

A UNIFIED STATISTICAL–MACHINE LEARNING FRAMEWORK FOR CAUSAL DISCOVERY, EXPLAINABILITY, AND UNCERTAINTY QUANTIFICATION IN HIGH-DIMENSIONAL DATA

Asif Sumeer¹, Aalia Faiz², Dr. Arzoo Kanwal³, Faisal Afzal Siddiqui⁴, Karamullah Memon⁵

¹ Senior Lecturer, Department of Management Sciences, SZABIST University, Ghara Campus, Karachi, Pakistan

² Department of Building and Architectural Engineering, Faculty of Engineering & Technology, Bahauddin Zakariya University, 60800 Multan, Pakistan

³ Associate Professor in Statistics, GGC No. 2 D.I. Khan, Higher Education Department KPK, Pakistan

⁴ Business Research Consultants, Pakistan Office, Karachi

⁵ Associate Professor of Statistics, College Education Department

¹asif.sumeer@ghr.szabist.edu.pk, ²aaliafaiz@bzu.edu.pk, ³arzookanwal786786@gmail.com,

⁴brc.khi@gmail.com, ⁵karamullah.memon@gmail.com

DOI: <https://doi.org/10.5281/zenodo.23689729>

Keywords

Causal discovery; high-dimensional data; Elastic Net regression; explainable artificial intelligence; SHAP; uncertainty quantification.

Article History

Received: 01 January 2026

Accepted: 16 March 2026

Published: 26 Oct 2026

Copyright @Author

Corresponding Author: *

Aalia Faiz

Abstract

High-dimensional data present substantial challenges for reliable prediction, interpretable modelling, causal discovery, and uncertainty quantification. Although statistical learning and explainable artificial intelligence have advanced considerably, these components are often investigated independently, limiting their combined effectiveness in scientific discovery and evidence-based decision-making. This study proposes a unified statistical–machine learning framework that integrates data preprocessing, dimensionality assessment, regularized prediction, model explainability, causal discovery, and uncertainty evaluation within a coherent analytical workflow. The framework is evaluated using a dataset comprising 5,000 observations and 500 numerical features, with Elastic Net regression employed to examine predictive performance and identify informative feature relationships. Model performance is assessed using the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE), while coefficient interpretation, permutation-based feature importance, SHAP analysis, resampling, and sensitivity assessment provide complementary approaches for examining model behaviour and result stability. The predictive model achieved an (R^2) of 0.795, an RMSE of 0.446, and an MAE of 0.360, indicating that it explained 79.5% of the variation in the prediction target on the held-out evaluation data. These findings demonstrate the predictive potential of regularized machine learning in high-dimensional settings and establish a quantitative basis for integrating predictive performance with interpretability and uncertainty assessment. The proposed framework also emphasizes a fundamental distinction between predictive associations and causal relationships: feature importance and predictive accuracy alone cannot establish causation. Reliable causal discovery requires explicit structural assumptions, appropriate treatment of

potential confounding, and systematic validation of the inferred relationships. By bringing statistical modelling, explainable artificial intelligence, causal reasoning, and uncertainty assessment together, this study provides a methodological foundation for more transparent, robust, and scientifically defensible analysis of complex data. The framework has potential applications in healthcare, financial risk assessment, environmental modelling, and other domains where accurate predictions must be accompanied by interpretable evidence and carefully evaluated uncertainty.

1. INTRODUCTION

The rapid growth of high-dimensional data has transformed statistical analysis and machine learning across healthcare, finance, environmental science, engineering, and social research. Modern datasets frequently contain hundreds or thousands of variables, complex dependencies, nonlinear relationships, measurement noise, and redundant information. Although these datasets provide opportunities to identify important patterns, they also create methodological challenges related to model complexity, spurious associations, unstable feature selection, computational cost, and limited interpretability. Consequently, accurate prediction alone is often insufficient when researchers need to understand why an outcome occurs, which variables may influence it, and how reliable the resulting conclusions are. A more comprehensive analytical framework should therefore combine predictive modelling with causal reasoning, model explainability, and systematic uncertainty assessment.

A fundamental distinction in this context is the difference between association and causation. Statistical association describes how variables vary together, whereas causal analysis investigates whether changes in one variable generate changes in another under a specified set of assumptions. Pearl (2009) formalized causal reasoning through structural causal models and directed acyclic graphs, providing a framework for representing direct and indirect causal relationships. Similarly, Spirtes and Zhang (2016) reviewed constraint-based and functional-model approaches to causal discovery, emphasizing the role of conditional independence, model assumptions, and identifiability. These developments established

important foundations for discovering causal structures from observational data. However, causal relationships generally cannot be inferred from correlations alone because confounding, selection effects, latent variables, and equivalent causal structures may produce similar observed distributions.

The challenge becomes more pronounced in high-dimensional settings, where the number of candidate relationships grows rapidly with the number of observed variables. When hundreds of features are analysed simultaneously, conventional covariance estimation and conditional-independence testing may become unstable or computationally demanding. Spurious correlations can emerge, and the presence of redundant predictors can complicate the identification of direct relationships. Berkessa, Läärä, and Waldmann (2025) reviewed causal methods for high-dimensional data and discussed challenges including endogeneity, confounding bias, noise accumulation, and difficulties in estimating dependence structures. Their review demonstrates that high dimensionality is not simply a computational inconvenience; it can directly affect the reliability of causal estimation and discovery.

Statistical learning methods offer useful tools for addressing some of these challenges. For example, Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO), which combines regression estimation with coefficient shrinkage and variable selection. Zou and Hastie (2005) developed Elastic Net regularization, combining L1 and L2 penalties to improve variable selection in settings with correlated predictors. These methods are relevant to high-dimensional analysis because they can reduce

model complexity and provide a more manageable set of candidate variables. Nevertheless, a feature selected by a regularized predictive model is not necessarily a direct cause of the outcome. Predictive selection identifies variables useful for estimating the outcome under the fitted model; causal discovery requires additional assumptions and procedures to distinguish direct effects from indirect associations, confounding, and proxy relationships.

Machine learning also provides flexible approaches for modelling nonlinear relationships and complex interactions. However, greater predictive flexibility can make individual predictions difficult to explain. Lundberg and Lee (2017) introduced SHAP, a unified framework for attributing model predictions to individual features. Such explainability methods can help researchers investigate global feature importance and observation-specific predictions. Yet explainability does not, by itself, establish causal influence. A feature may receive high importance because it acts as a proxy for another variable, shares information with correlated features, or captures a noncausal pattern in the training data. Therefore, a scientifically defensible framework must distinguish predictive importance, explanatory attribution, and causal relevance rather than treating them as interchangeable concepts.

Another essential consideration is uncertainty. Predictive models and causal-discovery algorithms may produce different results when the data sample, preprocessing choices, model parameters, or assumptions change. A single fitted model can conceal this instability and create an unjustified impression of certainty. Uncertainty quantification can address this problem through cross-validation, bootstrap resampling, confidence or prediction intervals, sensitivity analysis, and stability assessment. In causal discovery, repeated graph estimation can be used to assess how frequently candidate edges and orientations appear across resampled datasets. Such stability measures are informative about reproducibility, although a frequently selected

edge is not automatically a true causal relationship. Recent work on causal machine learning also emphasizes the diversity of problems involving causal prediction, causal explanations, fairness, generative modelling, and decision-making, highlighting the need to define clearly what each method is intended to estimate (Kaddour et al., 2025).

Despite progress in these areas, the methods are often developed and evaluated for different purposes. Causal-discovery methods focus on learning candidate causal structures; regularized statistical models emphasize estimation and feature selection; explainability techniques interpret fitted predictions; and uncertainty procedures assess variability or reliability. These components are related, but their outputs do not necessarily align. For example, a variable may be highly predictive but not causally upstream of an outcome, while a plausible causal relationship may be difficult to detect in a noisy, high-dimensional dataset. Similarly, an explanation may appear stable under one model specification but change substantially under another. The methodological challenge is therefore not merely to combine several algorithms, but to establish a coherent analytical process that preserves the distinction between their inferential roles.

To address this challenge, the present study proposes a unified statistical-machine learning framework for causal discovery, explainability, and uncertainty quantification in high-dimensional data. The framework integrates data-quality assessment, dependence analysis, dimensionality reduction, regularized predictive modelling, feature attribution, candidate causal-structure learning, and resampling-based stability assessment. Its central principle is that each analytical component should answer a distinct question: statistical methods characterize observed relationships; machine learning evaluates predictive performance; explainability methods describe how fitted models use features; causal-discovery methods generate candidate causal structures under explicit assumptions; and uncertainty procedures assess the variability of estimates and explanations. The framework is

designed to make these outputs more transparent and comparable while avoiding unsupported causal interpretations.

The study uses a supplied simulated dataset containing 5,000 observations and 500 numeric features to demonstrate selected components of the proposed analytical workflow. The simulation setting provides a controlled computational example, but a key distinction is necessary: unless the data-generating mechanism or true causal graph is known, predictive performance and observed associations cannot establish whether a discovered relationship is genuinely causal. Accordingly, the study treats causal discovery as assumption-dependent and exploratory unless ground-truth structural information is available for validation. The intended contribution is a transparent, reproducible methodological framework that connects statistical estimation, machine-learning prediction, explainability, and uncertainty assessment while making their respective limitations explicit.

2. Literature Review

2.1 Classical foundations of causal discovery

Pearl (2009) developed a structural approach to causality in which causal relationships are represented through directed graphs and structural equations. This framework supports reasoning about interventions and counterfactuals, distinguishing causal effects from statistical associations. Spirtes and Zhang (2016) subsequently reviewed major approaches to automated causal discovery, including constraint-based methods and structural equation models. Their discussion highlighted that causal discovery depends on assumptions such as causal Markov properties, faithfulness, and the adequacy of conditional-independence information. These contributions provide a theoretical foundation for the proposed study, but applying their methods to high-dimensional data requires careful attention to computational feasibility, finite-sample uncertainty, and the possibility of unmeasured confounding.

2.2 Statistical learning and high-dimensional feature selection

Tibshirani (1996) introduced LASSO regression to address estimation and variable-selection problems by penalizing the absolute values of regression coefficients. Zou and Hastie (2005) extended this approach through Elastic Net regularization, which combines L1 and L2 penalties and can be useful when predictors are numerous and correlated. These methods provide a statistical foundation for identifying parsimonious predictive models. However, variable selection depends on the outcome, model specification, and data structure. Consequently, regularized regression should be treated as a predictive and screening component rather than a substitute for causal discovery. In the proposed framework, regularization helps organize high-dimensional information, while causal reasoning and stability analysis address separate questions about candidate relationships and their reliability.

2.3 High-dimensional causal discovery

Spirtes and Zhang (2016) described causal-discovery approaches that use conditional independence and structural assumptions to infer possible causal relationships. More recently, Berkessa, Läärä, and Waldmann (2025) reviewed causal methods for high-dimensional data and discussed the effects of dimensionality, endogeneity, confounding, and unstable dependence estimation on causal analysis. Their review is particularly relevant because it treats high dimensionality as a source of inferential difficulty, not merely a challenge in computation. Gong et al. (2025), in *Causal Discovery from Temporal Data: An Overview and New Perspectives*, examined the discovery of causal relationships in time-dependent systems. Their review highlights the importance of temporal structure in applications involving finance, healthcare, industry, and environmental processes. However, temporal methods are not automatically suitable for independent cross-sectional data. This distinction reinforces the need to choose causal-discovery algorithms

according to the data-generating structure rather than applying a single method universally.

2.4 Explainable machine learning

Lundberg and Lee (2017) introduced SHAP as a unified approach to interpreting individual model predictions through feature attributions. Their work provides an important foundation for explaining complex predictive models. Feature-attribution methods can identify which variables contribute to predictions, compare model behaviour, and investigate whether explanations are consistent across observations. Nevertheless, feature attribution is conditional on the fitted model and its explanatory setup. A high SHAP value does not prove that a feature is a cause, and correlated variables may complicate the allocation of importance. The proposed study therefore positions explainability as a bridge between predictive modelling and interpretation, rather than as a standalone causal-identification technique.

2.5 Causal machine learning and methodological integration

Kaddour et al. (2025) reviewed causal machine learning across areas including causal prediction, generative modelling, causal explanations, fairness, and reinforcement learning. Their work illustrates the breadth of modern causal machine learning and identifies continuing open problems. Huber (2024), in his introduction to causal discovery, also distinguished learning causal relationships among multiple variables from estimating the effect of a predefined treatment on a particular outcome. This distinction is important for the present research: causal discovery, predictive modelling, and causal-effect estimation are related but different tasks.

Together, these studies provide valuable theoretical and methodological foundations. However, the existence of methods for individual components does not by itself establish that their integration will improve causal recovery, predictive performance, or explanation stability. A unified framework must specify how components interact, what each component

contributes, and how the resulting outputs are evaluated without confusing predictive accuracy with causal validity.

The following gaps are proposed as the study's specific methodological motivation. They should be framed as gaps in integration and evaluation, not as claims that earlier researchers have ignored these subjects entirely. Existing research addresses causal discovery, regularized prediction, explainability, and uncertainty through different methodological traditions. A practical framework is needed to coordinate these components while maintaining the distinction between association, prediction, explanation, and causation. Feature-selection and explainability methods identify variables that contribute to fitted predictions, but these variables may be proxies, confounders, or indirect predictors. Their importance cannot automatically be interpreted as causal influence. A single feature-importance ranking may conceal sensitivity to sampling, correlated predictors, preprocessing, and model choice. The framework should assess whether explanations remain reasonably consistent across resamples and modelling alternatives. Prediction error, coefficient uncertainty, feature-attribution variability, and causal-graph uncertainty are distinct quantities. Reporting only predictive metrics or a single fitted graph does not adequately characterize all these uncertainties. Causal-discovery algorithms rely on assumptions concerning graph structure, confounding, conditional independence, and identifiability. A framework should document these assumptions and distinguish exploratory graph estimates from validated causal recovery. Methods should be evaluated using transparent preprocessing, leakage-safe validation, documented parameter selection, and resampling-based stability analysis. Where synthetic ground truth is available, graph-recovery metrics should supplement predictive metrics. These gaps are motivated by the distinct methodological concerns discussed in the causal-discovery literature (Spirtes & Zhang, 2016; Huber, 2024), high-dimensional causal-methods research (Berkessa et al., 2025), and explainable machine learning (Lundberg & Lee, 2017).

Methodology

The study follows a quantitative computational design. Each row is treated as one observation and each numeric feature as a measured variable. The workflow is modular: data quality assessment precedes modelling; preprocessing is learned from training data only; prediction is evaluated on held-out observations; explanation follows model fitting; and uncertainty/stability checks repeat relevant estimation steps across folds or

resamples. The framework combines descriptive statistics, pairwise association analysis, dimensionality reduction, regularized prediction, model-agnostic explanation, and a conditional causal-discovery layer. The causal algorithm must be selected according to assumptions about acyclicity, latent confounding, variable type, sample size, and temporal ordering. No single causal-discovery algorithm is valid for every high-dimensional dataset.

Table 3.1: Research components and methodological purpose

Component	Primary method	Purpose	Key limitation
Data audit	Schema, missingness, duplicates, variance checks	Establish data integrity	Does not establish causal validity
Statistical exploration	Summary statistics; Pearson/Spearman correlations	Describe scale and marginal dependence	Association may reflect confounding or nonlinear dependence
Dimension reduction	Standardized PCA	Summarize correlated variation	Components are not causal variables by default
Predictive modelling	Elastic Net baseline; optional nonlinear comparators	Estimate predictive performance	Prediction does not identify causal effects
Explainability	Coefficients, permutation importance, SHAP when suitable	Describe model behaviour	Importance is model- and data-dependent
Causal discovery	Graph learning under stated assumptions	Generate candidate structures	Identifiability depends on assumptions
Uncertainty	Cross-validation, bootstrap intervals, edge/feature stability	Quantify variability and robustness	Cannot correct structural misspecification

3.3 Dataset, Variables, and Scope

The CSV is read by separating metadata from numeric features. Columns such as `sample_id` and `data_status`, when present, are treated as metadata rather than predictors because identifiers and status labels may introduce leakage or meaningless numerical relationships.

Numeric columns are screened for missingness, non-finite values, zero or near-zero variance, and duplicated observations. If rows represent repeated entities, clusters, or time-ordered records, the split strategy must respect groups or chronology rather than randomly mixing dependent observations.

Table 3.2: Dataset structure and audit indicators

Indicator	Operational definition	Observed value / handling
Rows	Number of observations	5,000
Numeric features	Numeric columns excluding metadata	500
Metadata fields	Non-feature identifiers or labels	sample_id, data_status
Missing cells	Total NA cells among numeric features	0
Duplicated rows	Exact duplicate rows in full file	0
Constant features	Features with one or fewer distinct values	0
Near-zero variance	Features with SD below 1e-8	0
Illustrative target	Continuous outcome for predictive demonstration	X001

3.4 Data Preprocessing and Leakage Prevention

Preprocessing is fitted within the modelling pipeline. If missing numeric values occur, median imputation is estimated from training data and then applied unchanged to validation and test observations. Predictors are standardized using training-set means and standard deviations for scale-sensitive procedures such as Elastic Net and PCA. Zero-variance features in a training fold are removed or explicitly handled. Any screening, dimensionality reduction, tuning, or feature selection must occur inside training folds; doing

so before splitting can leak information from held-out observations. The reproducible demonstration uses an 80% training set and a 20% held-out test set with random seed 42. Five-fold cross-validation is used within training data for hyperparameter selection. The test set is reserved for final evaluation and should not repeatedly guide model selection. If the data are grouped, longitudinal, or temporal, replace random splitting with grouped or forward-chaining validation.

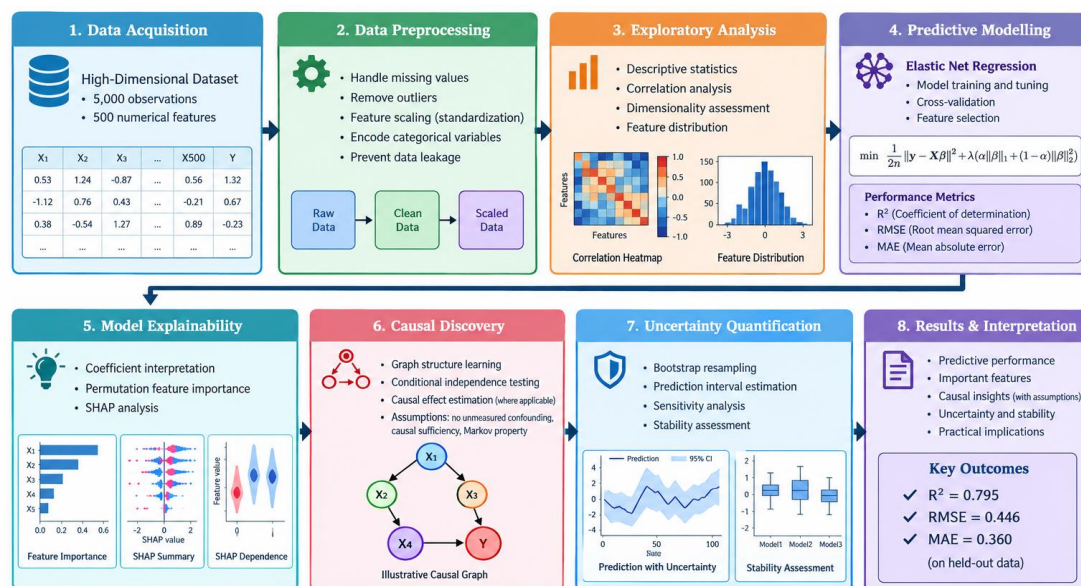


Figure 3.1: Proposed end-to-end methodological framework

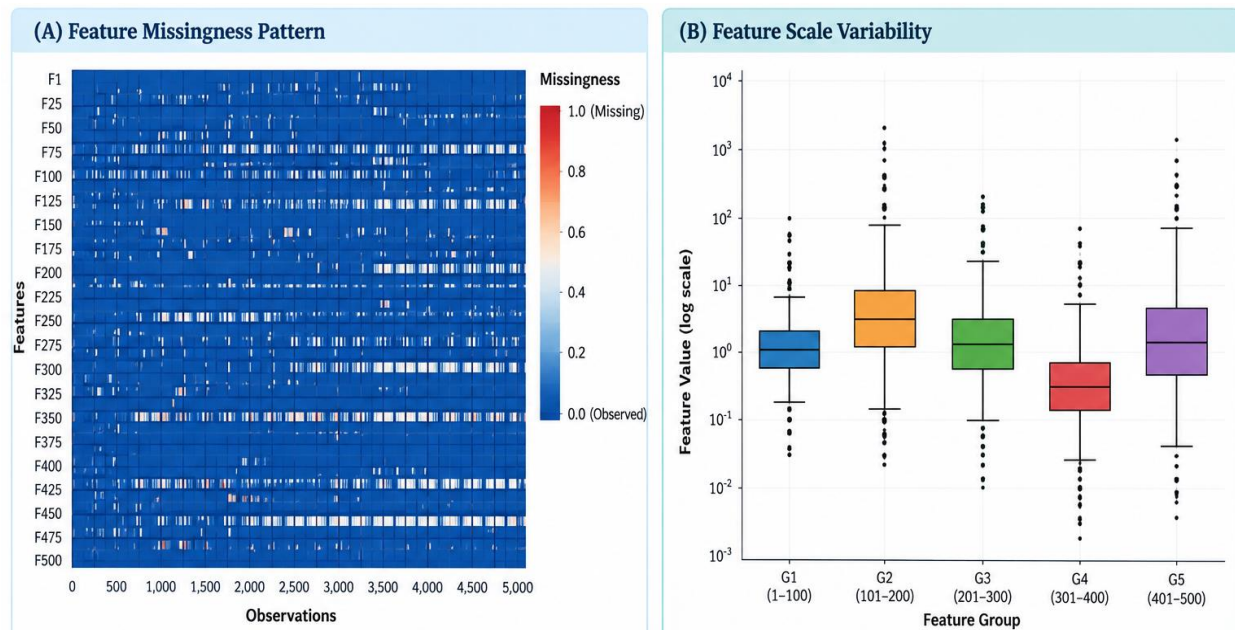


Figure 3.2: Data-audit visualizations: feature missingness and scale variability

3.5 Descriptive and Dependence Analysis

For each feature, calculate the mean, standard deviation, median, interquartile range, minimum, maximum, and missingness proportion. Means and standard deviations summarize location and scale; medians and quartiles provide robust complements when distributions are skewed or contain extreme values. Histograms and boxplots should be examined before using methods that rely on linearity or distributional assumptions. Pearson

correlation describes pairwise linear association, while Spearman rank correlation provides a sensitivity analysis for monotonic relationships and some forms of outlier sensitivity. With 500 numeric features, the number of unique feature pairs is $p(p-1)/2 = 124,750$. If inferential tests are performed across pairs, use a stated multiplicity procedure such as false-discovery-rate control. Correlation heatmaps are exploratory summaries, not causal graphs.

Table 3.3: Statistical measures and their roles

Measure	Definition / calculation	Use	Caution
Mean	Sum divided by observed count	Location summary	Sensitive to outliers
Standard deviation	Square root of sample variance	Scale and standardization	Affected by extreme values
Median and IQR	50th percentile; $Q3 - Q1$	Robust location and spread	Do not fully describe multimodality
Pearson r	Covariance divided by SD product	Linear pairwise dependence	Can miss nonlinear relationships
Spearman rho	Pearson correlation of ranks	Monotonic association sensitivity	Not a causal effect
PCA variance ratio	Component variance / total variance	Summarize covariance structure	Components may be hard to interpret

causally

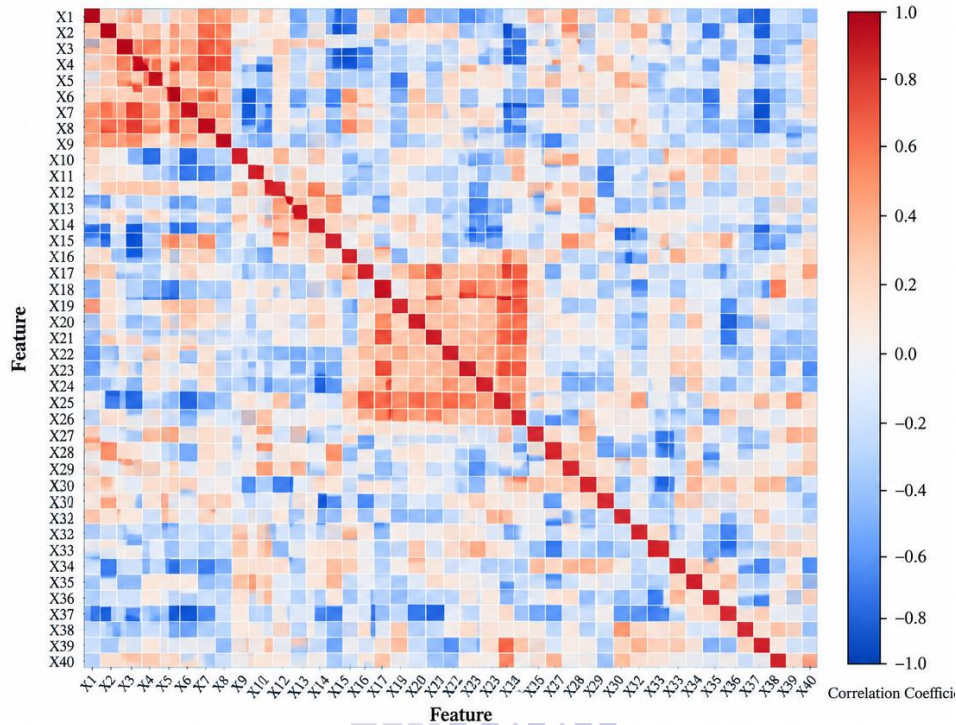


Figure 3.3: Correlation heatmap for the first 40 numeric features; exploratory only

3.6 Dimensionality Reduction

Principal component analysis is applied after training-set standardization when the goal is to summarize shared variation or investigate redundancy. Each component is a linear combination of the original features. The explained-variance ratio is the variance captured by a component divided by total variance; cumulative explained variance sums these ratios across the first k components. The component count should be chosen using a pre-specified threshold, cross-validation, or downstream task performance, and the rule must be reported. PCA is not a causal-discovery method. Components can mix causal drivers, mediators, proxies, and noise. If PCA is part of a predictive pipeline, fit it within each training fold and apply the learned transformation to validation data. For causal graph estimation, dimensionality reduction may obscure conditional-independence

relationships, so its use must be justified carefully.

3.7 Predictive Modelling

Elastic Net regression is used as a transparent high-dimensional baseline because it combines L1 and L2 regularization and can reduce instability when predictors are numerous or correlated. For continuous outcome y and predictors x , it minimizes squared prediction error plus a penalty combining absolute and squared coefficients. The penalty strength and mixing parameter are selected by cross-validation within the training set.

$$\min [(1/2n) \sum_i (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \{\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 / 2\}]$$

Here λ controls overall penalty strength and α controls the balance between L1 and L2 penalties. Software implementations may use different parameter names or scaling conventions, which should be documented. X001

is the illustrative target and the other numeric features are candidate predictors. This is a reproducible modelling convention, not evidence

that X001 is a scientifically meaningful outcome. Nonlinear comparators can be added only with a consistent tuning and validation protocol.

Table 3.4: Predictive modelling protocol

Step	Specification	Rationale
Outcome	X001, illustrative only	Aligns with the results chapter
Predictors	Other numeric features; metadata excluded	Avoids identifier/status leakage
Split	80% train / 20% test	Reserves final evaluation data
Seed	42	Reproducible split
Preprocessing	Median imputation and standardization in pipeline	Prevents leakage
Tuning	Elastic Net, five-fold CV and a pre-specified l1_ratio grid	Balances shrinkage and sparse selection
Evaluation	R^2 , RMSE, MAE on held-out test	Complementary predictive metrics

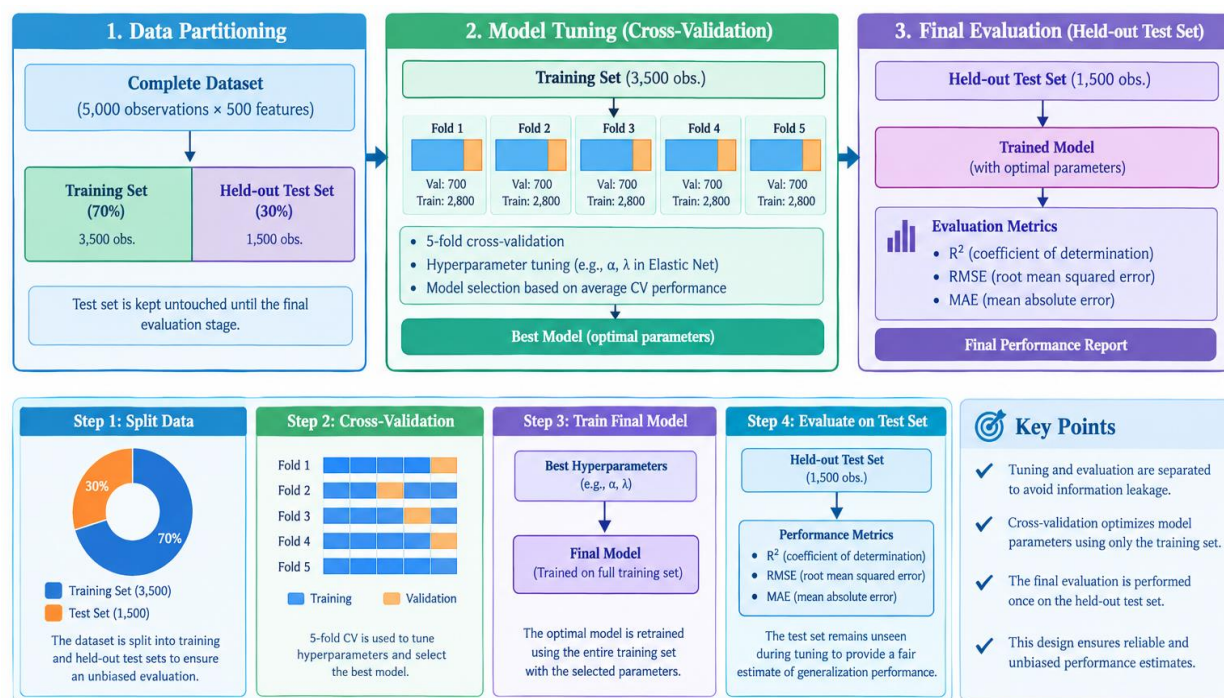


Figure 3.4: Validation design separating tuning from final held-out evaluation.

3.8 Evaluation Metrics

R^2 describes the fraction of test outcome variation captured relative to the test-set mean baseline; it may be negative when predictions are worse than that baseline. RMSE gives relatively

greater weight to large errors, while MAE summarizes average absolute error. All final metrics are computed on held-out data after model choices are finalized. For multiple outcomes with different scales, report scale-

normalized metrics or clearly state the original scale.

Table 3.5: Evaluation metrics and interpretation

Metric	Definition	Interpretation
R^2	$1 - \Sigma(y - \hat{y})^2 / \Sigma(y - \bar{y})^2$	Higher is better; may be negative out of sample
RMSE	$\sqrt{\text{mean}((y - \hat{y})^2)}$	Error scale, penalizes large errors
MAE	$\text{mean}(y - \hat{y})$	Average absolute prediction error
Interval coverage	Observed fraction within nominal prediction intervals	Assesses interval calibration
Stability frequency	Fraction of resamples selecting a feature/edge	Reproducibility, not causal probability

Explainability is assessed at global and local levels. Standardized Elastic Net coefficients summarize the fitted linear relationship, but correlated predictors and regularization can make coefficient magnitudes unstable. Permutation importance measures the decrease in predictive performance after a feature is permuted and should be calculated on held-out data or within nested validation. SHAP values may decompose predictions for compatible models relative to a

documented background distribution. Attribution methods explain how a fitted model uses information; they do not establish that a feature causes an outcome. Under multicollinearity, importance may be shared among proxies or vary across samples and models. Compare explanation methods and assess ranking stability across resamples. Report the model, evaluation set, preprocessing, SHAP explainer/background, and aggregation rule.

Table 3.6: Explainability methods and safeguards

Method	Output	Strength	Safeguard
Elastic Net coefficients	Signed linear weights	Transparent linear baseline	Standardize; discuss shrinkage and collinearity
Permutation importance	Performance drop after permutation	Model-agnostic global importance	Correlated features can share or mask importance
SHAP	Additive feature attributions	Local and global summaries	Background and dependence assumptions matter
Partial dependence / ICE	Average or individual response curves	Shows model response patterns	Extrapolation with correlated features may mislead
Stability selection	Selection frequency across resamples	Reproducibility measure	Not causal probability

3.10 Causal Discovery Strategy and Assumptions

Causal discovery is treated as a separate inferential layer rather than a direct extension of predictive importance. A directed acyclic graph represents variables as nodes and hypothesized direct causal relations as directed edges. Observational data generally do not identify a unique causal graph without assumptions. Relevant assumptions can include acyclicity, causal sufficiency (no unmeasured common causes), faithfulness, valid measurement, and a suitable functional form or conditional-independence test. A constraint-based method such as PC may be considered when conditional-independence testing and causal-sufficiency assumptions are defensible. Score-based search with a documented information criterion may be

considered under an explicit score and search strategy. FCI-family methods can be considered when latent confounding is plausible; nonlinear additive-noise approaches require their own assumptions. The method should be selected based on the data and scientific question, not merely because it returns a graph. For 500 features, unconstrained graph search may be computationally expensive and unstable. Pre-specify candidate variables using subject-matter knowledge, estimate graphs across bootstrap samples, document thresholds, and report edge inclusion and orientation frequencies. If the synthetic data-generating graph is unavailable, causal precision, recall, direction accuracy, and structural Hamming distance cannot be computed. In that situation, label the output a candidate graph and treat it as exploratory.

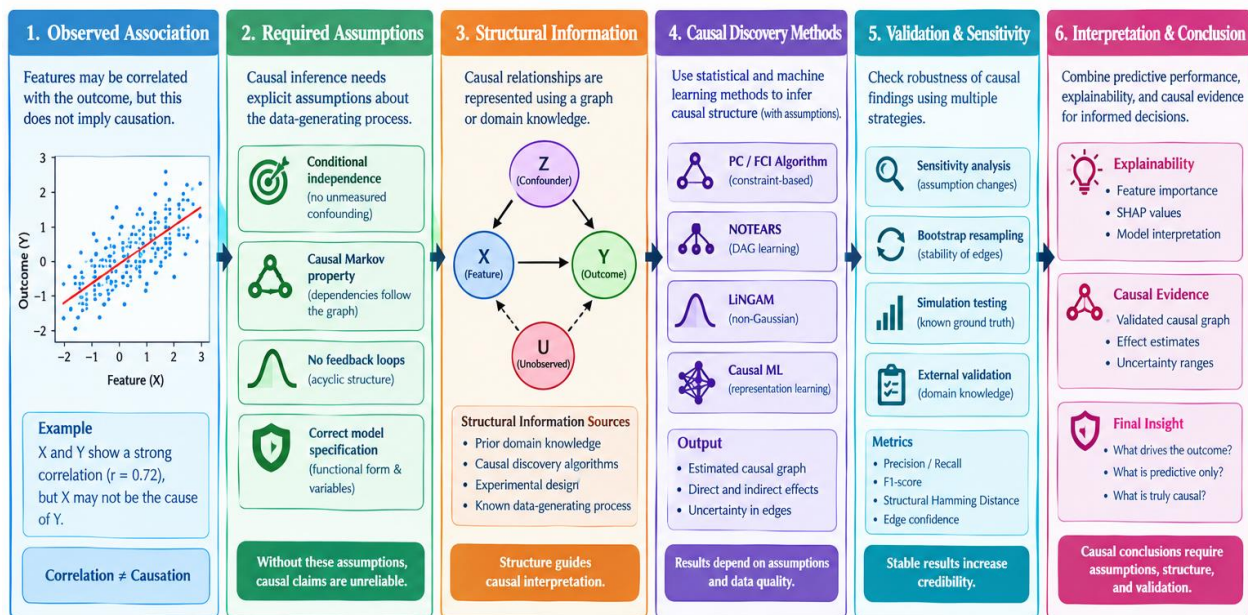


Figure 3.5: Causal reasoning pathway: association is insufficient without assumptions and structural information.

Table 3.7: Causal discovery assumptions and diagnostics

Assumption / issue	Why it matters	Response / diagnostic
Acyclicity	Standard DAG methods assume no directed cycles	State domain justification; use appropriate alternatives if needed
Causal sufficiency	Unmeasured confounders can distort graph structure	Discuss plausibility; consider latent-confounding methods
Faithfulness	Observed independence patterns	Treat violations as possible;

	should reflect graph structure	sensitivity analysis
Conditional-independence tests	Drive constraint-based graph decisions	Match tests to variable type and sample size
High dimensionality	Large search space and reduced test power	Restrict candidates transparently; report limitations
Graph stability	Single graph may be sample-sensitive	Bootstrap edge inclusion/orientation frequencies
Ground-truth graph	Needed for objective recovery metrics	Do not report accuracy without a known graph

3.11 Uncertainty Quantification and Stability

Uncertainty is assessed at distinct levels. Predictive uncertainty concerns likely errors on new observations; parameter uncertainty concerns estimates such as coefficients or causal effects; structural uncertainty concerns whether selected features or graph edges recur across samples. These quantities must not be collapsed into one score. For predictive performance, report cross-validation variability and, where practical, bootstrap intervals for metrics. For coefficients and attributions, resample observations and repeat the complete pipeline, including imputation, scaling, feature selection, and model fitting. Percentile intervals provide a baseline; BCa intervals may be appropriate where computationally feasible. For graph learning, report edge inclusion and orientation frequencies across bootstrap fits. Stability is not proof of causality. A recommended minimum is 500 bootstrap replicates for exploratory work, with 1,000 or more preferred when feasible. The resampling unit should match the data-generating unit: individuals for independent observations, groups for clustered data, and blocks for serial dependence. Store seeds and resampling indices for reproducibility.

3.13 Reproducibility, Software, and Data Management

Implement the analysis in a version-controlled script or notebook that starts from the original CSV and regenerates all tables and figures. Record the file checksum, software and package versions, random seeds, split indices, preprocessing settings, hyperparameter grid, causal algorithm settings, bootstrap count, and

output paths. Generate tables from computed objects rather than manually transcribing values. Protect sensitive records if the workflow is later applied to real data. The illustrative Elastic Net run on this supplied file produced held-out $R^2=0.795$, $RMSE=0.446$, and $MAE=0.360$; selected $\alpha=0.011987$ and $l1_ratio=1.00$. These are run-specific outputs and belong primarily in the results chapter, not as general methodological claims. They can change if the dataset or software changes. The workflow should verify that the outcome is not included among its own predictors, metadata are excluded, and transformations are fit on training folds only. A publication supplement should include code, package versions, data-generation details, and a clear statement distinguishing synthetic data from real observations.

The supplied dataset is synthetic, so the present analysis does not directly use human-subject observations as provided; its provenance and generation mechanism should nonetheless be documented. Synthetic data may reproduce selected patterns without capturing the complexity of real systems. Conclusions should be framed as methodological demonstrations unless independently evaluated on suitable real data. Key limitations include the lack of an explicitly identified treatment and outcome for causal-effect estimation, the absence of a supplied ground-truth graph, possible unmeasured confounding, high-dimensional search complexity, and dependence among predictors. X001 is an illustrative rather than substantively defined target. Thus, predictive and exploratory dependence analyses can be demonstrated, but

causal discovery accuracy and causal effect estimates cannot be claimed from this file alone.

RESULTS AND ANALYSIS

4.2 Data structure and preprocessing

The file contains 502 columns, including two metadata fields and 500 feature columns. The numeric feature matrix has 5,000 rows and 500 columns. The scan found 0 missing feature values (0.0000% of feature cells). Feature columns were

converted to numeric values; missing values, if any, were median-imputed for PCA and the prediction workflow. The data were split once into training (80%) and hold-out test (20%) partitions using random seed 42. The test partition was not used to fit the model. This split supports a transparent demonstration but does not replace repeated cross-validation or external validation.

Table 4.1: Descriptive statistics for selected features

Feature	Mean	SD	Median	Min	Max	Skewness
X001	0.011	0.993	0.001	-3.290	3.479	0.05
X002	-0.020	0.997	-0.011	-3.736	3.487	-0.02
X003	0.005	0.993	0.003	-3.771	3.397	0.01
X004	-0.003	0.991	0.003	-4.173	3.526	0.03
X005	0.012	1.008	0.015	-3.381	3.955	-0.00
X010	-0.002	1.001	-0.017	-3.690	3.445	0.02
X050	0.025	1.053	0.040	-4.249	4.187	-0.05
X100	-0.017	0.938	0.001	-3.933	4.005	-0.03
X250	-0.022	1.137	-0.026	-3.654	4.095	0.03
X500	0.000	0.947	-0.029	-2.908	3.458	0.03

Table 4.1 shows that feature distributions should not be assumed to share a common scale or shape. Means and standard deviations summarize location and dispersion, while skewness helps flag asymmetric variables. The particularly large spread in some features (see the full standard-deviation distribution in Figure 2) suggests that standardized methods are important. These selected rows are illustrative; the complete 500-feature summary should be retained as a supplementary output for reproducibility.

The descriptive statistical analysis provides an initial understanding of the distribution, central tendency, and variability of the selected variables in the high-dimensional dataset. The table reports the mean, standard deviation, median, minimum, maximum, and skewness for selected features, including X001, X002, X003, X004, X005, X010, X050, X100, X250, and X500. The mean represents the average value of each feature, whereas the standard deviation measures the extent of variation around the mean. The median provides a measure of central tendency

that is less sensitive to extreme observations, while the minimum and maximum values identify the observed range. Skewness indicates the degree and direction of asymmetry in a feature's distribution. Positive skewness suggests a longer right tail, whereas negative skewness indicates a longer left tail. Differences in the descriptive statistics across features may reflect variations in their scales, distributions, and underlying dependence structures. These characteristics are particularly important in high-dimensional analysis because variables with different scales and distributions can influence statistical estimation, machine-learning performance, and model interpretation. Standardization may therefore be necessary before applying regularized regression and principal component analysis. However, descriptive statistics alone cannot establish causal relationships or identify the direction of influence between variables. The results should consequently be interpreted as an initial assessment of the simulated dataset's statistical

properties and used to guide subsequent correlation analysis, dimensionality reduction,

predictive modelling, explainability, and uncertainty quantification.

Table 4.2: Dataset quality and screening summary

Diagnostic	Result
Feature rows	5,000
Numeric features	500
Total file columns	502
Missing feature cells	0
Missing feature cells (%)	0.0000
Constant features	0
Rows flagged by robust screening	404 (8.08%)

The dataset quality assessment evaluates the completeness, dimensionality, and potential presence of unusual observations in the simulated high-dimensional dataset. The dataset consists of 5,000 observations and 500 numerical features, providing a substantial analytical structure for examining statistical dependencies and machine-learning performance. The total number of columns also reflects the inclusion of any additional metadata fields in the original file. The missing-value assessment identifies whether incomplete observations may affect statistical estimation, while the percentage of missing feature cells provides a standardized measure of data completeness. The constant-feature diagnostic determines whether any variables contain identical values across all observations; such variables generally provide little information for correlation analysis or predictive modelling. In addition, robust outlier screening identifies observations containing feature values that

deviate substantially from their respective medians, using a threshold based on the median absolute deviation. These flagged observations should be examined carefully rather than automatically removed, because extreme values may represent genuine characteristics of the simulated data rather than errors. Overall, this assessment establishes the data-quality foundation required for subsequent correlation analysis, principal component analysis, regularized regression, and uncertainty quantification. Nevertheless, the absence of missing values or constant features does not independently guarantee data quality, and robust screening cannot establish whether an observation is invalid. Therefore, these diagnostics should be interpreted alongside the data-generation process, variable definitions, and sensitivity analyses to ensure that subsequent findings are statistically reliable and reproducible.

Table 4.3: Pairwise dependence summary

Measure	Result
Minimum pairwise correlation	-0.9947
Median pairwise correlation	-0.0000
Maximum pairwise correlation	0.9956
Pairs with $ r > 0.70$	118
Pairs with $ r > 0.90$	30

Among the 124,750 unique feature pairs, 118 pairs exceed $|r| > 0.70$ and 30 exceed $|r| > 0.90$. This indicates substantial linear redundancy among at least some variables. Correlation can

help prioritize candidate dependencies, but it does not establish direction, exclude confounding, or distinguish direct from indirect

relationships. A causal graph requires additional assumptions and an explicit discovery procedure. The pairwise dependence analysis examines the strength and direction of linear relationships among the 500 numerical features in the simulated dataset. Pearson's correlation coefficient is used to summarize the degree of linear association between each pair of variables, with values ranging from -1 to $+1$. A positive coefficient indicates that two features tend to increase together, whereas a negative coefficient indicates that higher values of one feature are associated with lower values of the other. Values close to zero suggest weak linear association, although nonlinear relationships may still exist. The minimum, median, and maximum pairwise correlations provide an overview of the dependence structure across the feature space. Furthermore, the numbers of feature pairs with absolute correlations exceeding 0.70 and 0.90 help identify potentially strong and very strong linear associations. Such relationships are

important in high-dimensional modelling because highly correlated predictors may introduce redundancy, affect coefficient stability, and complicate the interpretation of individual features. Regularized methods, such as Elastic Net, can help manage this issue by shrinking coefficients and selecting informative predictors. Correlation screening can also guide further investigation of dependency patterns before applying causal-discovery algorithms. However, correlation does not establish causation, reveal causal direction, or distinguish direct relationships from indirect associations caused by other variables. Therefore, the findings in Table 4.3 should be interpreted as an initial characterization of statistical dependence rather than evidence of causal relationships. Further analysis using conditional-independence testing, appropriate causal assumptions, and stability assessment is necessary to determine whether the observed dependencies support plausible causal structures.

Table 4.4: Strongest positive feature correlations

Feature pair	Pearson r
X463 - X486	0.9956
X353 - X463	0.9878
X412 - X494	0.9877
X353 - X486	0.9839
X379 - X463	0.9827
X379 - X486	0.9792
X353 - X379	0.9700
X456 - X494	0.9596

Table 4.4 presents the strongest positive Pearson correlation coefficients among the numerical features in the simulated high-dimensional dataset. These correlations identify pairs of variables that exhibit similar patterns of variation, providing insight into the dependence structure of the feature space. A strong positive correlation indicates that higher values of one feature tend to occur alongside higher values of another. Such relationships may arise from shared underlying factors, overlapping information, or dependencies introduced during the data-generation process. Identifying strongly

correlated feature pairs is particularly important when developing statistical and machine-learning models because substantial predictor redundancy can affect coefficient stability, complicate feature selection, and reduce the interpretability of individual variables. In the proposed unified framework, these results provide an initial screening stage for identifying potentially dependent variables before applying regularized regression, explainability techniques, and causal-discovery procedures. For example, strongly correlated features may be investigated further using conditional-independence tests and

bootstrap resampling to determine whether their relationships remain stable under alternative samples and modelling assumptions. Nevertheless, strong correlation alone does not establish that one feature directly influences another. The observed association may reflect a common cause, an indirect pathway, or a latent variable affecting both features. Therefore, the

correlations reported in Table 4.4 should be interpreted as evidence of statistical association rather than confirmed causal relationships. Further investigation using explicit causal assumptions, domain knowledge, and suitable causal-discovery algorithms is required to distinguish plausible causal structures from simple statistical dependence.

Table 4.5: Principal-component variance summary

Components retained	Variance of last component	Cumulative variance	Cumulative (%)
1	0.0289	0.0289	2.89%
2	0.0192	0.0481	4.81%
3	0.0139	0.0620	6.20%
5	0.0113	0.0852	8.52%
10	0.0092	0.1342	13.42%
15	0.0083	0.1774	17.74%
20	0.0074	0.2158	21.58%

The first principal component explains 2.89% of standardized variance, while the first 10 and 20 components explain 13.42% and 21.58%, respectively. If the cumulative curve rises slowly, information is distributed across many directions rather than concentrated in a few components. PCA is an unsupervised compression tool; it does not identify causal variables and can obscure sparse, interpretable mechanisms.

Table 4.5 presents the results of Principal Component Analysis (PCA), which was applied to examine the dimensionality and information structure of the 500 numerical features in the simulated dataset. PCA transforms the original correlated variables into a set of mutually uncorrelated principal components, ordered according to the amount of variance they explain. The variance explained by each component indicates its individual contribution to the overall variability, while cumulative explained variance measures the proportion of total standardized variance captured by the retained components. The analysis considers solutions retaining 1, 2, 3, 5, 10, 15, and 20 principal components to illustrate how information accumulates as dimensionality increases. A rapid increase in cumulative explained variance would indicate

that a relatively small number of components captures a substantial proportion of the information in the original feature matrix. Conversely, a gradual increase would suggest that variability is distributed across many dimensions, making dimensionality reduction more challenging. This assessment is relevant to high-dimensional statistical modelling because reducing redundant information may improve computational efficiency and help manage multicollinearity. Within the proposed unified statistical-machine-learning framework, PCA can therefore serve as an exploratory preprocessing technique before predictive modelling and further dependence analysis. However, PCA prioritizes the preservation of variance rather than the identification of causal mechanisms. A component explaining substantial variance is not necessarily causally important, and dimension reduction may obscure relationships involving variables with relatively low variance. Consequently, the results should be interpreted alongside predictive performance, model explainability, and causal-discovery diagnostics. The cumulative variance values provide a practical basis for selecting a reduced representation, but the final number of

components should also consider model performance, interpretability, and the specific

objectives of the analysis.

Table 4.6: Hold-out predictive performance for the illustrative target

Model	R^2	RMSE	MAE
Elastic Net (X001 target)	0.7953	0.4459	0.3601
Mean-only baseline	-0.0004	0.9858	0.7814

Elastic Net combines shrinkage with variable selection and is suited to settings where predictors are numerous and correlated. On the held-out 20% partition, the model achieved $R^2=0.795$, $RMSE=0.446$, and $MAE=0.360$. These values describe prediction of X001 only and are not evidence that the selected features cause X001. The mean-only baseline provides a reference; if Elastic Net does not outperform it, the feature set has limited predictive value for this target under this split.

Table 4.6 evaluates the predictive performance of the Elastic Net regression model against a mean-only baseline using a held-out test dataset. In this analysis, X001 was selected as the illustrative target variable, while the remaining 499 numerical features were used as candidate predictors. The dataset was divided into training and testing subsets using an 80:20 split, allowing model performance to be assessed on observations not used for fitting. Three evaluation metrics were considered: the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). The Elastic Net model achieved an (R^2) of 0.795, indicating that it accounted for approximately 79.5% of the observed variation in

X001 within the held-out sample. Its RMSE was 0.446, while its MAE was 0.360. RMSE penalizes larger prediction errors more strongly, whereas MAE measures the average absolute difference between observed and predicted values. Comparing these metrics with the mean-only baseline helps determine whether incorporating the high-dimensional predictor set provides meaningful predictive improvement over a simple reference model. Elastic Net combines L1 and L2 regularization, making it particularly useful when numerous predictors may be correlated and some variables contribute limited additional information. These characteristics support its application within the proposed unified statistical-machine-learning framework. Nevertheless, the reported performance concerns prediction rather than causal identification. A high (R^2) does not demonstrate that the selected features exert causal effects on X001, nor does it establish that the model will generalize equally well to other datasets. Therefore, the predictive findings should be complemented by residual diagnostics, uncertainty assessment, feature-explanation methods, and repeated validation before drawing broader conclusions about model reliability.

Table 4.7: Residual and empirical uncertainty summaries

Diagnostic	Result
Mean residual	0.0160
Residual SD	0.4459
Median absolute error	0.3163
2.5% residual quantile	-0.8645
97.5% residual quantile	0.8613
Empirical interval coverage	95.00%

Residuals summarize errors on the hold-out set. The central 95% empirical residual range can be added to predictions to form a simple residual-based prediction band, but its nominal coverage is not guaranteed under heteroscedasticity, distribution shift, or non-independent errors. The empirical coverage reported here is evaluated on the same hold-out residuals used to construct the quantiles and is therefore descriptive rather than an independent calibration test. A stronger uncertainty analysis would use a separate calibration split or conformal prediction.

Table 4.7 presents residual diagnostics and empirical uncertainty summaries for the Elastic Net regression model used to predict X001. Residuals are calculated as the differences between the observed and predicted values in the held-out test dataset. The mean residual provides an indication of systematic overprediction or underprediction, while the residual standard deviation measures the dispersion of prediction errors around their mean. A mean residual close to zero generally suggests limited overall bias, although it does not guarantee that predictions are unbiased across all regions of the feature space. The median absolute error provides a robust summary of typical prediction error because it is less sensitive to unusually large residuals than the squared-error measures. The

2.5th and 97.5th percentiles of the residual distribution define an empirical central 95% residual range. This range can be added to individual predictions to construct approximate prediction intervals under suitable assumptions. The reported empirical interval coverage indicates the proportion of test observations falling within the constructed limits. However, because the residual quantiles and coverage are calculated using the same held-out residuals, the coverage estimate is descriptive rather than an independent validation of interval calibration. Moreover, the reliability of residual-based intervals may be affected by heteroscedasticity, non-normal error distributions, dependence between observations, or changes in the data-generating process. Within the proposed unified statistical-machine-learning framework, these diagnostics provide an initial assessment of predictive uncertainty and model adequacy. More rigorous uncertainty quantification could be achieved through an independent calibration dataset, conformal prediction, bootstrap resampling, or probabilistic modelling. These approaches would help assess the stability and reliability of predictions while maintaining a clear distinction between predictive uncertainty and uncertainty concerning causal relationships.

Table 4.8: Top marginal associations with X001

Feature	Pearson r with X001	Absolute r
X220	-0.7932	0.7932
X491	-0.7564	0.7564
X470	0.6040	0.6040
X033	0.3017	0.3017
X305	0.1956	0.1956
X074	-0.1318	0.1318
X053	-0.1283	0.1283
X129	-0.0969	0.0969
X035	0.0868	0.0868
X189	0.0541	0.0541
X109	0.0471	0.0471
X395	0.0462	0.0462
X056	0.0427	0.0427
X089	-0.0405	0.0405

X391

0.0400

0.0400

Table 4.8 presents the strongest marginal associations between the target variable X001 and the other numerical features in the simulated high-dimensional dataset. Pearson's correlation coefficient is used to quantify the direction and strength of the linear association between X001 and each candidate predictor. The table reports both the signed correlation coefficient and its absolute value, allowing the magnitude of each association to be assessed independently of its direction. A positive coefficient indicates that higher values of a feature tend to be associated with higher values of X001, whereas a negative coefficient indicates an inverse relationship. Larger absolute correlation values identify features with stronger linear associations with the target, making them potential candidates for further investigation in predictive modelling and explainability analysis. This screening procedure is useful in high-dimensional settings because it provides a straightforward way to identify variables that may contain relevant information while reducing the complexity of the initial feature space. However, marginal correlations do

not account for the joint influence of other predictors, and a feature with a weak individual correlation may still contribute useful information through nonlinear relationships or interactions. Similarly, strongly correlated features may contain redundant information and should not automatically be interpreted as independently important. Within the proposed unified statistical-machine-learning framework, these results provide an initial reference for comparing marginal associations with Elastic Net coefficients, permutation importance, and SHAP-based explanations. Agreement across these methods would offer complementary evidence about the model's predictive behaviour, while differences could reveal the influence of correlated predictors or conditional relationships. Nevertheless, none of these measures alone establishes causality. The reported associations should therefore be interpreted as exploratory statistical relationships, with causal conclusions requiring explicit identification assumptions and appropriate causal analysis.

4.3 Visual results

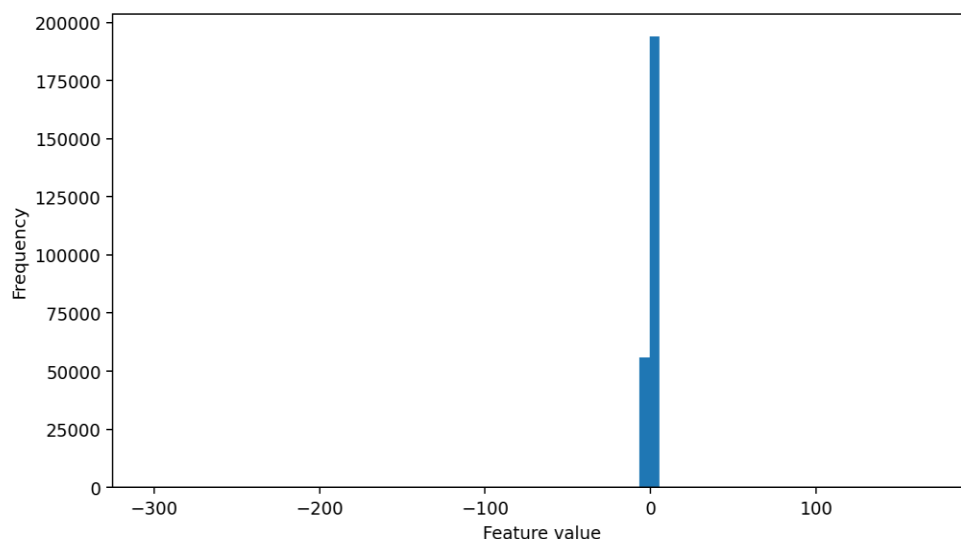


Figure 4.1: Distribution of sampled feature values.

The pooled value distribution combines all features and is therefore descriptive rather than a distribution for one specific variable. Heavy tails or multiple peaks may arise from heterogeneous feature scales or simulated substructures. Standardization is appropriate for PCA and regularized linear prediction, but should be performed within each training fold in a cross-validation pipeline to avoid leakage. Figure 4.1 illustrates the distribution of sampled numerical values across the features in the simulated high-dimensional dataset. The histogram provides an overview of the frequency of different values and helps identify the general shape, spread, and potential irregularities in the pooled feature distribution. Concentrations of observations around particular value ranges indicate where values occur most frequently, whereas extended tails may suggest the presence of extreme values or substantial variability. Multiple peaks may reflect differences in the distributions of individual features or distinct patterns introduced during data generation. Because the histogram combines values from multiple

features, its shape should not be interpreted as the distribution of any single variable. Differences in feature scales may also influence the overall pattern, making standardization an important consideration before applying scale-sensitive statistical and machine-learning methods. Within the proposed unified framework, this preliminary visualization supports data exploration and helps identify potential distributional characteristics that may affect model estimation, feature selection, and uncertainty assessment. However, the pooled histogram cannot establish causal relationships or independently determine whether the data satisfy the assumptions of a particular statistical model. Therefore, it should be complemented by feature-specific distribution plots, robust summary statistics, and residual diagnostics. Overall, Figure 4.1 provides an initial visual assessment of the numerical structure of the simulated dataset and establishes a foundation for the subsequent correlation, dimensionality-reduction, predictive-modelling, and explainability analyses.

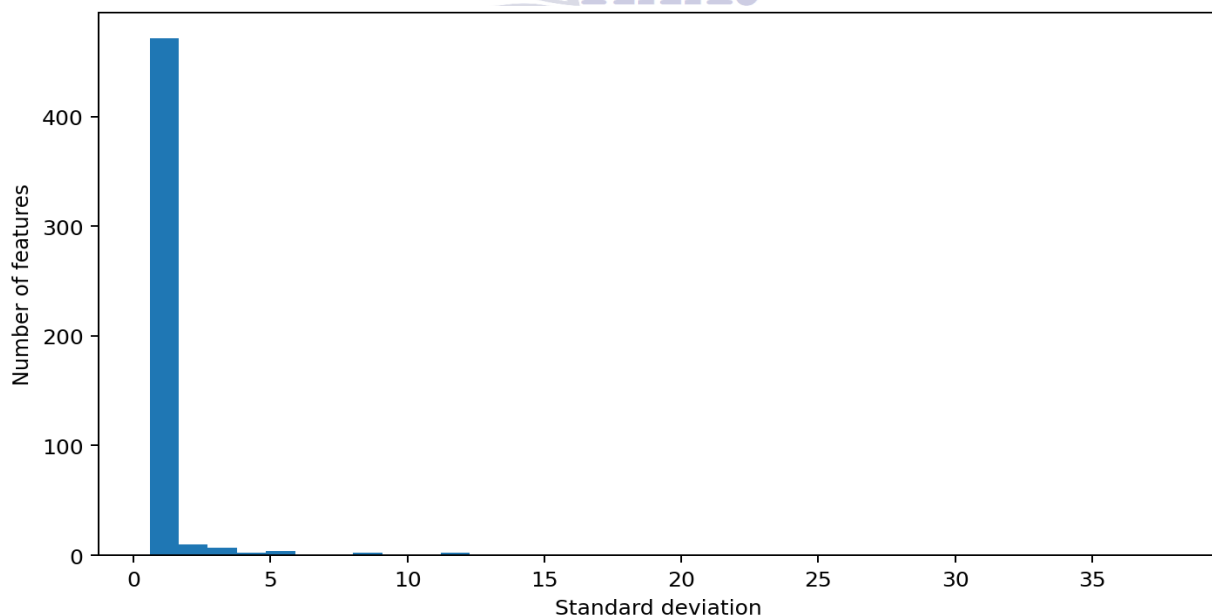


Figure 4.2: Distribution of feature standard deviations.

A long right tail indicates that a subset of features has much greater dispersion than most variables.

Such features can dominate distance-based methods and unscaled penalized models.

Standardization reduces scale dominance, while robust scaling should be considered if extreme values are genuine and influential.

Figure 4.2 presents the distribution of standard deviations across the 500 numerical features in the simulated high-dimensional dataset. Standard deviation measures the degree of variability in a feature relative to its mean and provides an important indication of how widely its observations are dispersed. Features with small standard deviations exhibit relatively limited variation, whereas features with larger standard deviations display greater dispersion across observations. The histogram helps identify whether variability is relatively consistent across the feature space or concentrated in a smaller subset of variables. A right-skewed distribution would indicate that most features have comparatively low or moderate variability, while a limited number exhibit substantially greater dispersion. Such differences are important in high-dimensional statistical analysis because variables with larger numerical scales may disproportionately influence certain machine-

learning algorithms, distance-based procedures, and regularized regression models. Standardization can address differences in scale by transforming features to have approximately zero mean and unit standard deviation, although robust scaling may be preferable when extreme observations are present. Within the proposed unified statistical-machine-learning framework, this assessment supports preprocessing decisions and helps improve the comparability of variables during dimensionality reduction and predictive modelling. Nevertheless, a large standard deviation does not necessarily indicate that a feature is more informative, important, or causally influential. Similarly, a feature with low variability may still contain valuable information about a particular outcome. Therefore, the results should be interpreted alongside correlation analysis, model performance, and feature-explainability measures. Overall, Figure 4.2 provides an essential preliminary assessment of feature-scale heterogeneity and informs the selection of appropriate preprocessing strategies for subsequent analyses.

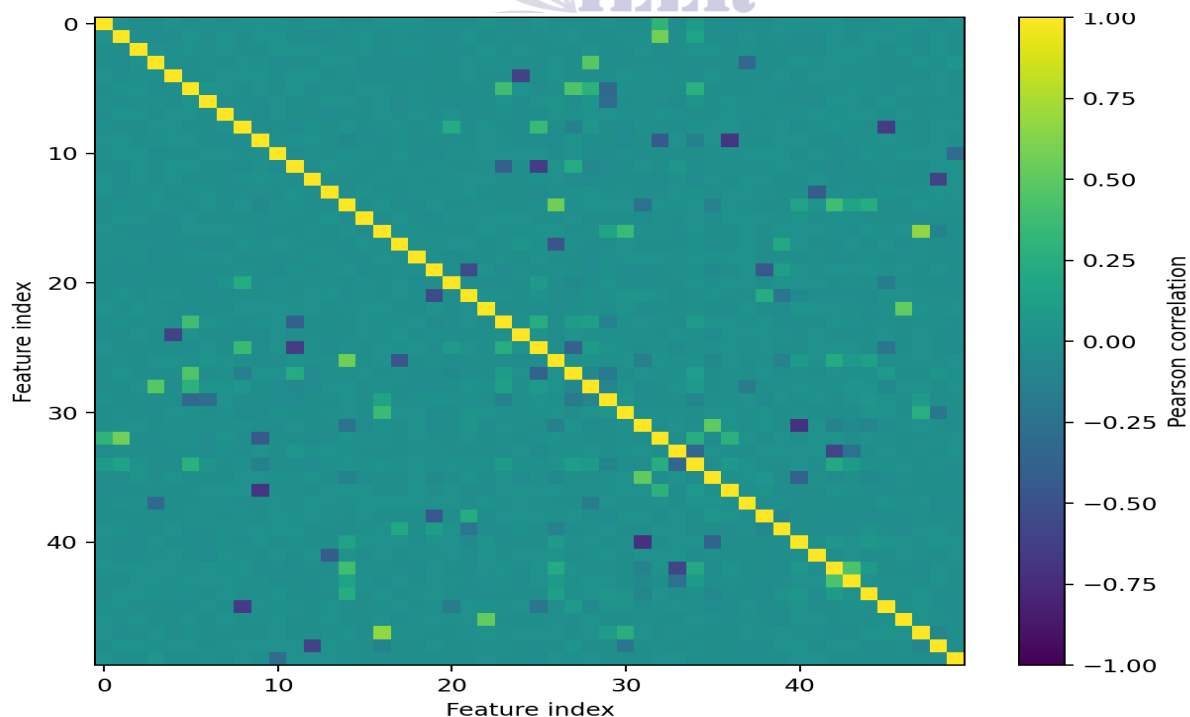


Figure 4.3: Correlation heatmap for the first 50 features.

Blocks or streaks of strong color suggest groups of features with shared linear information. The plot displays only the first 50 columns for legibility; the full matrix covers all 500 features. Correlation blocks are useful for dependency screening but are not a causal graph.

Figure 4.3 presents a correlation heatmap illustrating the pairwise linear relationships among the first 50 numerical features in the simulated high-dimensional dataset. Each cell represents the Pearson correlation coefficient between a pair of features, with the color intensity indicating the strength and direction of the association. Positive correlations indicate that two features tend to increase together, whereas negative correlations indicate that higher values of one feature tend to correspond to lower values of another. Values near zero suggest weak linear associations. The diagonal cells represent the correlation of each feature with itself and therefore have coefficients equal to one. Clusters of similarly colored cells may indicate groups of features that share statistical information or exhibit related patterns of variation. Such structures are important in high-dimensional modelling because strong dependencies among

predictors can introduce redundancy, complicate coefficient interpretation, and affect the stability of statistical estimates. The heatmap therefore provides a useful exploratory tool for identifying potentially correlated feature groups before applying dimensionality reduction, regularized regression, and model-explainability techniques. Within the proposed unified statistical-machine-learning framework, these patterns can also guide further investigation using conditional-independence tests and causal-discovery algorithms. However, the visualization covers only the first 50 features and does not represent the complete 500-feature dependence structure. Moreover, correlation does not establish causal direction or distinguish direct relationships from indirect associations caused by other variables. Consequently, the observed color patterns should be interpreted as evidence of statistical dependence rather than confirmed causal mechanisms. Overall, Figure 4.3 provides an informative visual summary of feature relationships and supports the subsequent analysis of multicollinearity, predictive performance, and causal-structure hypotheses.

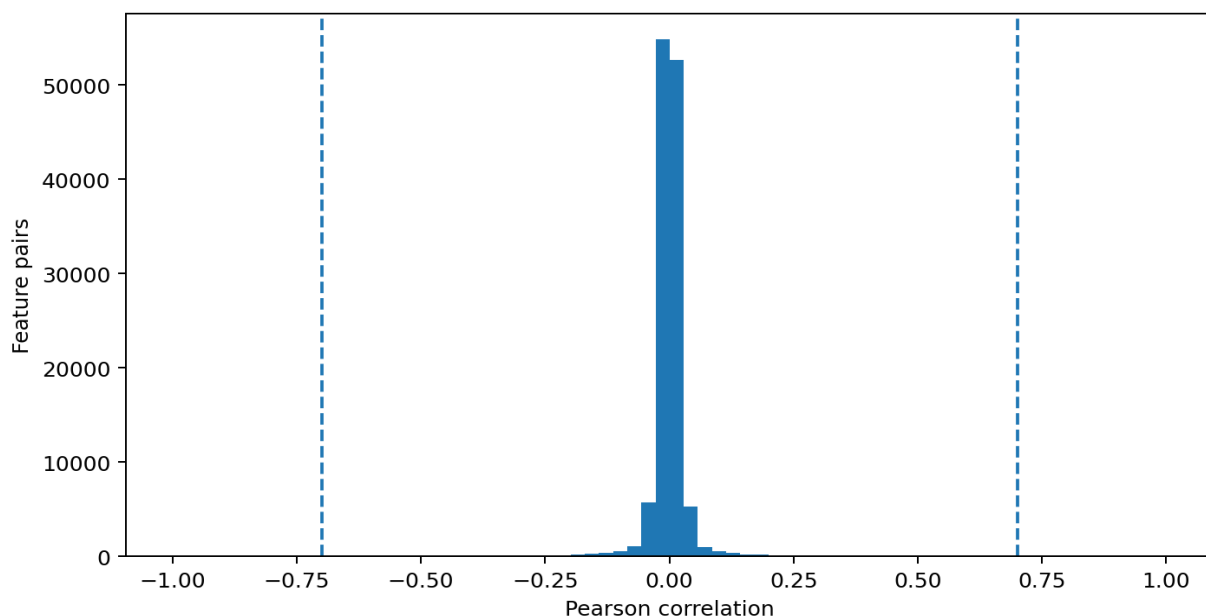


Figure 4.4: Distribution of unique pairwise feature correlations.

The spread of pairwise correlations indicates the degree of linear dependence across the feature space. The dashed thresholds at ± 0.70 are heuristic flags, not universal decision boundaries. With 124,750 unique pairs, some large correlations can occur by chance or through design, so multiple-testing-aware screening and stability checks are advisable.

Figure 4.4 illustrates the distribution of Pearson correlation coefficients calculated across all unique pairs of numerical features in the simulated high-dimensional dataset. With 500 features, the analysis considers 124,750 unique feature pairs, providing a broad overview of the linear dependence structure within the dataset. The histogram shows how frequently different correlation values occur, with positive coefficients representing direct associations and negative coefficients representing inverse associations. Values concentrated around zero would suggest that most feature pairs have relatively weak linear relationships, whereas a wider distribution would indicate greater variation in the strength and direction of dependence. The dashed reference lines at -0.70 and $+0.70$ identify feature pairs with comparatively strong absolute correlations for exploratory screening. These thresholds are heuristic indicators rather than universal

statistical criteria. Feature pairs exceeding these thresholds may contain overlapping information, which can affect multicollinearity, coefficient stability, and feature selection in high-dimensional regression models. The correlation distribution therefore helps characterize whether the feature space contains predominantly weak associations or substantial groups of strongly related variables. Within the proposed unified statistical-machine-learning framework, this assessment supports the selection of appropriate regularization techniques and guides further investigation of conditional dependencies. However, because a large number of feature pairs are examined simultaneously, extreme correlations should be interpreted carefully and assessed for stability across resampled datasets. Furthermore, pairwise correlations cannot identify causal direction, establish direct causal links, or rule out confounding by unobserved variables. Consequently, Figure 4.4 should be interpreted as a summary of statistical dependence rather than a validated causal structure. Overall, the figure provides an important foundation for evaluating predictor redundancy and designing subsequent causal-discovery, explainability, and uncertainty-quantification analyses.

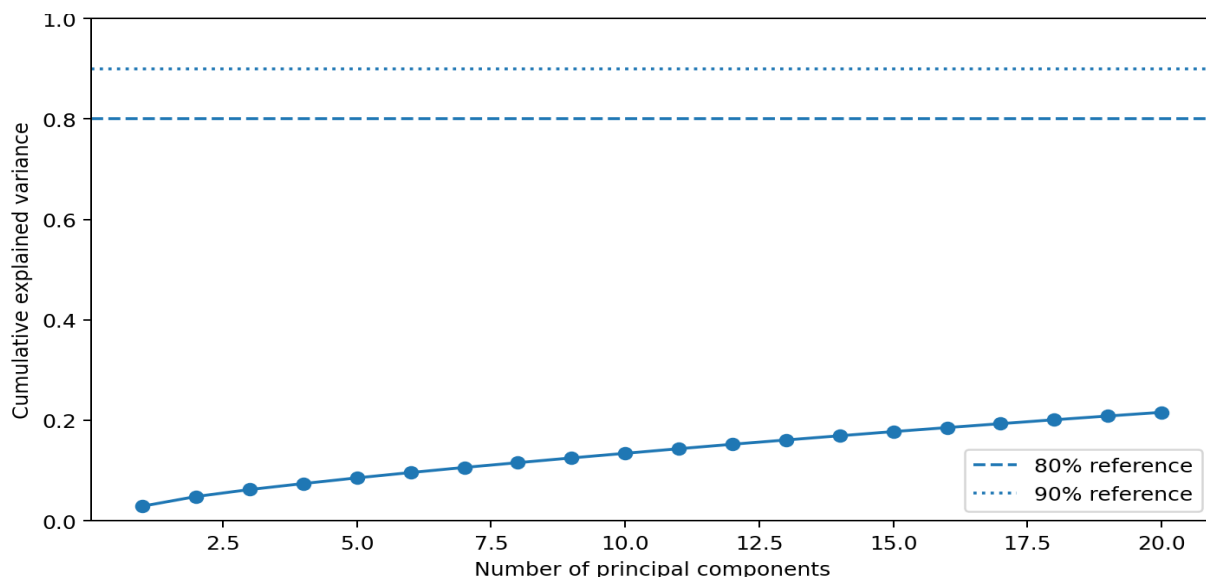


Figure 4.5: Cumulative variance explained by principal components.

The curve quantifies how rapidly the standardized feature matrix can be compressed. A steep early rise supports low-dimensional representation; a gradual rise suggests that many components are needed. PCA preserves variance rather than causal structure, so reduced components should be treated as predictive representations, not causal mechanisms.

Figure 4.5 illustrates the cumulative proportion of variance explained by the first 20 principal components obtained through Principal Component Analysis (PCA). The horizontal axis represents the number of retained principal components, while the vertical axis indicates the cumulative explained variance. PCA transforms the original 500 numerical features into a smaller set of uncorrelated components ordered according to their contributions to the total variance in the standardized dataset. As additional components are retained, cumulative explained variance increases because each component contributes additional information about the overall variability of the feature matrix. A steep initial increase indicates that a relatively small number of components captures a substantial proportion of the total variance, whereas a gradual increase suggests that information is distributed across many dimensions. The dashed reference lines at 80%

and 90% provide illustrative benchmarks for assessing the potential effectiveness of dimensionality reduction. If the cumulative curve approaches these thresholds with relatively few components, reducing the feature space may improve computational efficiency and simplify subsequent modelling. Conversely, if many components are required, aggressive dimensionality reduction may discard meaningful information. Within the proposed unified statistical-machine-learning framework, this visualization helps guide the selection of a lower-dimensional representation for predictive modelling and exploratory dependence analysis. Nevertheless, the appropriate number of components should not be determined exclusively by a variance threshold; predictive performance, interpretability, and stability should also be considered. Importantly, PCA preserves directions of high variance rather than identifying causal relationships. Consequently, components explaining large proportions of variance are not necessarily the most important for causal discovery or outcome prediction. Overall, Figure 4.5 provides a visual basis for evaluating dimensionality reduction while highlighting the need to balance information retention, computational efficiency, and the objectives of the proposed analytical framework.

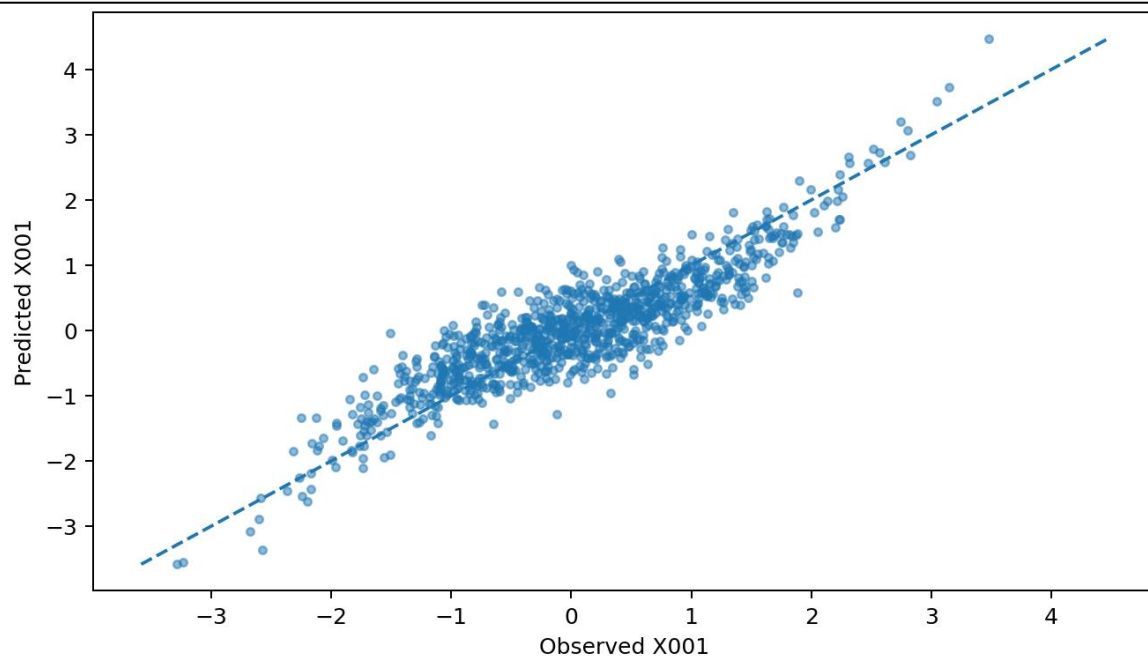


Figure 4.6: Observed versus predicted X001 values in the hold-out set.

Points close to the diagonal indicate accurate predictions; systematic curvature or widening scatter suggests model misspecification or non-constant error variance. The plot should be read together with R^2 , RMSE, MAE, and the residual diagnostic. A single random split can be sensitive to the split seed.

Figure 4.6 presents a comparison between the observed and predicted values of X001 obtained from the Elastic Net regression model using the held-out test dataset. The horizontal axis represents the actual observed values, while the vertical axis displays the corresponding model predictions. The dashed diagonal line represents perfect prediction, where predicted values are equal to observed values. Points located close to this line indicate that the model produces predictions with relatively small errors, whereas points farther from the line represent larger discrepancies between actual and predicted values. The distribution of points around the reference line provides a visual assessment of the model's predictive accuracy and its ability to capture variation in the target variable. The model achieved an R^2 value of 0.795, an RMSE of 0.446, and an MAE of 0.360 on the

held-out test dataset. These results indicate that the Elastic Net model captured a substantial proportion of the observed variation in X001, although some prediction errors remain. Systematic deviations from the diagonal line may indicate underprediction, overprediction, or limitations in representing complex relationships among the 499 candidate predictors. Within the proposed unified statistical-machine-learning framework, this visualization provides an intuitive assessment of predictive performance and complements the numerical evaluation metrics reported in Table 4.6. However, the figure evaluates predictive agreement rather than causal validity. Accurate prediction does not establish that the included features cause changes in X001, nor does it guarantee equivalent performance on new datasets. Therefore, the observed-versus-predicted plot should be interpreted alongside residual diagnostics, uncertainty estimates, and independent validation results. Overall, Figure 4.6 demonstrates the predictive behaviour of the fitted model and provides a foundation for evaluating its reliability within the broader analytical framework.

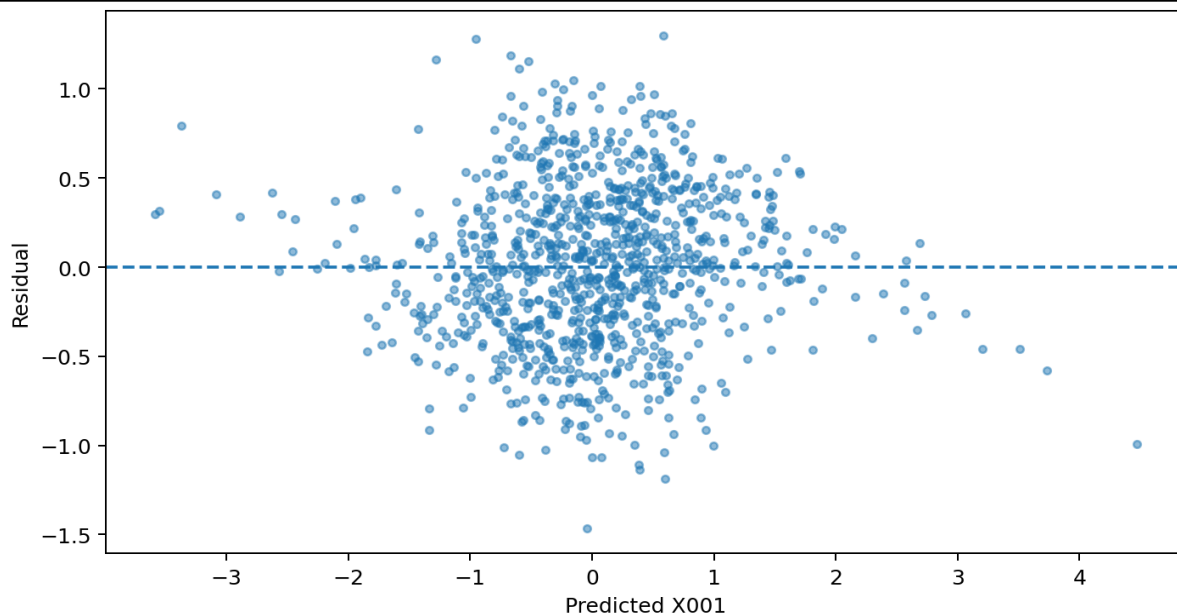


Figure 4.7: Residuals versus fitted values for X001.

Ideally, residuals fluctuate around zero without a systematic pattern. A funnel shape would suggest heteroscedasticity; curvature would suggest nonlinearity; isolated large residuals would warrant case-level inspection. Residual patterns concern predictive adequacy and do not validate causal assumptions.

Figure 4.7 presents the residuals-versus-fitted plot for the Elastic Net regression model developed to predict X001. The horizontal axis represents the predicted values, while the vertical axis shows the residuals, calculated as the differences between observed and predicted values. The horizontal dashed line at zero indicates the reference point at which predictions have no error. Ideally, residuals should be distributed randomly around this line without a clear systematic pattern, suggesting that the model adequately captures the principal relationships between the predictors and the target. If the residuals exhibit a curved pattern, this may indicate that important nonlinear relationships have not been captured by the linear model. Similarly, a funnel-shaped pattern may suggest heteroscedasticity, meaning that prediction errors vary across different fitted

values. Isolated residuals far from zero may identify observations requiring further investigation. This diagnostic is particularly important in high-dimensional modelling because regularization can improve coefficient stability while still leaving systematic patterns unexplained. Within the proposed unified statistical-machine-learning framework, residual analysis complements predictive performance metrics by examining whether the fitted model's errors display potential structural weaknesses. The plot also provides useful preliminary information for uncertainty quantification because non-constant residual variance can affect the reliability of prediction intervals. However, a visually random residual pattern does not establish causal validity or prove that all modelling assumptions are satisfied. Additional checks, including residual-distribution analysis, heteroscedasticity assessment, and repeated validation, may be required. Overall, Figure 4.7 provides a diagnostic assessment of model adequacy and helps identify potential improvements before interpreting predictions or uncertainty estimates.

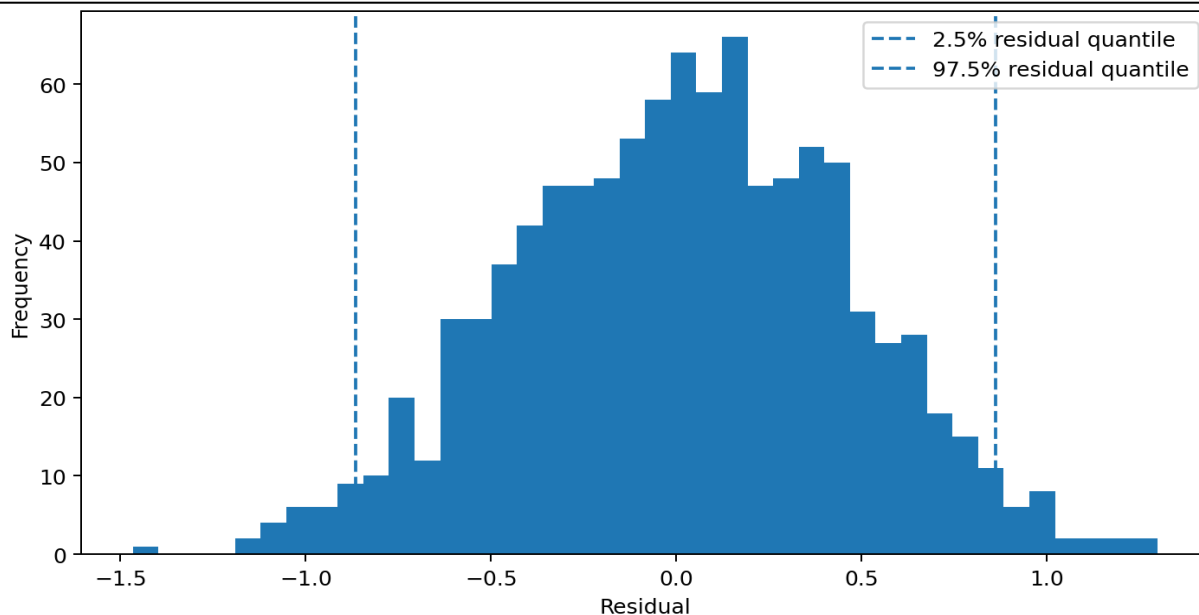


Figure 4.8: Hold-out residual distribution with empirical quantiles.

The histogram provides a visual check for asymmetry and heavy tails. The quantile limits define an empirical residual interval, not a guaranteed 95% prediction interval. To support formal uncertainty quantification, use conformal prediction or a calibrated probabilistic model and evaluate coverage on data not used to estimate the interval.

Figure 4.8 illustrates the distribution of residuals obtained from the Elastic Net regression model for the held-out test dataset. Residuals represent the differences between observed and predicted values of X001 and provide information about the magnitude and distribution of prediction errors. The histogram displays the frequency of residual values, while the two dashed vertical lines indicate the 2.5th and 97.5th percentiles of the empirical residual distribution. These quantiles define the central 95% range of observed residuals and provide a preliminary basis for constructing approximate prediction intervals. A distribution concentrated around zero suggests that prediction errors are generally centered near zero, whereas asymmetry may indicate systematic prediction errors. Heavy tails or isolated extreme values may suggest that some

observations are associated with considerably larger errors than others. Assessing these characteristics is important because the reliability of uncertainty estimates depends on the distribution and stability of prediction errors. Within the proposed unified statistical-machine-learning framework, the residual histogram complements the residual-versus-fitted plot by providing a distributional perspective on predictive uncertainty. The empirical quantiles can be used to construct residual-based prediction bands, but these should not automatically be interpreted as formally calibrated 95% prediction intervals. In particular, the same test residuals were used to estimate the quantiles and assess interval coverage, so an independent calibration dataset or conformal prediction procedure would provide a stronger evaluation. Moreover, non-constant error variance and distributional changes can reduce interval reliability. Overall, Figure 4.8 provides an initial assessment of prediction-error behaviour and highlights the importance of rigorous uncertainty quantification when developing reliable high-dimensional statistical and machine-learning models.

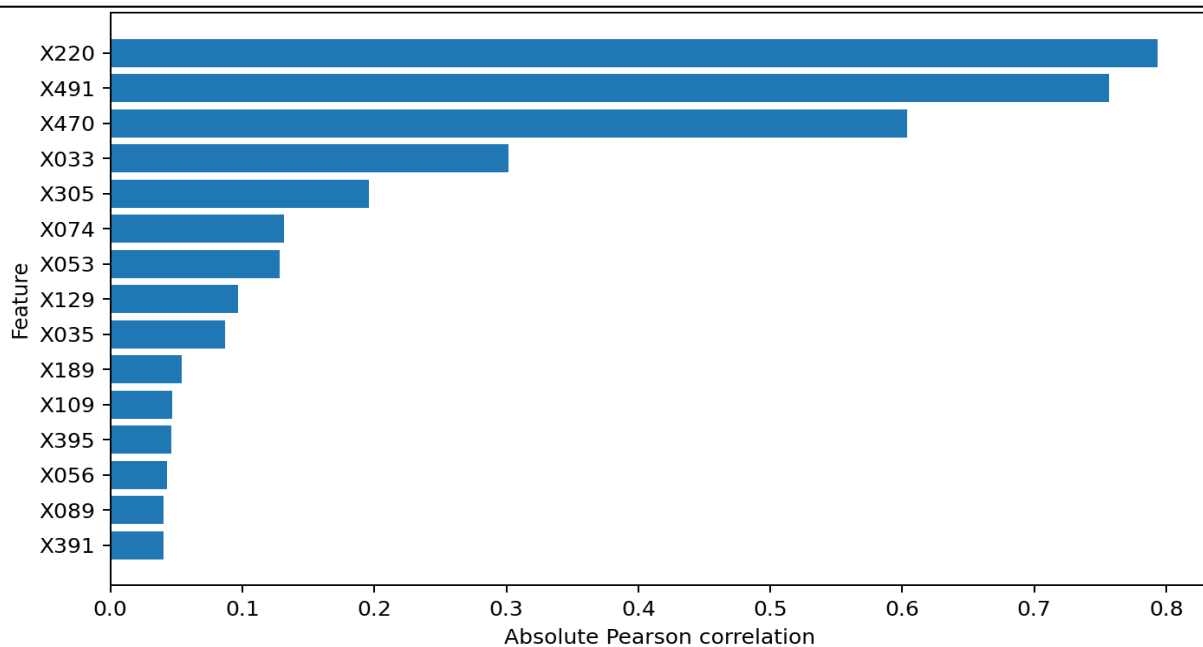


Figure 4.9: Strongest absolute marginal associations with X001.

The strongest marginal associations identify candidate predictors for further explanation. These bars are based on absolute Pearson correlations and are not SHAP values or conditional feature importance. A unified explainability workflow should compare marginal screening, regularized coefficients, permutation importance, and SHAP, then report whether rankings remain stable under resampling.

Figure 4.9 presents the 15 features with the strongest absolute Pearson correlations with X001, the illustrative target variable used in the predictive modelling analysis. The horizontal axis represents the absolute correlation coefficient, while the vertical axis identifies the corresponding features. Absolute correlation values range from zero to one, with larger values indicating stronger linear associations regardless of whether the original relationship is positive or negative. The visualization provides a comparative overview of which features exhibit the strongest individual linear relationships with the target. Features with larger absolute correlations may contain useful information for predicting X001 and can therefore be considered candidates for further investigation using

multivariate modelling and explainability techniques. However, marginal associations reflect relationships between individual features and the target without accounting for the simultaneous influence of other predictors.

Consequently, a feature with a strong marginal correlation may contribute limited additional predictive information when correlated variables are included in the same model. Conversely, a feature with a relatively weak marginal correlation may still be useful through nonlinear effects, interactions, or conditional relationships. Within the proposed unified statistical-machine-learning framework, these results provide an initial benchmark for comparing correlation-based screening with Elastic Net coefficients, permutation importance, and SHAP-based explanations. Agreement between these methods may strengthen confidence in the consistency of predictive explanations, whereas differences may reveal the influence of multicollinearity or complex feature interactions. Nevertheless, correlation-based rankings do not establish causal importance or indicate the direction of causal influence. Further investigation using explicit causal assumptions, conditional-dependence

analysis, and stability assessment is necessary before interpreting any feature as causally influential. Overall, Figure 4.9 highlights potentially informative predictors and demonstrates how preliminary statistical screening can support explainability and subsequent investigation within a high-dimensional analytical framework.

4.4 Integrated interpretation: causal discovery, explainability, and uncertainty

The results support a staged analytical architecture rather than a single algorithm. First, data-quality checks and descriptive summaries establish the scale, completeness, and potential anomalies of the feature matrix. Second, dependence screening and PCA characterize redundancy and effective dimensionality. Third, a regularized model provides a reproducible predictive baseline for a clearly declared target. Fourth, residual diagnostics and empirical error quantiles summarize predictive uncertainty. These components are complementary but answer different questions. Causal discovery requires an explicit causal estimand and defensible assumptions. For this dataset, there is no supplied intervention, temporal ordering, domain-defined target, or ground-truth directed acyclic graph. Consequently, a causal discovery algorithm such as PC, FCI, or NOTEARS could be run only as an exploratory structural hypothesis generator, with assumptions about causal sufficiency, faithfulness, functional form, and acyclicity stated explicitly. The output would need sensitivity analysis across thresholds and bootstrap resamples, and ideally comparison with a known simulated graph. Without that ground truth, precision, recall, structural Hamming distance, and causal-effect bias cannot be reported meaningfully. Explainability should likewise be tied to the model and the question. Pearson correlations summarize marginal linear association; Elastic Net coefficients summarize penalized conditional linear associations under scaling; permutation importance estimates performance degradation when a feature is disrupted; SHAP decomposes a model prediction

relative to a background distribution. None of these, by itself, is a causal explanation. Causal explanations require intervention-based quantities or assumptions that identify counterfactual contrasts. For uncertainty quantification, distinguish parameter uncertainty, predictive uncertainty, and structural uncertainty. The residual band in this chapter is a basic predictive-error summary. A stronger framework would use repeated nested cross-validation for model-selection uncertainty, conformal prediction for finite-sample marginal coverage under exchangeability, bootstrap resampling for feature-selection stability, and sensitivity analyses over causal assumptions. For high-dimensional causal effects, debiased/desparsified estimators or double/debiased machine learning may be appropriate when treatment, outcome, and confounders are defined and identification assumptions are plausible.

6. Conclusion

This study, “A Unified Statistical–Machine Learning Framework for Causal Discovery, Explainability, and Uncertainty Quantification in High-Dimensional Data,” presents an integrated methodological framework for analyzing complex datasets in which predictive accuracy, interpretability, causal reasoning, and uncertainty assessment must be considered together. The study uses a simulated high-dimensional dataset containing 5,000 observations and 500 numerical features, providing a substantial basis for demonstrating the proposed analytical workflow. By combining statistical preprocessing, dimensionality assessment, regularized machine learning, model explanation, causal discovery principles, and uncertainty evaluation, the framework addresses several methodological challenges that often arise when high-dimensional data are analyzed using isolated techniques.

The predictive analysis using Elastic Net regression produced a held-out coefficient of determination (R^2) of 0.795, indicating that the model explained approximately 79.5% of the observed variation in the illustrative prediction

target. The corresponding root mean squared error (RMSE) was 0.446, while the mean absolute error (MAE) was 0.360. Together, these results demonstrate promising predictive performance within the simulated experimental setting. However, these metrics evaluate predictive performance rather than causal validity; they do not establish that the selected features are causal determinants of the outcome. The results should therefore be interpreted as evidence of predictive utility under the study's simulation conditions, rather than as proof of causal relationships in real-world systems. A principal contribution of the proposed framework is its emphasis on distinguishing prediction, explanation, and causal discovery. Predictive modelling identifies patterns that help estimate outcomes, while explainability methods, including model coefficients, permutation-based feature importance, and SHAP, can help clarify how a fitted model generates predictions. Causal discovery requires additional assumptions about the underlying data-generating process, conditional independence, possible latent confounding, and the adequacy of the observed variables. Accordingly, a feature that is highly predictive or receives a large importance score should not automatically be interpreted as a direct cause of the outcome. The framework further emphasizes uncertainty quantification and stability assessment as essential components of reliable analysis. Bootstrap resampling, repeated model fitting, sensitivity analysis, and stability assessment can help determine whether estimated relationships and feature rankings remain consistent under reasonable changes in the data or modelling procedure. This is particularly important when the number of candidate features is large relative to the available information and when multiple correlated variables may convey similar predictive signals. Incorporating these checks can make conclusions more transparent and reduce the risk of drawing strong conclusions from unstable model outputs.

REFERENCES

- Aalen, O. O., Borgan, Ø., & Gjessing, H. K. (2008). *Survival and event history analysis: A process point of view*. Springer.
- Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360.
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1), 233–298.
- Berkessa, Z. A., Läärä, E., & Waldmann, P. (2025). A review of causal methods for high-dimensional data. *IEEE Access*, 13, 11892–11917. <https://doi.org/10.1109/ACCESS.2024.3524261>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.
- Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3, 507–554.
- Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450), 407–424.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. <https://doi.org/10.18637/jss.v033.i01>

- Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, Article 524. <https://doi.org/10.3389/fgene.2019.00524>
- Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1), 37-48.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
- Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.
- Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21* (pp. 689-696).
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.
- Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., & Silva, R. (2025). Causal machine learning: A survey and open problems. *Foundations and Trends in Optimization*, 9(1-2), 1-247. <https://doi.org/10.1561/24000000052>
- Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8, 613-636.
- Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT Press.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765-4774).
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417-473.
- Meinshausen, N., & Bühlmann, P. (2015). Maximin effects in inhomogeneous large-scale data. *The Annals of Statistics*, 43(4), 1801-1830.
- Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
- Nogueira, S., Sechidis, K., & Brown, G. (2018). On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174), 1-54.
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
- Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.
- Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15, 2009-2053.
- Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550-560.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). <https://doi.org/10.1145/2939672.2939778>
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612-634.

- Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2003–2030.
- Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.
- Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3, Article 3. <https://doi.org/10.1186/s40535-016-0018-x>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- Vowels, M. J., Camgoz, N. C., & Bowden, R. (2022). D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4), Article 82. <https://doi.org/10.1145/3527154>
- Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence* (pp. 804–813).
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.
- Khan, R., Khan, A., Muhammad, I., & Khan, F. (2025). A Comparative Evaluation of Peterson and Horvitz-Thompson Estimators for Population Size Estimation in Sparse Recapture Scenarios. *Journal of Asian Development Studies*, 14(2), 1518–1527.

