Explainable Artificial Intelligence: A Practical Framework for Interpreting and Trusting Machine Learning Models

Debabrata Pruseth

AI Architect & Applied AI Researcher
Singapore

Author Note

This article is a research-style companion version of author's blog post "A Beginner-Friendly Guide to Explainable AI (XAI)" and associated GitHub notebook experimentation "explainable-ai-demo-shap-lime-eli5-dalex-PDP-ICE".

This research presents an applied framework for Explainable Artificial Intelligence (XAI) grounded in both conceptual analysis and an implementation-oriented demonstration using SHAP, LIME, ELI5, DALEX, Partial Dependence Plots, and Individual Conditional Expectation analysis. The work positions explainability as a practical capability for interpreting machine learning decisions, validating model behavior, detecting risk, and supporting trust in AI-assisted decision systems.

The implementation workflow uses a supervised income-classification task based on a tabular **UCI Adult Income dataset** with 32,561 records and 12 model features, including age, education level, marital status, occupation, relationship, race, sex, capital gain, capital loss, working hours, and country. The notebook demonstrates a decision-tree-based explainability pipeline and evaluates the model using recall, precision, F1 score, accuracy, and AUC. The reported model performance is accuracy 0.84947 and AUC 0.873046, supporting its use as a demonstration model for interpretability analysis rather than as a production-grade income prediction system.

This study is implementation-oriented and does not claim a new benchmark, new algorithm, or universal explanation method. Its contribution lies in organizing multiple explainability techniques into a practical framework for model interpretation, stakeholder trust, and responsible AI governance.

Abstract

Machine learning models are increasingly used in decision-making systems where predictive performance alone is insufficient. In domains such as finance, healthcare, hiring, cybersecurity, public administration, and enterprise risk management, stakeholders require

explanations that make model behavior interpretable, auditable, and trustworthy. Explainable Artificial Intelligence (XAI) addresses this need by providing techniques that reveal how models use input features, how predictions are formed, and where model behavior may require additional scrutiny.

This research develops a practical framework for interpreting and trusting machine learning models using a multi-method explainability workflow. The study combines conceptual analysis with an applied implementation using SHAP, LIME, ELI5, DALEX, Partial Dependence Plots, and Individual Conditional Expectation analysis. The implementation uses a tabular income-classification task with 32,561 observations and 12 features. A decision-tree classifier is analyzed through global feature importance, local prediction explanation, decision-path interpretation, model-level performance evaluation, and feature-response analysis. The model achieves an accuracy of 0.84947 and AUC of 0.873046, providing a sufficiently realistic case study for examining explanation behavior without overclaiming predictive validity.

The proposed framework treats explainability as a lifecycle capability rather than a single post-hoc visualization. The analysis shows that no individual XAI method is sufficient on its own: ELI5 provides readable global summaries, SHAP supports additive feature attribution, LIME explains local decision neighborhoods, DALEX supports model diagnostics, and PDP/ICE reveal feature-response patterns. The study concludes that trustworthy machine learning requires triangulated explanations, governance controls, stakeholder-specific communication, and continuous validation.

Keywords: Explainable AI, XAI, SHAP, LIME, ELI5, DALEX, Partial Dependence Plot, ICE Plot, Interpretable Machine Learning, Decision Tree, AI Governance, Responsible AI, Model Transparency, Trustworthy AI, Feature Attribution

---

1. Introduction

Artificial intelligence systems increasingly support decisions that affect individuals, organizations, and public institutions. Machine learning models are now used to estimate credit risk, classify medical images, detect fraud, recommend products, prioritize security alerts, automate document review, and support operational forecasting. In these contexts, a prediction is rarely sufficient by itself. Stakeholders often need to understand why a model reached a specific conclusion, which input factors influenced the decision, and whether the output should be trusted.

The central difficulty is that many machine learning systems are optimized for predictive performance rather than interpretability. As models become more complex, the reasoning behind their outputs can become difficult to inspect directly. This opacity creates the well-known black-box problem: the model produces an output, but the decision logic remains unclear to the user, auditor, regulator, or domain expert.

Explainable Artificial Intelligence seeks to address this problem by making machine learning behavior more transparent. However, explainability should not be treated as a single tool or chart. A SHAP plot, LIME explanation, feature-importance table, or decision-tree diagram

may each reveal part of the model's behavior, but none provides a complete basis for trust. Trustworthy AI requires the integration of explanation techniques with model validation, stakeholder interpretation, governance, and human oversight.

This research develops a practical framework for interpreting and trusting machine learning models using a multi-method XAI workflow. The framework is grounded in an applied notebook implementation that demonstrates SHAP, LIME, ELI5, DALEX, PDP, and ICE on a supervised classification task. The implementation uses a decision-tree model whose root and intermediate splits include variables such as relationship, capital gain, education level, hours per week, marital status, capital loss, and age, illustrating how tree-based models can expose decision logic directly while still benefiting from additional post-hoc explanation methods.

The central research question is:

How can multiple explainability techniques be combined into a practical framework for interpreting machine learning models and supporting stakeholder trust?

---

2. Research Objective

The objectives of this study are:

1. To develop a practical framework for explainable machine learning.
2. To demonstrate how multiple XAI methods complement one another.
3. To distinguish global model interpretation from local prediction explanation.
4. To evaluate the role of SHAP, LIME, ELI5, DALEX, PDP, and ICE in applied model understanding.
5. To identify common failure modes and interpretation risks in XAI workflows.
6. To connect technical explainability with responsible AI governance and stakeholder trust.

---

3. Conceptual Foundation

Explainability refers to the capacity of a machine learning system to provide human-understandable reasons for its behavior. In practical AI systems, explainability serves several functions. It helps technical teams debug models, enables business stakeholders to understand decision factors, supports auditability, reveals potential bias, and allows domain experts to assess whether model behavior is plausible.

A useful distinction exists between **global explainability** and **local explainability**. Global explainability describes the overall behavior of a model across the dataset. Local explainability describes the factors influencing a specific prediction. For example, ELI5 feature importance can summarize which features are influential at the model level, while SHAP, LIME, or ELI5 local prediction explanations can describe why a single observation

was classified in a particular way. The notebook explicitly distinguishes these levels by warning that global importance should not be confused with instance-level explanation.

Explainability also differs from correctness. A model may provide an explanation for a prediction while still being biased, unstable, or wrong. Therefore, explanations must be interpreted as evidence for model understanding, not as proof of model validity.

---

4. Proposed Framework: Multi-Method Explainability Trust Framework

This study proposes a Multi-Method Explainability Trust Framework composed of six layers:

Layer	Purpose	Example Methods
Data and Feature Context	Understand input variables and encoding	Dataset inspection, feature review
Model Behavior Analysis	Examine learned decision structure	Decision tree visualization, split analysis
Global Explanation	Identify overall influential features	ELI5, SHAP summary, DALEX variable importance
Local Explanation	Explain individual predictions	SHAP force plot, LIME, ELI5 prediction explanation
Feature-Response Analysis	Analyze prediction sensitivity	PDP, ICE
Governance Interpretation	Convert explanations into accountable decisions	Documentation, audit review, human oversight

The framework emphasizes triangulation. A single explanation method should not be treated as definitive. Instead, explanations should be compared across complementary methods. If ELI5, SHAP, and DALEX consistently identify similar influential variables, confidence in the interpretation increases. If methods disagree, the disagreement should trigger further model inspection.

---

5. Implementation-Oriented Methodology

The applied workflow begins with a tabular classification dataset containing 32,561 records and 12 features. The features include demographic, employment, education, and financial

variables. Several categorical features are numerically encoded, including workclass, marital status, occupation, relationship, race, sex, and country.

A decision-tree classifier is trained for income classification. The decision tree provides an interpretable baseline because its structure exposes explicit splitting rules. The model's top-level decision logic includes relationship status at the root, followed by capital gain, education level, hours per week, marital status, capital loss, and age in later branches.

The model is evaluated using classification metrics. The reported results are:

Metric	Value
Recall	0.54689
Precision	0.760479
F1 Score	0.636237
Accuracy	0.84947
AUC	0.873046

These results suggest that the model performs reasonably well for demonstration purposes, but the moderate recall indicates that the classifier misses a meaningful proportion of positive-class cases. Therefore, the model is appropriate for XAI demonstration, not for high-stakes deployment without further validation.

---

6. Technical Analysis and Practical Findings

6.1 Decision-Tree Interpretability

The decision tree provides direct structural interpretability. The root split uses relationship status, indicating that this variable strongly partitions the dataset. Subsequent splits involving capital gain, education level, hours worked, marital status, and age show how the model builds hierarchical decision rules.

This illustrates the advantage of intrinsically interpretable models: their internal decision paths can be inspected without external explanation tools. However, tree interpretability decreases as trees become deeper, and simple inspection does not fully explain feature interactions, local prediction uncertainty, or fairness concerns.

6.2 ELI5 for Global and Local Explanation

ELI5 is used to generate readable feature-importance summaries and local prediction explanations. The notebook notes that ELI5's `show_weights()` reveals which features the model relied on most overall, while `show_prediction()` explains one specific prediction using actual feature values.

Weight	Feature
0.5030	Relationship
0.2258	Capital Gain
0.2116	Education-Num
0.0448	Capital Loss
0.0124	Hours per week
0.0020	Age
0.0004	Marital Status
0	Country
0	Occupation

Figure 1: ELI5 Output for Adult Income Model

Figure 2: ELI5 Decision Tree Output for Adult Income Model (Adult #10)

A critical limitation is that ELI5 reports impurity-based importance for decision-tree classifiers. These values are not equivalent to SHAP values and may not capture additive feature effects or interaction structure. This distinction is important for responsible interpretation because feature-importance rankings can differ across methods.

6.3 SHAP for Additive Feature Attribution

SHAP is used to explain how features push a prediction upward or downward relative to a baseline. The notebook explains SHAP force plots using a base value, positive red contributions, and negative blue contributions. Positive SHAP values increase the probability of the positive class, while negative values reduce it.

SHAP is particularly useful because it supports both local and global explanation. Local SHAP values explain a single prediction, while aggregated SHAP values can reveal broader feature influence patterns.

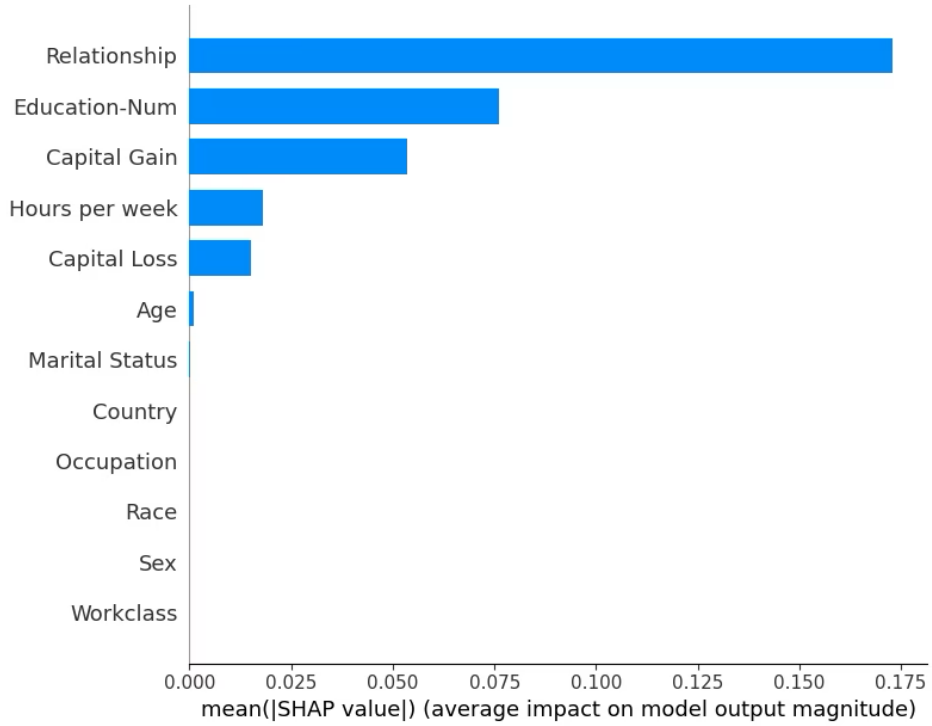

Figure 3: SHAP Bar Plot for Adult Income Model

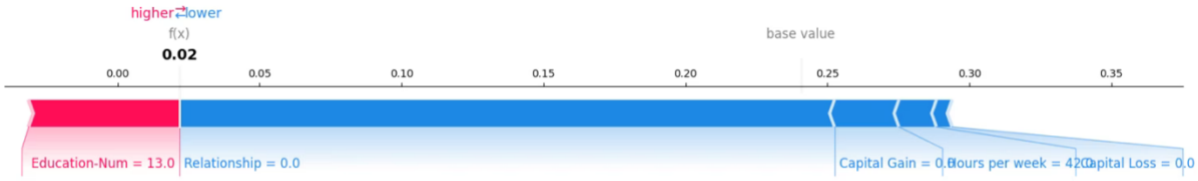

Figure 4:SHAP Force Plot for Adult Income Model (Adult #10)

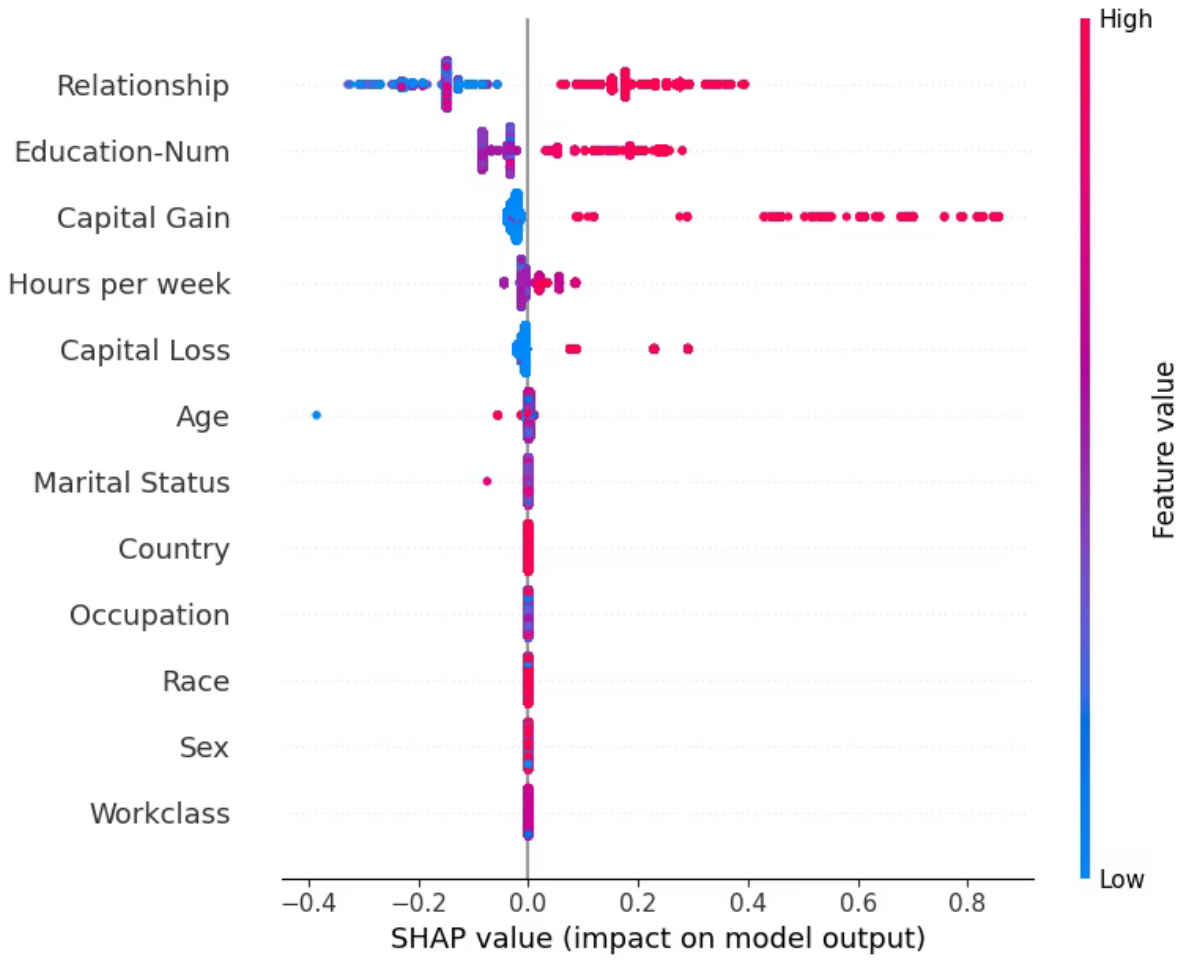

Figure 5: SHAP Beeswarm Plot for Adult Income Model

6.4 LIME for Local Neighborhood Approximation

LIME provides local model-agnostic explanations by approximating model behavior around a specific observation. Its main value lies in explaining why one individual prediction was produced. However, because LIME depends on local perturbations, its explanations can vary depending on sampling configuration, feature encoding, and neighborhood definition.

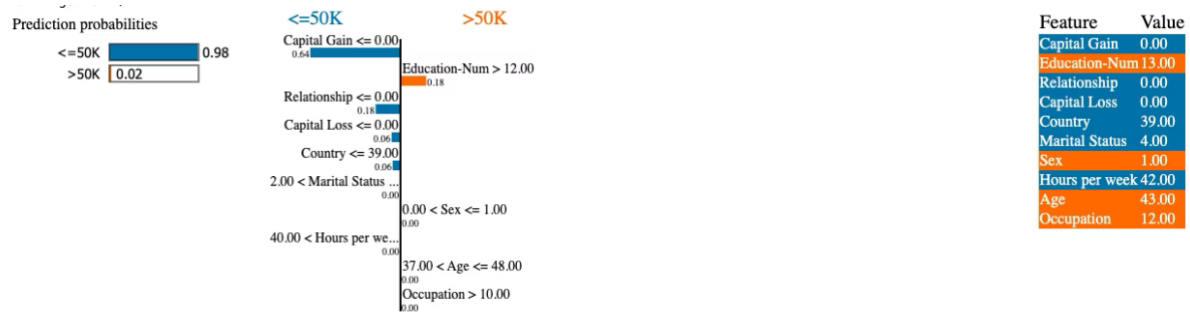

Figure 6: LIME Output for Adult Income Model (Adult #10)

🔍 Local Explanation Comparison for Instance #10

Feature	SHAP Value	LIME Weight
Relationship	−0.2308	−0.1768
Education-Num	0.0530	0.1830
Capital Gain	−0.0227	−0.6410
Hours per week	−0.0132	−0.0019
Capital Loss	−0.0055	−0.0631
Age	0.0002	0.0015
Marital Status	−0.0000	−0.0039
Workclass	0.0000	–
Occupation	0.0000	0.0002
Race	0.0000	–

Figure 7: SHAP vs LIME comparison (Adult #10)

6.5 DALEX for Model Diagnostics

DALEX is incorporated as a diagnostic framework for examining model behavior and prediction-level explanations. The notebook execution includes DALEX prediction-explanation functionality and also exposes library warnings related to future pandas indexing behavior, illustrating an important reproducibility consideration: explanation workflows depend not only on model logic but also on package versions and library behavior.

```
  -> data               : 26048 rows 12 cols
  -> target variable    : 26048 values
  -> model_class        : sklearn.tree._classes.DecisionTreeClassifier (default)
  -> label              : Decision Tree (Income)
  -> predict function   : <function yhat_proba_default at 0x7b74f85e19e0> will be used (default)
  -> predict function   : Accepts pandas.DataFrame and numpy.ndarray.
  -> predicted values   : min = 0.0, mean = 0.241, max = 1.0
  -> model type         : classification will be used (default)
  -> residual function  : difference between y and yhat (default)
  -> residuals          : min = -0.998, mean = -5.51e-18, max = 0.978
  -> model_info         : package sklearn

A new explainer has been created!
```

	recall	precision	f1	accuracy	auc
Decision Tree (Income)	0.54689	0.760479	0.636237	0.84947	0.873046

Figure 8: DALEX Output for Adult Income Model

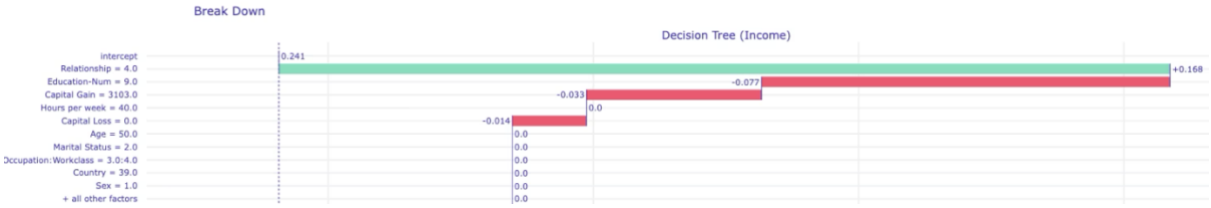

Figure 9: DALEX Local Breakdown Plot for Adult Income Model (Adult #10)

6.6 PDP and ICE for Feature-Response Analysis

Partial Dependence Plots and Individual Conditional Expectation curves help evaluate how predictions respond to changes in specific features. PDP shows an average feature effect, while ICE shows instance-level response curves. Together, they help distinguish broad model trends from heterogeneous individual behavior.

This is especially important in tabular classification problems because average effects may hide subgroup-specific variation. A PDP may suggest that a feature has a monotonic effect overall, while ICE curves may reveal that the effect differs across individuals.

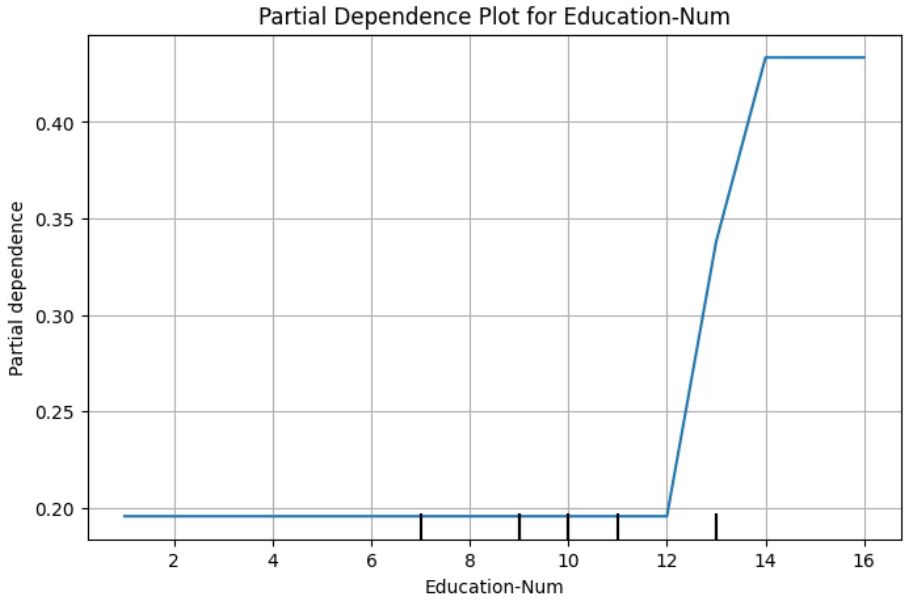

Figure 10: PDP Output for Education (Feature)

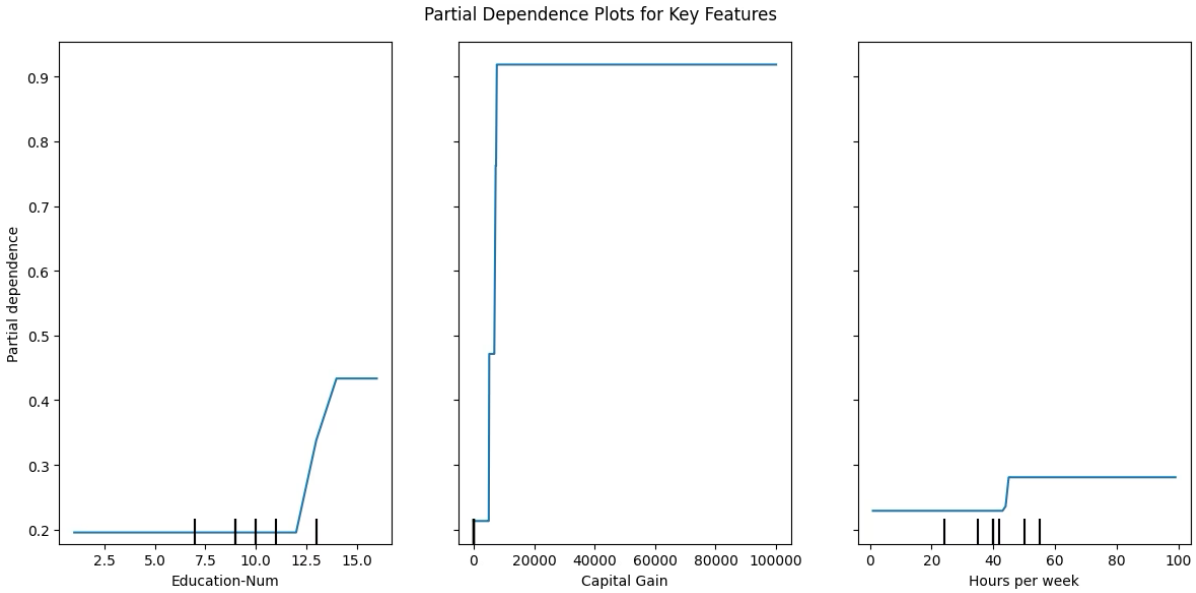

Figure 11: PDP output for multiple features

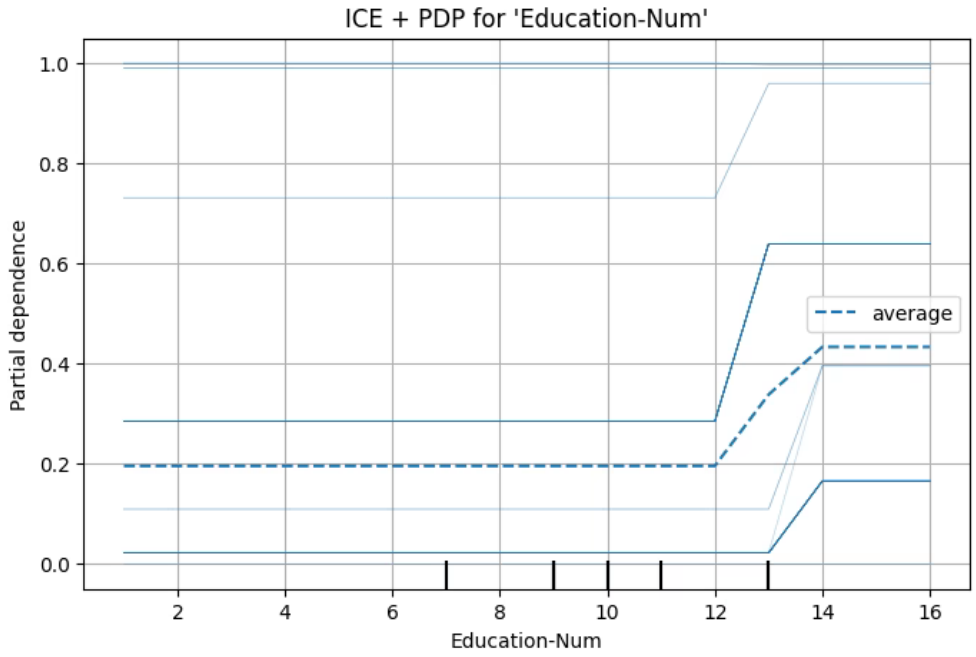

Figure 12:PDP and ICE Plot for a single feature 'Education'

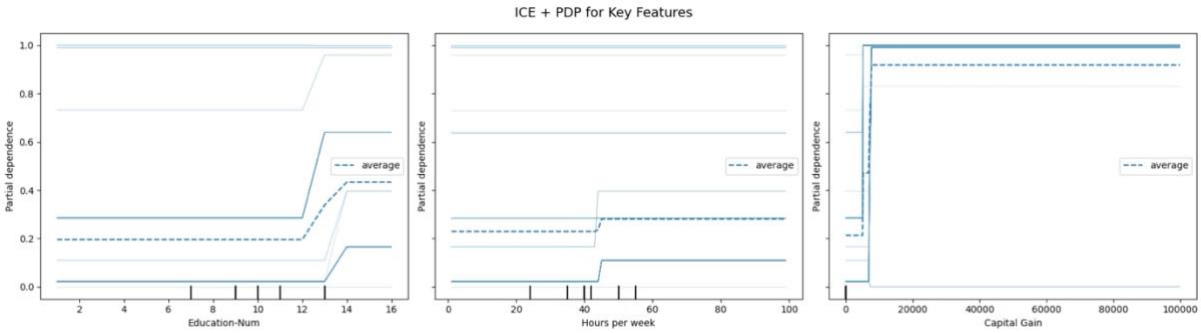

Figure 13: PDP and ICE for multiple features

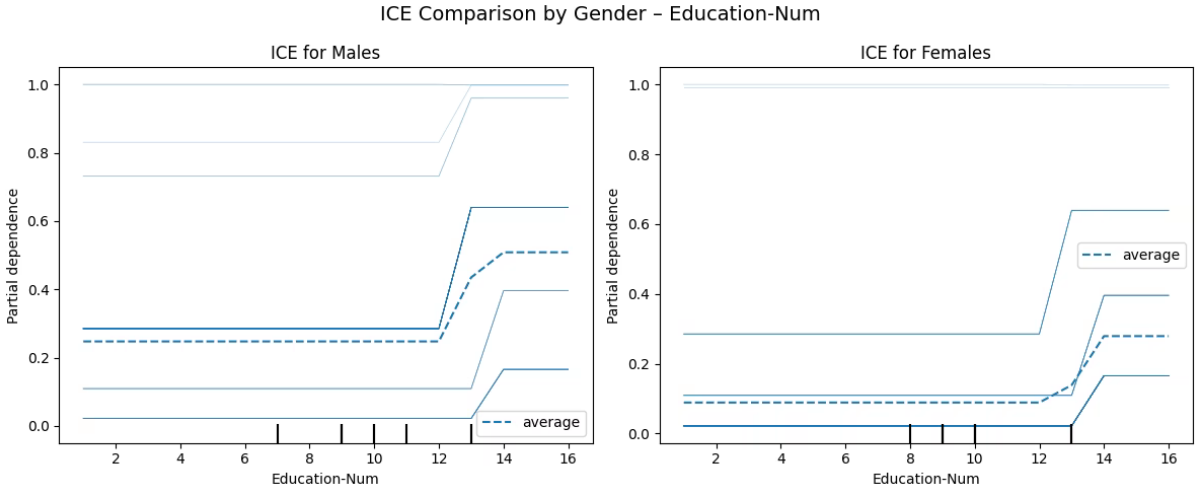

Figure 14: ICE for Fairness Comparison

7. Discussion

The implementation demonstrates that explainability is strongest when multiple methods are combined. Decision trees provide structural transparency, but they do not automatically provide complete interpretability. ELI5 offers readable summaries, but impurity-based feature importance can be misleading if interpreted as causal influence. SHAP provides additive attribution, but SHAP values must be interpreted in relation to the model and background data. LIME explains local neighborhoods, but its results may vary under perturbation. DALEX supports diagnostic interpretation, but results depend on implementation details. PDP and ICE reveal feature-response behavior, but they do not establish causality.

The main practical insight is that XAI should function as an interpretability stack rather than a single method. Each method answers a different question:

Question	Suitable Method
Which features matter overall?	ELI5, SHAP summary, DALEX
Why did this prediction occur?	SHAP force plot, LIME, ELI5 local explanation
What decision path was followed?	Decision-tree visualization
How does a feature affect predictions?	PDP
Do individuals respond differently to feature changes?	ICE
Is the explanation stable and trustworthy?	Cross-method comparison and governance review

This multi-method approach improves interpretability because explanations can be triangulated. Agreement across methods strengthens confidence; disagreement exposes uncertainty and motivates further analysis.

---

8. Research Contribution

This study contributes:

1. A practical XAI framework integrating SHAP, LIME, ELI5, DALEX, PDP, and ICE.
2. An implementation-grounded explanation workflow for tabular classification.
3. A distinction between global explanation, local explanation, and feature-response analysis.
4. A cautionary interpretation of model metrics, especially the difference between accuracy and recall.

5. A governance-oriented view of explainability as a trust-building capability.
6. A responsible AI interpretation model that avoids treating explanations as proof of correctness.

---

9. Limitations

This study has several limitations. The implementation uses a single tabular classification task and a decision-tree model. The findings therefore should not be generalized to all machine learning architectures without further testing. The dataset contains encoded categorical variables, which may reduce semantic clarity when interpreting feature effects. The reported model recall is moderate, indicating that the model may not identify all positive-class examples reliably.

The analysis does not perform fairness testing, subgroup evaluation, causal inference, or production monitoring. SHAP, LIME, ELI5, DALEX, PDP, and ICE are used for interpretability, not for proving that the model is fair, causal, or deployment-ready.

---

10. Governance and Responsible Use

Explainability must be governed carefully. Explanations can create a false sense of certainty if stakeholders mistake interpretability for correctness. A responsible XAI workflow should document feature definitions, preprocessing decisions, model assumptions, explanation methods, package versions, evaluation metrics, and known limitations.

For high-impact use cases, explainability should be combined with:

- bias and fairness testing,
- model drift monitoring,
- human review,
- audit trails,
- threshold analysis,
- subgroup performance evaluation,
- and domain-expert validation.

The use of sensitive or proxy-sensitive features such as race, sex, relationship status, marital status, and country should be assessed carefully before any real-world deployment.

---

11. Future Work

Future work should extend the framework across additional model classes, including logistic regression, random forests, gradient boosting, neural networks, and foundation-model-based classifiers. The framework should also be evaluated across multiple datasets and domains.

Further research should add fairness metrics, explanation stability tests, counterfactual explanations, causal analysis, and model-monitoring dashboards. A production-grade version of this workflow should also include reproducible environment files, version-controlled model artifacts, and automated explanation reporting.

---

12. Conclusion

This research presented a practical framework for Explainable Artificial Intelligence grounded in an applied implementation using SHAP, LIME, ELI5, DALEX, PDP, and ICE. The study demonstrated that explainability is not a single visualization or post-hoc technique but a multi-layered process involving model structure, feature attribution, local explanation, feature-response analysis, governance, and stakeholder communication.

The implementation showed how a decision-tree classifier can be interpreted through both intrinsic and post-hoc explanation methods.

The central conclusion is that trustworthy machine learning requires explanation triangulation. ELI5, SHAP, LIME, DALEX, PDP, and ICE each provide partial views of model behavior. Their combined use supports stronger interpretation, but explanations must remain subject to validation, governance, and human oversight.

References

Pruseth, D. (2026). *A Beginner-Friendly Guide to Explainable AI (XAI).* *Debabrata Pruseth AI Blog.*

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of KDD*.

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *arXiv*.

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*.

Suggested Citation

Pruseth, D. (2026). *Explainable Artificial Intelligence: A Practical Framework for Interpreting and Trusting Machine Learning Models*. Debabrata Pruseth AI Blog.