbon-footprint minimisation. Animate. By equipping graduates with this ML foundation, universities will prepare them to orchestrate data-driven improvements across design, operations, and sustainability. The proposed syllabus transforms ML from a buzzword into a rigorous, hands-on engineering toolset.

Keywords: *Machine Learning, Industry 4.0, Chemical Engineering 4.0, Data Analytics*