IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4649925/
Vol. 2025 (34), Article ID 4649925

Machine Learning in Research Practice, From Small Data to Meaningful Insights*

Grzegorz SŁOWIŃSKI

VIZJA University, Warsaw, Poland,

Correspondence should be addressed to: Grzegorz SŁOWIŃSKI, g.slowinski@vizja.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

Machine learning (ML) is increasingly recognized as a powerful tool for scientific discovery, yet practical guidance for working with small to medium-sized datasets is limited. This article fills that gap by presenting lessons drawn directly from the author's hands-on experience across natural and social sciences, including biomedical signal analysis, survey research, and behavioral prediction. Unlike a literature review, it reflects real-world applications, illustrating what works in practice and what pitfalls to avoid.

The study outlines a structured ML workflow emphasizing careful data preparation, model selection, rigorous validation, and interpretability. Both shallow and deep models are considered, with advanced techniques such as SHAP used to reveal how models make decisions and extract meaningful insights.

Results highlight that effective ML depends less on algorithmic complexity and more on disciplined methodology. Interpretable models, integration with domain knowledge, and thoughtful validation often outperform more sophisticated alternatives on modest datasets. Common challenges—including data leakage, default-model overreliance, and misconceptions of ML as an automatic solution—are addressed with practical examples.

This article serves as a guide for researchers who want to apply ML responsibly and effectively, demonstrating how real-world experience can transform small or imperfect datasets into scientifically meaningful insights.

Keywords: machine learning, small data, scientific research, interpretability, applied experience

Introduction

Over the past ten years, I have been working in academia, initially focusing primarily on teaching computer science, especially university-level programming courses. A sustained interest in artificial intelligence (AI) emerged around five years ago, leading to my first applied machine learning projects. Early studies included an analysis of the Dry Beans dataset (Słowiński, 2021) and a credit card fraud detection project (Słowiński & Wąsik, 2022), providing practical experience with typical AI workflows, such as preprocessing, handling imbalanced data, and evaluating model performance.

My work later expanded to include a project analyzing Raman spectra for distinguishing COVID from non-COVID samples (Szymborski et al., 2024). The most significant acceleration in AI-related research occurred over the past year following a change in institutional affiliation, involving intensive projects in the social sciences and psychology (Słowiński, Szopiński, & Wilczewski, in press) focused on behavioral prediction, survey analysis, and explainable machine learning.

Across these stages, early applied AI projects, biomedical signal analysis, and recent social science research, I have accumulated a broad set of practical insights. These experiences form the basis for the reflections and methodological guidance presented in this article.

Cite this Article as: Grzegorz SŁOWIŃSKI, Vol. 2025 (34) "Machine Learning in Research Practice, From Small Data to Meaningful Insights " Communications of International Proceedings, Vol. 2025 (34), Article ID 4649925, https://doi.org/10.5171/2025.4649925

The aim of this article is to provide guidance for researchers without extensive AI expertise. It serves as a practical introduction to the potential of AI in scientific research, helping scholars from diverse fields, humanities, technical, or natural sciences,gain insight into its applications. Rather than prescribing specific solutions, the article illustrates possibilities and supports informed reflection on AI's role in research.

While detailed results cannot be disclosed, the article draws on these experiences to show how AI can support scientific investigations and be integrated effectively into the research process. It emphasizes both opportunities and limitations, highlighting potential risks and offering practical guidance on maximizing benefits while minimizing errors and unintended consequences.

Artificial Intelligence: A Broad View

In recent years, artificial intelligence (AI) has become closely associated, at least in public perception, with large language models (LLMs) such as GPT, Claude, LLaMA, or Gemini. Many researchers outside computer science now use these tools for drafting documents, analyzing texts, summarizing literature, or exploring ideas. While generative models represent an important stage in AI's evolution, they form only one branch of a much broader and older research landscape.

The term artificial intelligence predates LLMs by nearly seven decades. Coined by John McCarthy during the Dartmouth Summer Research Project on AI (McCarthy et al., 1955), it marked the symbolic beginning of AI as a scientific field. In its early decades (1950s–1980s), AI research focused on symbolic reasoning, formal logic, manually designed rules, and expert systems, attempts to encode intelligence directly. With the rise of machine learning in the 1990s and 2000s, AI shifted toward statistical pattern recognition and data-driven models capable of generalizing from examples rather than explicit instructions.

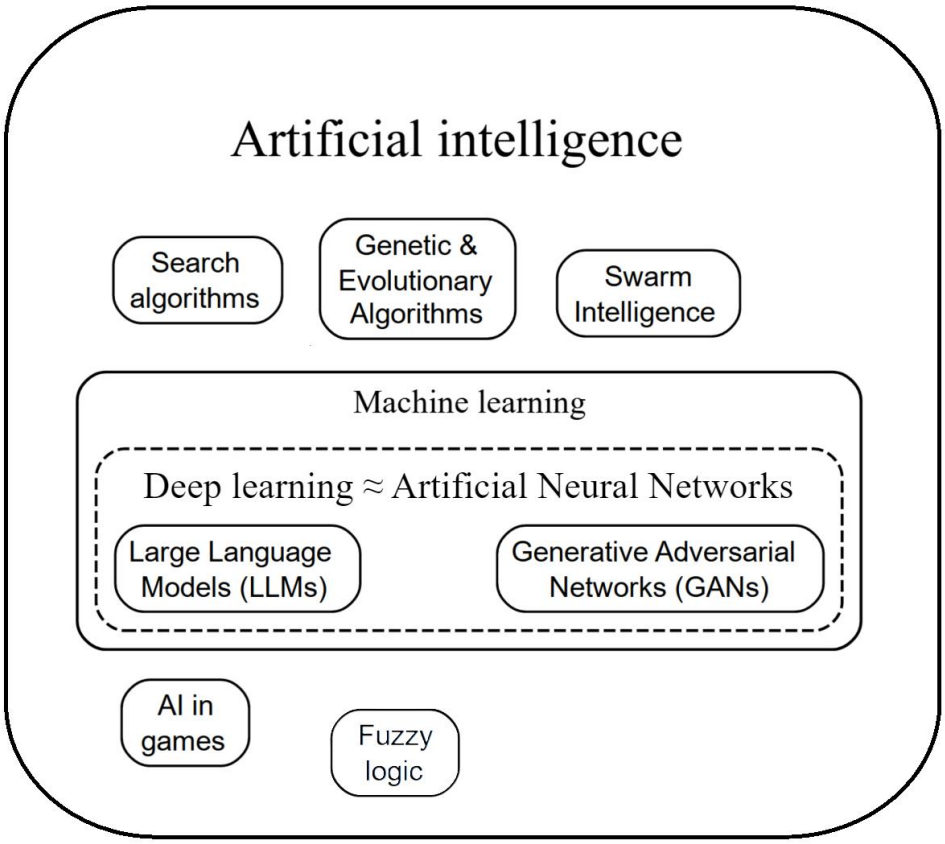

Figure 1 Categorization of concepts within AI, inspired by Hurbans (2020).

A synthetic overview of AI helps researchers appreciate this diversity. Hurbans (2020) presents a structured survey of AI approaches, from rule-based systems and classical machine learning to neural networks, deep learning, reinforcement learning, and generative models. His schematic illustration (Figure 1) depicts AI as a large circle with nested subsets corresponding to these techniques, providing a clear framework to understand the spectrum from human-coded rules to autonomously learned patterns. Personally, Hurbans' work has been invaluable in organizing these concepts, which is why I highlight it here.

Machine Learning: Shallow and Deep Approaches

Machine learning (ML) differs from traditional programming: rather than applying pre-defined rules to data, ML trains models on input–output pairs, allowing algorithms to infer patterns and generalize to new, unseen data (Figure 2). This principle underpins both shallow and deep learning approaches.

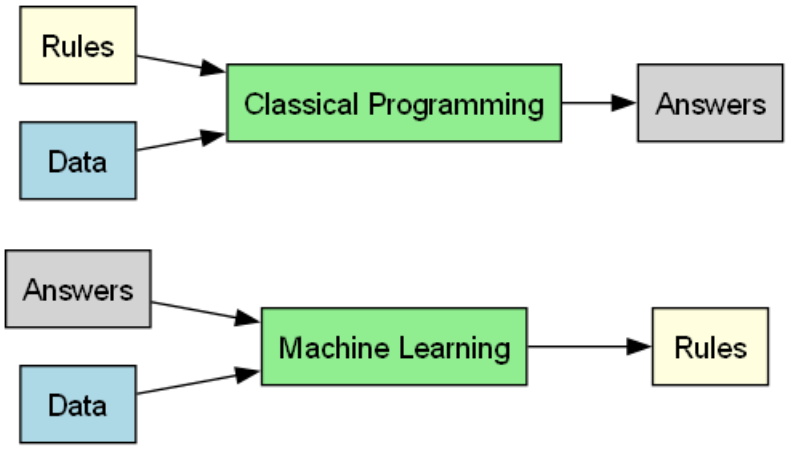

Figure 2. Illustration of the difference between traditional programming and machine learning. In traditional programming, rules are defined by humans and applied to inputs to generate outputs; in machine learning, rules are inferred from known input–output pairs to produce a model capable of predicting outputs for new inputs, own elaboration.

Deep learning, often associated with neural networks, receives the most attention due to its apparent similarity to an “artificial brain” and its role in large language models (Vaswani et al., 2017). However, model complexity must be balanced with data availability. Complex models with many parameters require large datasets; otherwise, they risk overfitting, performing well on training data but poorly on new samples. Conversely, overly simple models may underfit, failing to capture meaningful patterns (Figure 3).

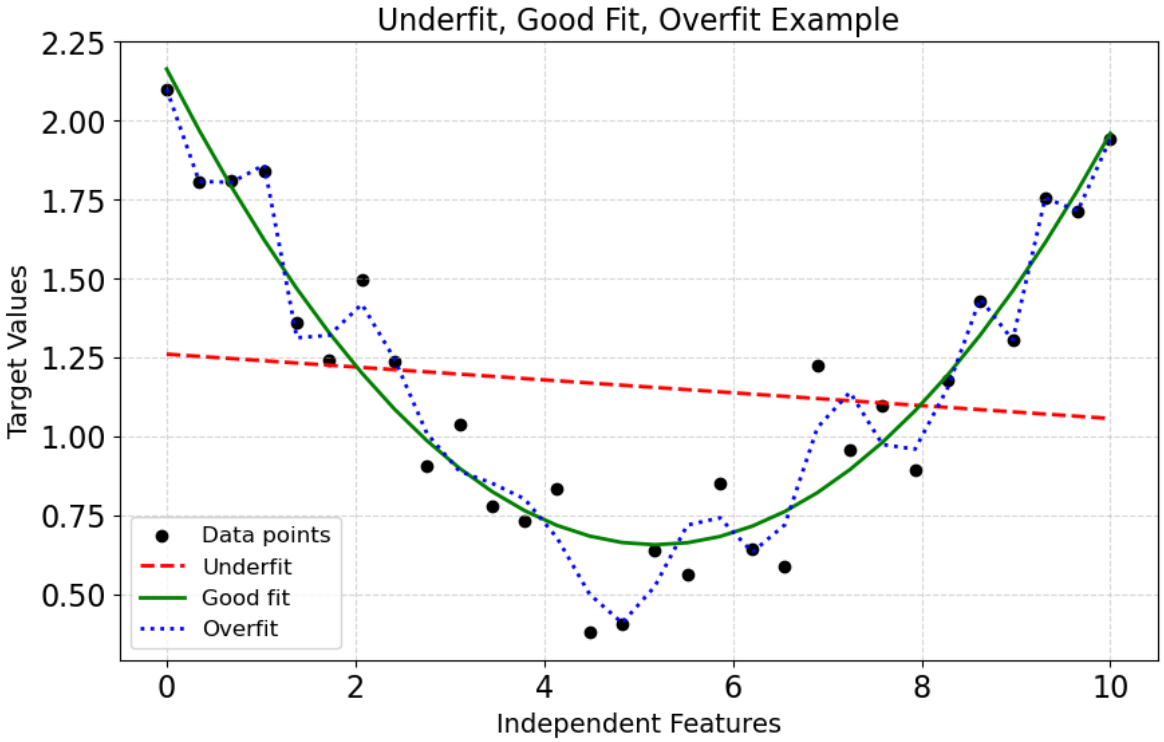

Figure 3. Illustration of underfitting, optimal fitting, and overfitting, *own elaboration.*

In many scientific applications, datasets are small or moderate in size. Here, shallow models often provide the best solution: they are interpretable, computationally efficient, and capable of learning meaningful patterns without overfitting, provided that proper validation is applied. Shallow learning supports hypothesis generation, variable relationship identification, uncertainty quantification, and integration with domain knowledge, complementing deeper approaches where appropriate.

Training and Testing: Ensuring Generalization

A core principle in ML is testing models on data not seen during training. A dataset is typically split into training and test sets; the model is trained on the former and evaluated on the latter. This ensures the model generalizes rather than memorizes patterns. Learning curves (Figure 4) help assess whether a model is underfitted, overfitted, or well-generalized. Proper separation of training and test data is essential for creating reliable models capable of providing meaningful insights from new observations.

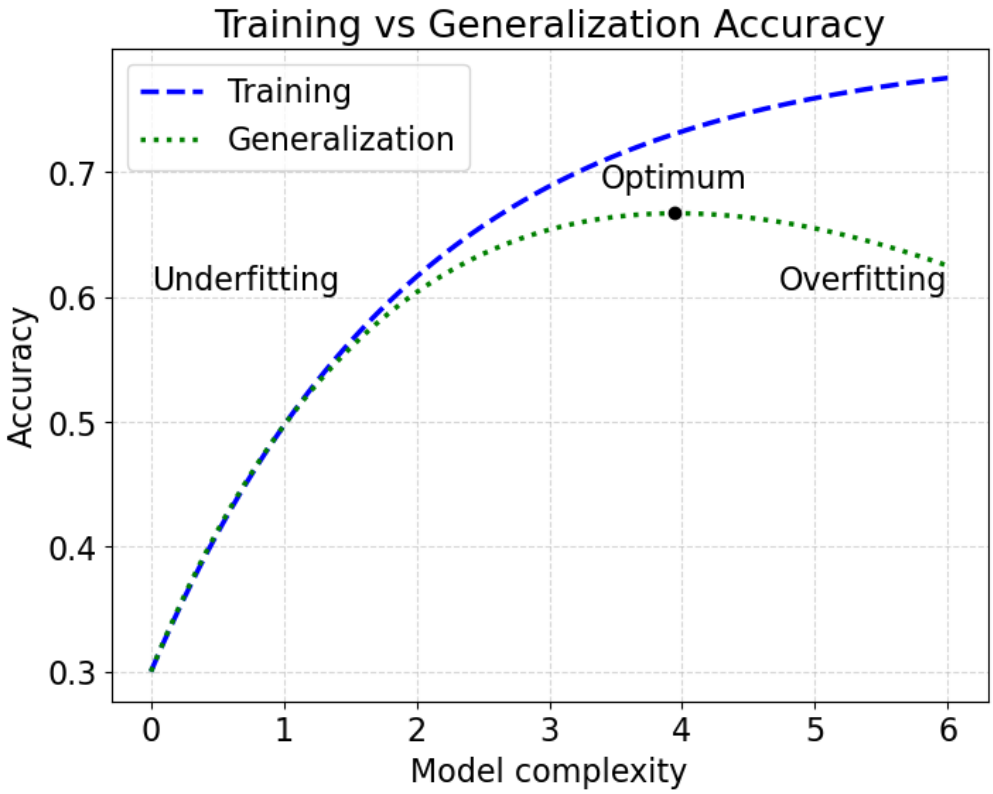

Figure 4. Learning curves showing underfitting, overfitting, and optimal fitting based on training and test performance, own elaboration.

Typical workflow

A typical workflow in applied machine learning involves several key steps, each crucial for obtaining meaningful and interpretable results.

Data Preparation: Raw data must be carefully cleaned and transformed to ensure compatibility with the modeling process. This includes selecting relevant features, encoding categorical variables, standardizing numerical features, preprocessing text fields, handling missing or inconsistent values, and documenting any transformations. Proper data preparation ensures that models can learn effectively and that subsequent analyses are reliable.

Exploratory Data Analysis (EDA): Initial examination of the dataset includes checking variable distributions, identifying multicollinearity through correlation analysis, and visualizing relationships using plots such as correlation matrix or pairwise plots (Figure 5). These steps help detect potential issues early and provide intuitive insight into the data structure, guiding informed decisions in the modeling stage.

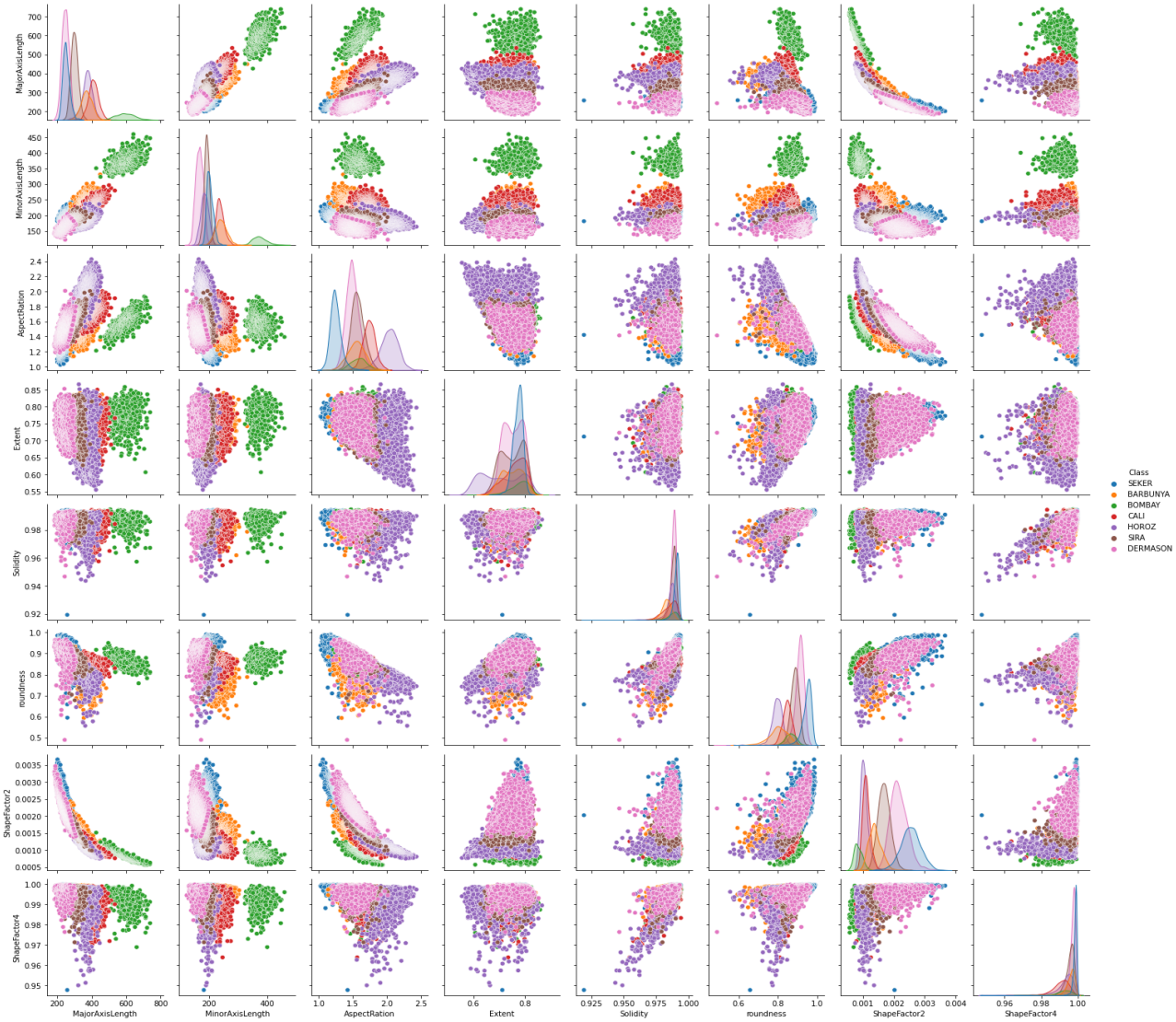

Figure 5. Pairplot of dry bean features. Reproduced from the author's previous work (Słowiński, 2021), CC BY 4.0.

At this early stage, Principal Component Analysis (PCA) can also be a useful exploratory tool. By projecting high-dimensional data onto the first two principal components, PCA provides a two-dimensional visualization of the sample distribution, allowing researchers to see how distinct or overlapping different classes appear in the reduced space. This makes PCA particularly informative for classification problems, offering an initial sense of class separability before any model is trained (Figure 6).

It is important to note, however, that PCA captures only linear relationships in the data and maximizes variance rather than class separation. As a result, good separation in PCA space does not necessarily imply that the classes are easily separable in the original feature space, and poor separation may simply reflect nonlinear boundaries that PCA cannot represent. Moreover, PCA is not directly compatible with categorical variables unless they are encoded appropriately. Despite these limitations, PCA remains a valuable first diagnostic step, providing an intuitive visual summary of complex datasets.

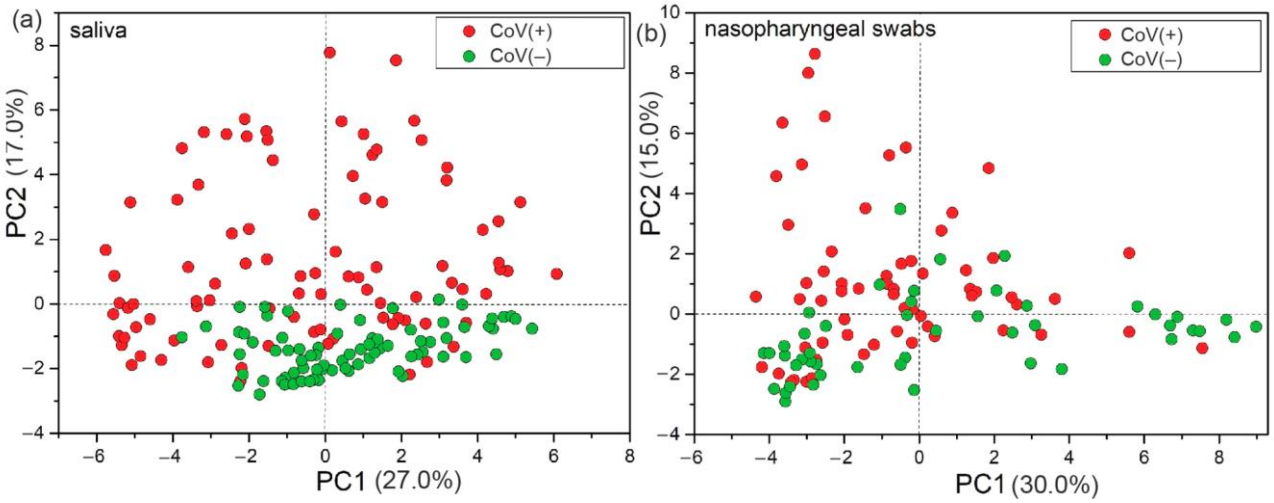

Figure 6. Visualization of saliva (a) and nasopharyngeal swab (b) datasets after PCA transformation to 2D space. Red dots represent samples infected with the SARS-CoV-2 virus (CoV(+)), and green dots represent non-infected samples (CoV(−)). Adapted from Szymborski et al. (2024), CC BY 4.0.

Model Building and Evaluation: Depending on the task: binary or multiclass classification, or regression, models are trained and assessed using appropriate metrics. For regression, L2 loss is typically the primary measure, with mean squared error (MSE) or mean absolute error (MAE) providing complementary information. In classification tasks, accuracy alone may be insufficient, especially for imbalanced datasets. Precision, recall, F1-score, and confusion matrices offer a more complete evaluation of model performance. Single-number metrics summarize predictive quality but do not reveal *how* a model errs. Confusion matrices (figure 7), by contrast, make error patterns visible. For example, whether certain classes are systematically confused with others, which can guide model refinement or suggest structural issues in the data. This deeper view is often essential when the objective is not only to obtain high accuracy but also to understand the behavior and limitations of the model. For instance, in figure 7 the most frequent misclassifications occur between the Dermason and Sira varieties, indicating that these two classes are particularly difficult for the model to distinguish. Such insights can inform system design: a practitioner might consider augmenting the feature set with additional attributes capable of separating these two categories more clearly, for example, color-related characteristics or other domain-specific measurements that capture their distinctive properties.

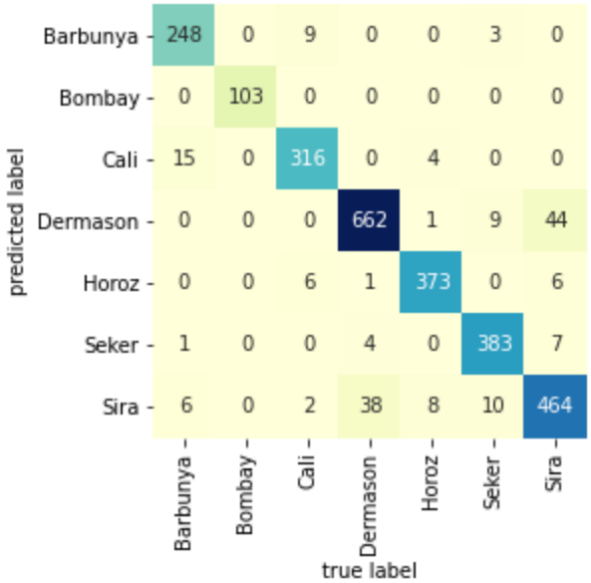

Figure 7. Confusion matrix for the random forest classifier on the dry beans test subset, illustrating the distribution of correct and incorrect predictions across outcome categories, reproduced from the author's previous work Słowiński (2021), CC BY 4.0.

Model Interpretation: White-box models, such as linear regression or decision trees, allow direct inspection of predictors' influence. Black-box models require interpretability techniques such as Feature Importance and SHAP (SHapley Additive exPlanations), which decompose predictions into feature contributions, showing both magnitude and direction of influence. This enables researchers to understand which features drive outcomes and how changes in feature values affect predictions, enhancing the interpretability of complex models.

Validation and Hyperparameter Optimization: k-fold cross-validation ensures that performance estimates are stable and not dependent on a particular train-test split. Hyperparameter tuning via grid search systematically explores combinations of model parameters to maximize predictive performance and generalizability. These steps enhance the rigor and reliability of applied ML workflows, helping ensure that models are both accurate and scientifically interpretable.

Lessons Learned

Applying machine learning in scientific research offers several lessons that are often overlooked by beginners.

First, models can only learn patterns that exist in the data. Machine learning is not magic; it detects actual regularities. Sophisticated algorithms, from linear models to ensembles, cannot extract meaningful relationships from noise. If a predictor is unrelated to the target, no model will uncover a connection.

Second, interpretation matters more than prediction. In scientific work, the goal is not merely to maximize accuracy, but to gain insights into the phenomenon under study. Predictive performance can usually be improved later with additional data, refined measurements, or new variables. More important is understanding how predictors relate to outcomes and using models as instruments for learning rather than as ends in themselves.

Another important aspect is the strategic use of models to enhance understanding. A common approach involves first training a model with reasonably good predictive performance to capture meaningful structure, and then examining its decisions using interpretability techniques, such as SHAP, to identify influential variables and relationships. Often, multiple models with different assumptions, linear, tree-based, and regularized, are applied. When these models highlight similar patterns, confidence increases that the detected relationships reflect genuine structure rather than artifacts of a specific algorithm. The author arrived at this approach gradually and, after considerable practice, achieved fluency in applying it effectively. Its full implementation was realized in a recent project on the adaptation of international students, where the results proved highly satisfactory to all co-authors, reinforcing the intention to apply this strategy consistently in future studies.

This ensemble approach reduces the risk of overinterpreting algorithm-specific behaviors and supports robust scientific conclusions. Machine learning thus becomes a tool for generating knowledge, uncovering mechanisms, and guiding hypothesis development, rather than merely optimizing predictive metrics.

A Brief Overview of Commonly Used Models

At this point, it is useful to provide a short overview of the types of models most frequently applied to tabular scientific data. François Chollet (2017) suggests that for tabular datasets, the state-of-the-art baseline is typically a gradient-boosted decision tree model, such as XGBoost. However, my own experience in scientific collaborations does not fully confirm this as a universal rule. XGBoost and its relatives tend to dominate when datasets are large, often comprising tens or hundreds of thousands of observations, allowing these highly flexible models to learn subtle nonlinear patterns. In academic projects, however, datasets are usually much smaller. Many empirical studies are constrained by cost, limited population sizes, or experimental limitations: surveys with a few hundred respondents, patient studies with 150–300 individuals, or experiments where observations are expensive to produce. In such settings, simpler models often match or outperform more complex alternatives while providing clearer interpretability.

Most machine-learning algorithms come in variants suitable for three fundamental problem types: binary classification (e.g., participated vs. did not participate in a boycott), multiclass classification (e.g., predicting the variety of a dry bean), and regression (e.g., predicting a student's life-satisfaction score on a 1–6 scale). Common frameworks such as scikit-learn provide both classification and regression versions of most algorithms.

Another important distinction is the degree of interpretability, often framed as white-box vs. black-box models. White-box models are transparent: we can inspect how they arrive at predictions. Linear models and their regularized variants, linear regression, logistic regression, and Bayesian linear models, remain prominent examples. Regularization, such as Elastic Net (combining L1 and L2 penalties), helps control overfitting when predictors are numerous relative to observations. Linear model coefficients are directly interpretable: positive coefficients push the prediction upward, negative ones push it downward, and their magnitude indicates the strength of the effect.

Another white-box example is the single decision tree, figure 8. While a single tree rarely achieves the highest predictive accuracy compared to ensemble methods, it offers unique interpretability. Small trees (typically 3–4 levels) can be visualized as a graph, allowing researchers to inspect initial splits and identify the strongest predictors. Even if overall accuracy is moderate, the clarity of a decision tree makes it invaluable for exploratory analysis, hypothesis verification, and communicating results.

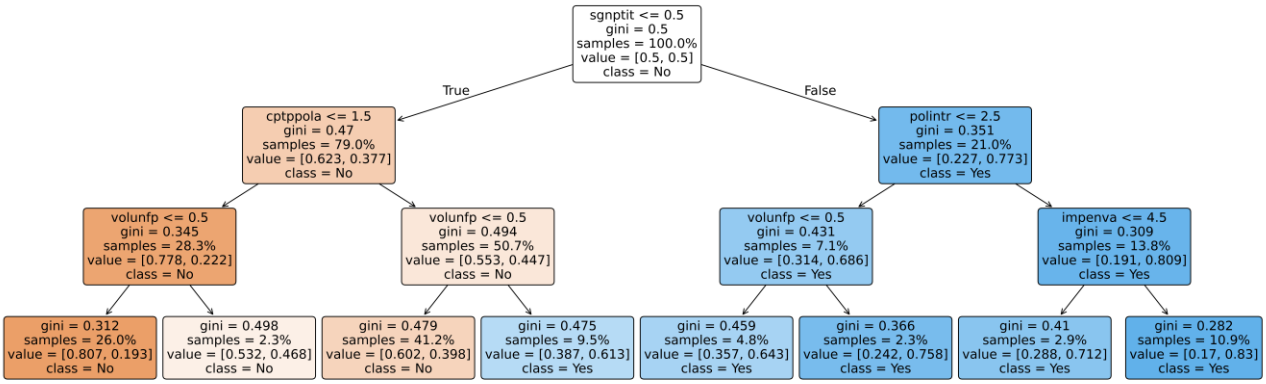

Figure 8. Decision tree for predicting boycott behavior, reproduced from Słowiński et al. (in press) with permission.

Black-box models, by contrast, are more opaque. Extracting information about their internal logic is not straightforward, though ensemble methods such as Random Forests and Gradient Boosting (including XGBoost) can be interpreted to some degree. In smaller datasets, XGBoost may not provide a significant advantage over simpler alternatives.

An important advantage of tree-based models, both individual trees and ensemble algorithms, is their inherent ability to capture nonlinear relationships and complex interaction effects. Because they partition the feature space recursively, they can represent highly irregular decision boundaries without requiring manual feature engineering. This ability often makes random forests and gradient boosting among the strongest-performing models in practical machine learning tasks, especially when nonlinear patterns dominate the data. Their strong predictive metrics reflect not only robustness but also the computational challenge they pose, which is why they are often considered among the most demanding model families in applied research.

Another important family of algorithms within supervised learning is Support Vector Machines (SVMs), which are well known for their solid theoretical foundations and their ability to model nonlinear relationships through kernel functions. In my own applied work, however, SVMs have proven less practical. They are often computationally expensive, particularly when applied to larger datasets, and in most of my experiments they do not rank among the top-performing models in terms of predictive accuracy. Unlike tree-based approaches, they also lack straightforward interpretability, which limits their usefulness in research contexts where explanation and variable-level insight are essential.

Black-box model interpretation

To interpret black-box models, Feature Importance can quantify how much each predictor contributes to model decisions, for instance, by counting its use in tree splits or measuring impurity reduction (e.g., Gini index). This helps researchers understand which variables strongly influence predictions. However, Feature Importance does not indicate the direction of effect, whether high or low values increase or decrease the outcome.

Advanced interpretability techniques, particularly SHAP (SHapley Additive exPlanations), address this limitation. SHAP decomposes model predictions into contributions from individual features, showing both magnitude and direction of influence. It can be applied to virtually any model, including black-box ensembles, and produces intuitive visualizations that support understanding how features drive predictions.

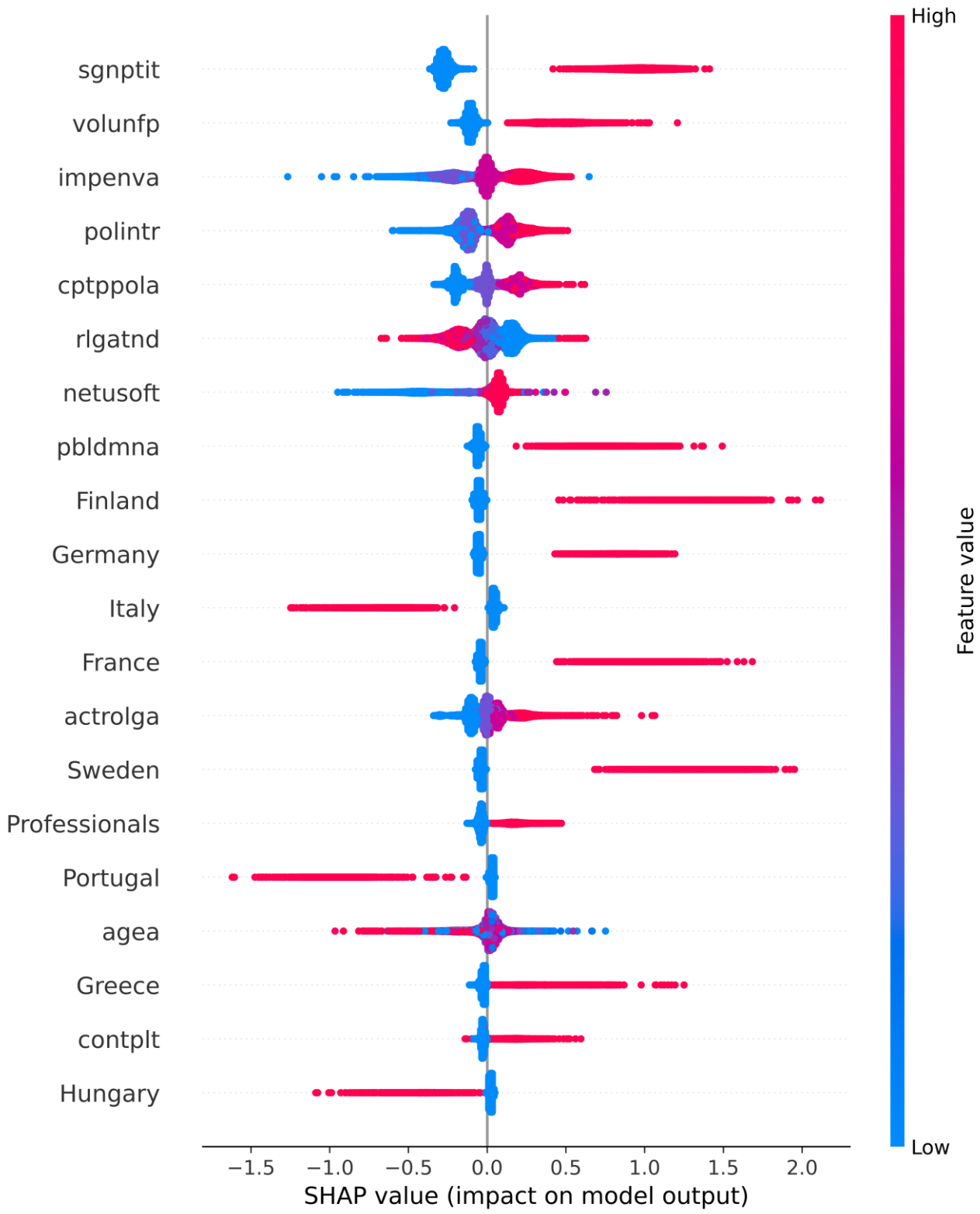

Figure 9. Summary SHAP plot of the 20 most influential features in the Gradient Boosting Classifier. Each dot represents a respondent; x-axis shows the SHAP value, y-axis ranks features by importance, color indicates feature value (red = high, blue = low). reproduced from Słowiński et al. (in press) with permission.

SHAP plots are often color-coded: for example, red points indicate high feature values, blue points indicate low values, and their horizontal position shows the effect on the model output. This enables researchers to see not only

which features matter most, but also how changes in feature values impact predictions, which is crucial when the goal is to understand data relationships rather than merely optimize predictive performance.

Moreover, SHAP visualizations can reveal whether a model is capturing nonlinear patterns. As illustrated in figure 9, many predictors exhibit relatively simple relationships: points of similar color cluster consistently on one side of the vertical axis, indicating a mostly monotonic effect. However, variables such as *netusoft* (internet use intensity) and *rlgatnd* (religious activity) display more complex, curved patterns, suggesting nonlinear or threshold-like effects that a linear model would struggle to represent. Detecting such structure would require explicit transformations in a linear framework, whereas tree-based models can uncover these relationships automatically.

Common Pitfalls and Limitations in Applied Machine Learning

Data preparation is a crucial but often underestimated aspect of applied machine learning. Many newcomers focus primarily on model performance, sometimes perceiving it as almost magical, while the painstaking work of preparing data remains in the background. Raw datasets can be surprisingly challenging to handle, containing redundant or irrelevant columns, missing values, or inconsistent entries, common in surveys conducted across multiple languages or with open-ended questions. Cleaning and standardizing such data often requires custom mappings or dictionaries.

It is frequently said that data preparation consumes the majority of effort, often more than 80%, especially when the data provider is not fully familiar with the dataset or research goals. Close collaboration between the data analyst and the data-generating researchers is therefore essential. Exploratory checks, such as examining correlations or addressing missing values, help ensure the dataset is suitable for modeling and interpretation.

Another limitation is insufficient sample size. Machine learning relies on detecting patterns within a representative population. Small datasets, sometimes fewer than 100 observations, may not provide enough information for meaningful learning, although modest datasets of 300 - 400 samples can often yield reasonable results depending on the context and strength of the underlying relationships. In natural sciences, effects are typically stronger and easier to detect, while in social sciences, weaker effects and uncontrolled factors increase the need for larger samples.

Data leakage is another common issue, occurring when a model has access to information during training that would not be available in real-world predictions. This can arise if features are derived from the target or from future data, or if repeated measurements from the same sample are treated as independent observations. Leakage can give an illusion of high model performance, but the resulting model fails to generalize. Ensuring proper separation between training and test sets is therefore critical.

A related problem is what could be called Model Zoo Behavior. Modern ML tools make it easy to run many models with minimal expertise, which can lead beginners to apply numerous algorithms without sufficient preprocessing, scaling, or hyperparameter tuning. In practice, focusing on a smaller set of well-chosen models and dedicating effort to data preparation and thoughtful interpretation usually produces more meaningful insights.

Finally, the Automation Illusion, the misconception that ML can magically generate results from random or poorly structured data, is common among newcomers and collaborators. Machine learning can detect complex or subtle patterns, but it cannot produce meaningful predictions where no informative signal exists. Awareness of these limitations is essential for applying ML responsibly and effectively in scientific research.

Final Remarks

This article presented a practical and methodologically oriented overview of how machine learning can meaningfully support scientific research. While public discussions increasingly equate artificial intelligence with generative models and large language models, the examples and reflections provided here emphasize that AI is a much broader field, one that long predates today's LLMs and continues to include symbolic approaches, classical machine learning, shallow models, and deep neural networks. Generative AI represents only one branch of this continuum. Understanding this wider landscape helps researchers appreciate the full range of tools available for scientific inquiry.

Drawing on experience from projects in natural sciences, biomedical analysis, and social science research, the article showed that effective use of ML depends less on algorithmic novelty and more on methodological discipline. High-quality data preparation, awareness of dataset size constraints, and rigorous separation of training and test data remain fundamental for building models that generalize reliably. Small or moderately sized datasets,

common across many research domains, often favor shallow or regularized models, which provide interpretability, stability, and integration with domain expertise.

The article also highlighted common challenges encountered by newcomers: the tendency to rely on default model collections ("model zoo"), misconceptions about ML as an automated solution, unrealistic expectations of predictive power with limited data, and the risk of data leakage. Through examples and illustrations, it was shown that meaningful insights depend on thoughtful model selection, appropriate validation, and transparent reporting practices rather than on high-complexity architectures.

Finally, the discussion underscored the value of interpretability techniques, such as SHAP, which bridge the gap between statistical prediction and substantive understanding. When combined with principled workflows: data preparation, exploratory analysis, careful modeling, and interpretability. machine learning becomes a powerful methodological partner in research rather than a black-box tool.

Overall, the article encourages researchers from diverse disciplines to adopt ML not as a technological shortcut, but as a structured analytical framework that enhances scientific reasoning. Used responsibly and critically, machine learning can support the transition from small or imperfect data to meaningful and trustworthy insights.

References

- Chollet, F. (2017) *Deep learning with Python*. Manning Publications, Shelter Island.
- Hurbans, R. (2020) *Grokking artificial intelligence algorithms*. Manning Publications. Shelter Island.
- McCarthy, J., Minsky, M., Rochester, N. and Shannon, C.E. (1955) *A proposal for the Dartmouth summer research project on artificial intelligence*. [Online] Dartmouth College. Available at: http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html (accessed 13 December 2025).
- Słowiński, G. (2021) 'Dry beans classification using machine learning', in *Proceedings of the 29th International Workshop on Concurrency, Specification and Programming (CS&P'21)*, pp. 166–173. CEUR Workshop Proceedings. Available at: http://ceur-ws.org/Vol-2951/paper3.pdf (accessed 13 December 2025).
- Słowiński, G. and Wąsik, K. (2022) 'Credit card fraud detection using machine learning', in Such-Myślnik-Pyrgiel, M., Rogula-Myślnik-Kozłowska, V., Piec, R. and Feltynowski, M. (eds.) *Digital security: From global to local problems. State and society*. Warsaw: Main School of Fire Service, pp. 117–132.
- Słowiński, G., Szopiński, T. and Wilczewski, M. (in press) 'Modeling the probability of consumer boycott participation among Europeans using artificial intelligence', *Contemporary Economics*.
- Szymborski, T.R., Berus, S.M., Nowicka, A.B., Słowiński, G. and Kamińska, A. (2024) 'Machine learning for COVID-19 determination using surface-enhanced Raman spectroscopy', *Biomedicines*, 12(1), 167. https://doi.org/10.3390/biomedicines12010167
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 5998–6008. Curran Associates, Inc. Available at: https://arxiv.org/abs/1706.03762 (accessed 13 December 2025).