Causal Inference: A Comparative Analysis Of IPW And AIPW Methods With Machine Learning Models

Habeeb Olawale HAMMED

Université Côte d'Azur

Abstract. Causal inference plays a crucial role in understanding the effects of interventions, which is relevant in healthcare where randomized controlled trials are often infeasible. In this study, we employ two widely used methodologies: Inverse Probability Weighting (IPW) and Augmented Inverse Probability Weighting (AIPW) in Causal Inference. A directed acyclic graph (DAG) is constructed to model causal relationships, ensuring a valid adjustment set for estimation. We explore various outcome models, including Linear Regression, Random Forest, and Gradient Boosting, to assess the robustness of AIPW. The results indicate that AIPW provides a more stable and efficient estimate compared to IPW, particularly in handling non-linearities and reducing variance. Robustness checks, including placebo tests and subset data refuter analysis, confirm the validity of the estimated treatment effect. Further comparisons between CausalML and EconML implementations reveal slight variations in treatment effect estimation, emphasizing the importance of model choice. We demonstrate the effectiveness of AIPW over IPW, highlighting the need for double-robust estimators in observational studies to improve causal inference accuracy.

1 Introduction

Causal learning focuses on cause-and-effect relationships. Instead of seeking correlations between features, causal learning aims to determine asymmetric causal relationships. Causal learning can be divided into two main areas: Causal Discovery and Causal Inference. The former focuses on identifying the causal structure among variables, while the latter estimates the magnitude of causal effects once causal relationships have been established. In this work, we are specifically interested in causal inference.

2 Basic Foundation of Causal Inference

In what follows, we introduce the foundational notations and graphical representations necessary for causal inference.

We consider a set of units $i = 1, \ldots, n$, each of them possibly associated with the following variables:

- **Covariates ($X_i \in \mathbb{R}^p$)**: A vector of p-dimensional observed characteristics.
- **Treatment Assignment ($T_i \in \{0, 1\}$)**: Indicates whether unit i receives treatment ($T = 1$) or remains untreated ($T = 0$).

- **Outcome ($Y_i \in \mathbb{R}$)**: Observed result or effect of interest for unit i i.e., we restrict ourselves to the use of binary treatment.

- **Noise (e_i)**: Independent variability associated with each variable X_i, which captures unexplained random effects.

Confounder: Observed or unobserved covariates that influences both T and Y. It is a potential source of bias in estimating the causal effect between T and Y

.

A Directed Acyclic Graph (DAG) is used to represent causal relationships among variables. It is denoted as $(G = (V, E))$ where

- V: The set of nodes (variables), for example T, Y, X.
- E: The set of directed edges (arrows) that represent causal relationships between variables. An example of DAG is illustrated by Fig 1.

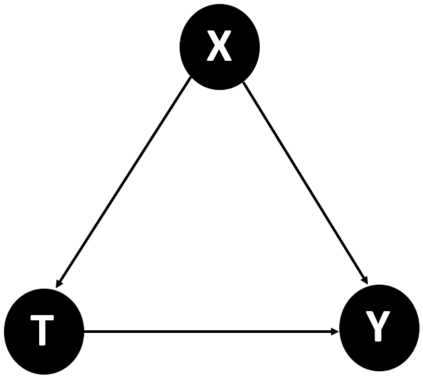


Fig. 1: A Directed Acyclic Graph (DAG) illustrating causal relationships ([3]) Here X is a confounder.

Note that DAG 1 does not contain any cycles; i.e, it is impossible to start at a node, follow the directed edges, and return to the same node.

The factorization formula was formally introduced by Pearl [4], who laid the groundwork for modern causal graphical models. It allows to compute the probability of the entire system of variables represented by a DAG. It describes how the joint probability distribution of observed variables can be decomposed into a product of conditional probabilities.

The factorization assumes that each variable X_i is generated as a function of its parents (Pa_i) i.e., the direct causes of X_i, and an independent noise term (e_i):

$$P(X) = \prod_{i=1}^{n} P(X_i \mid \mathrm{Pa}_i, e_i) \tag{1}$$

This decomposition is essential for deriving conditional independence and identifying causal effects. Using the relationship in 1, the decomposition becomes:

$$P(X, Y, T) = P(Y \mid T, X, e_i).P(T \mid X, e_i).P(X \mid e_i) \tag{2}$$

3 Potential Outcome Framework

The Potential Outcome Framework, introduced by Rubin [5], formalizes causal effects by considering outcomes under different treatment scenarios.

For each individual, only one of the potential outcomes is observed i.e., it is treated or not. The observed outcome for unit i can be defined as:

$$Y_i = T_i \cdot Y_i^{(1)} + (1 - T_i) \cdot Y_i^{(0)}, \tag{3}$$

where:

- (Y_i) represent the observed outcome for individual i
- (T_i) represent the treatment indicator for individual i
- (Y_i^1) signifies the potential outcome if treated
- (Y_i^0) signfies the potential outcome if not treated.

Several metrics have been proposed to quantify the treatment effects depending on the scope.

Average Treatment Effect (ATE)

The Average Treatment Effect(ATE) measures the average causal effect of a treatment across an entire population. It's the most common metrics for estimating causal effects and is defined as:

$$ATE = \mathbb{E}[Y_1 - Y_0] \tag{4}$$

This measures the expected difference in outcomes if everyone in the population were treated [Y_1] versus if no one were treated [Y_0].

Conditional Average Treatment Effect (CATE)

This metrics quantifies the average causal effect of a treatment for a specific subgroup, given certain covariates (X). As such, it can help identify some variations in treatment effects across subgroups. It is mathematically defined as:

$$CATE(X = x) = \mathbb{E}[Y_1 - Y_0 \mid X = x] \tag{5}$$

4 Causal Inference Techniques

Estimating causal effects is a central challenge in both experimental and observational studies. The gold standard procedure involves randomly assigning subjects to treatment or control groups to ensure that differences in outcomes are due to treatment and not confounding factors; Randomized Controlled Trials. However, this is often impractical due to financial, ethical, or logistical constraints. When randomization is not possible, causal effects can be estimated from observational data. To do that, methods to adjust for confounding variables are required.

In this work, we are specifically interested in two techniques: Inverse Probability Weighting(IPW) [1] and the Augmented Inverse Probability Weighting(AIPW) [2].

4.1 Inverse Probability Weighting (IPW)

Objective: To estimate causal effects by creating a pseudo-population where treatment assignment is independent from observed covariates, mimicking a randomized controlled trial.

Parameters:

- **Propensity Score ($e(X)$)**: The probability of receiving the treatment given covariates X, defined as:

$$e(X) = P(T = 1 \mid X) \tag{6}$$

Key Features

- It balances covariates between treated and untreated groups.
- It provides unbiased estimates of the ATE 4 if the propensity score model 6 is correctly specified.

Steps:

1. **Estimate the Propensity Score**: Use logistic regression or machine learning algorithms to estimate $e(X)$ 6.

2. **Compute Weights**:

$$w_i = \begin{cases} \frac{1}{e(X_i)} & \text{if } T_i = 1 \\ \frac{1}{1-e(X_1)} & \text{if } T_i = 0 \end{cases} \tag{7}$$

3. **Apply Weights**: Fit a weighted regression model where the outcome (Y_i) is regressed on the treatment (T_i) using the computed weights (w_i) as observation weights. This is done as follows:

$$\hat{Y} = \beta_0 + \beta_1 T_i,$$

   where:
   - β_0: Baseline outcome (intercept).
   - β_1: Treatment effect in the weighted pseudo-population.
4. **Estimate the ATE** using the IPW estimator defined as:

$$\hat{\mu}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{T_i Y_i}{e(X_i)} - \frac{(1 - T_i) Y_i}{1 - e(X_i)} \right]. \tag{8}$$

   This estimator uses weights based on the inverse of the propensity score to balance treated and untreated units in the population.

4.2 Augmented Inverse Probability Weighting (AIPW)

While IPW provides a robust framework for adjusting for confounding, it can be sensitive to misspecification of the propensity score model. To enhance its effectiveness, an extension of IPW [1] has been proposed: **Augmented Inverse Probability Weighting (AIPW)** [2]. This method incorporates double robustness by combining IPW [1] with outcome regression models.

Objective: To improve the efficiency and robustness of causal effect estimation by combining IPW with outcome regression models.

Parameters:

- **Propensity Score** ($e(X)$) 6: as in IPW.
- **Outcome Model** ($\hat{Y}(X, T)$): A model predicting the outcome Y based on covariates X and treatment T.

Key Features:

- **Double Robustness**: It provides consistent estimates if either the propensity score model or the outcome model is correctly specified.
- **Improved Efficiency**: It achieves lower variance compared to IPW alone.

Steps:

1. **Estimate the Propensity Score** same as for IPW 6.

2. **Fit the Outcome Model**: Predict Y using covariates X and treatment T.

$$\hat{Y}(X_i, 1) = \mathbb{E}[Y_i \mid T_i = 1, X_i], \quad \hat{Y}(X_i, 0) = \mathbb{E}[Y_i \mid T_i = 0, X_i]. \tag{9}$$

$\hat{Y}(X_i, 1)$ and $\hat{Y}(X_i, 0)$ are predicted outcomes under treatment and control. These are often obtained by fitting a regression model:

$$\hat{Y}(X_i, t) = \beta_0 + \beta_1 T_i + \boldsymbol{\beta}_X X_i, \tag{10}$$

where $t \in \{0, 1\}$.

3. **Compute the AIPW Estimator**:

$$\hat{\mu}_{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{Y}(X_i, 1) - \hat{Y}(X_i, 0) + \frac{T_i \cdot (Y_i - \hat{Y}(X_i, 1))}{e(X_i)} - \frac{(1 - T_i) \cdot (Y_i - \hat{Y}(X_i, 0))}{1 - e(X_i)} \right] \tag{11}$$

The terms $\frac{T_i \cdot (Y_i - \hat{Y}(X_i,1))}{e(X_i)}$ and $\frac{(1-T_i) \cdot (Y_i - \hat{Y}(X_i,0))}{1-e(X_i)}$ adjust for discrepancies between observed and predicted outcomes.

4.3 Comparison of IPW and AIPW

The key differences between IPW and AIPW are summarized in Table 1.

IPW	**AIPW**
Relies solely on the propensity score model.	Uses both propensity score and outcome models.
Sensitive to misspecification of the propensity score model.	Consistent if either model is correctly specified.
Can have high variance with extreme weights.	Typically achieves lower variance.
Simpler, requiring only the propensity score model.	More complex, requiring two models.
Not double robust; needs a correct propensity score model.	Double robust; valid if either propensity or outcome model is correct.

Table 1: Comparison of IPW and AIPW [1].

IPW and AIPW share several similarities:

- Both methods are designed for observational studies where randomization is not feasible.
- Both rely on the assumption of no unmeasured confounders (i.e., all variables affecting treatment and outcome are observed).
- Both methods can work with binary, continuous, or mixed outcomes.
- Both use propensity scores (directly or indirectly) to adjust for confounding variables, ensuring valid comparisons between treated and untreated groups.

5 Experiments and Results

We generated 5000 units of synthetic dataset with the following characteristics to mimic real-world treatment assignments and outcomes:
Covariates:

- Age (Continuous)
- Sex (Binary: 0 = Female, 1 = Male)
- Heart Rate (Continuous)
- Blood Pressure (Continuous)

Confounders:

- Comorbidity Index (Categorical: 0 = Low, 1 = Moderate, 2 = High)
- Lifestyle Factors (Binary: 0 = Healthy, 1 = Unhealthy)

Treatment Assignment: Drug Administered(Binary: 0 = Not Treated, 1 = Treated)
Outcome Variable: Proarrhythmic Risk

We ensured that the treatment assignment model is well-calibrated, so that individuals have a balanced spread of treatment probabilities. Fig 2 shows the distribution of the treatment assignment probabilities.

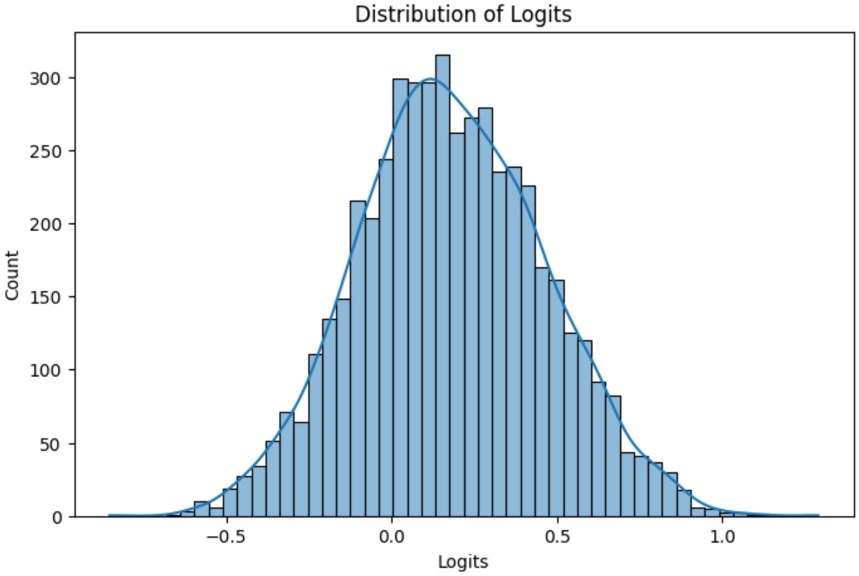


Fig. 2: The distribution of treatment assignment probabilities.

Causal DAG Representation: A Directed Acyclic Graph (DAG) is used to represent the assumed causal relationships between variables.

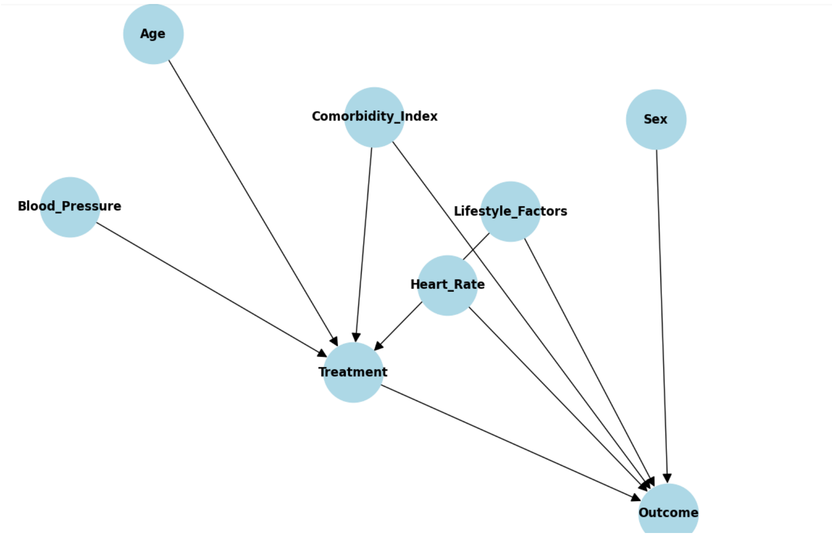


Fig. 3: A Directed Acyclic Graph (DAG) illustrating causal relationships among variables of the synthetic dataset.

From Fig 3, we can see that the treatment assignment is influenced by: Age, Blood Pressure, Comorbidity Index, and Lifestyle Factors. Similarly, the Outcome is influenced by: Treatment, Comorbidity Index, Lifestyle Factors, Heart Rate, and Sex. This DAG ensures that we account for all possible confounding factors when estimating the causal effect of the treatment on the outcome.

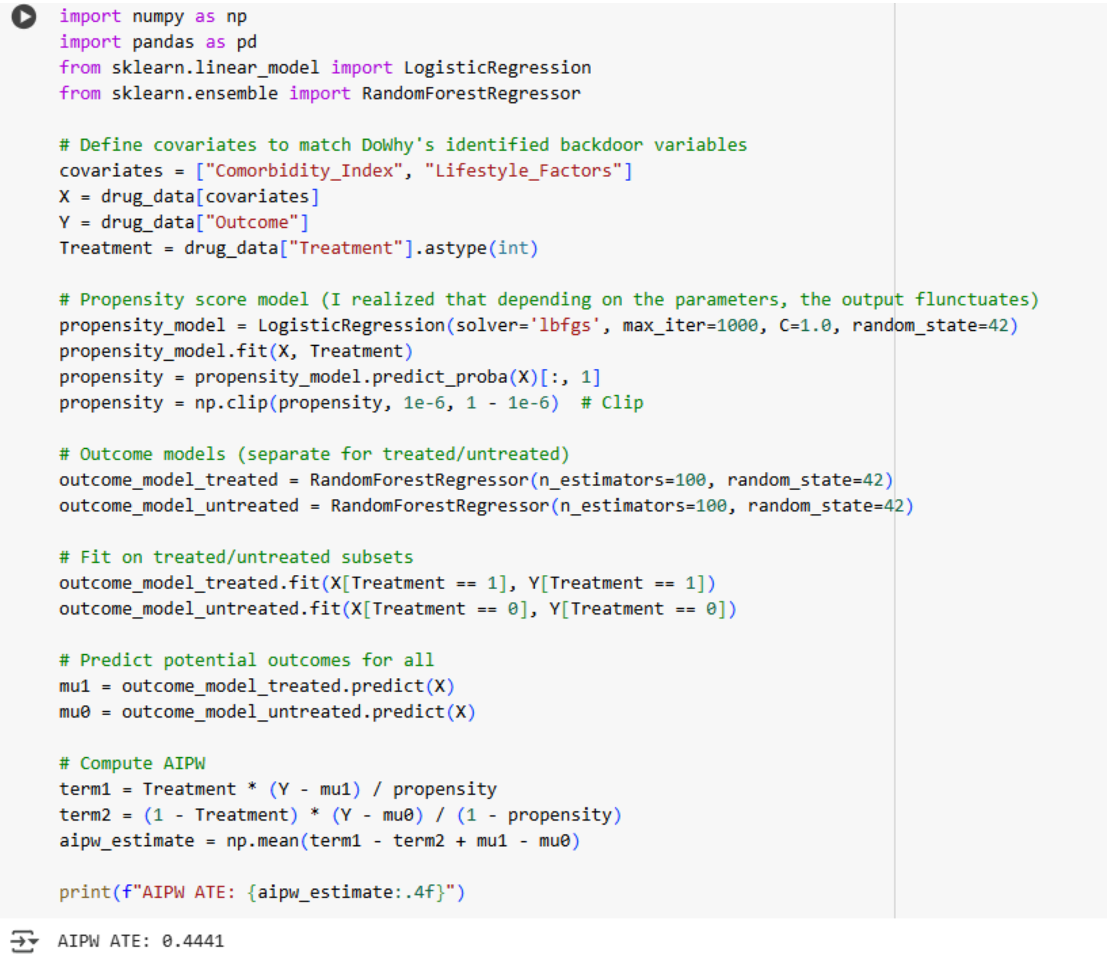
Experimentation & Results:
The dataset was generated as shown in the DAG 3. We study the causal effect of treatment on the outcome using two main methods: Inverse Probability Weighting (IPW) & Augmented Inverse Probability Weighting (AIPW). We also experimented with different modeling choices for propensity score estimation and outcome regression models, while also conducting robustness checks to validate the reliability of our estimates. The results of the causal estimates using the doWhy package are summarized below:

Interpretation:

1. **IPW Estimate Is Higher than AIPW**
   - IPW ATE estimate gives 0.5316, while AIPW ATE estimate has a mean value of 0.4879.
   - IPW is likely **overestimating the effect** due to **higher sensitivity to extreme weights**.

- AIPW is more robust since it combines **propensity scores and outcome models** to correct for variance.

We also did a manual implementation 4 of the AIPW ATE estimate to display a thorough grasp of the concept. The result computed was similar to the one obtained by the doWhy package. The resulting slight variance can be linked to the hyperparameters tuning.



```
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestRegressor

# Define covariates to match DoWhy's identified backdoor variables
covariates = ["Comorbidity_Index", "Lifestyle_Factors"]
X = drug_data[covariates]
Y = drug_data["Outcome"]
Treatment = drug_data["Treatment"].astype(int)

# Propensity score model (I realized that depending on the parameters, the output flunctuates)
propensity_model = LogisticRegression(solver='lbfgs', max_iter=1000, C=1.0, random_state=42)
propensity_model.fit(X, Treatment)
propensity = propensity_model.predict_proba(X)[:, 1]
propensity = np.clip(propensity, 1e-6, 1 - 1e-6)  # Clip

# Outcome models (separate for treated/untreated)
outcome_model_treated = RandomForestRegressor(n_estimators=100, random_state=42)
outcome_model_untreated = RandomForestRegressor(n_estimators=100, random_state=42)

# Fit on treated/untreated subsets
outcome_model_treated.fit(X[Treatment == 1], Y[Treatment == 1])
outcome_model_untreated.fit(X[Treatment == 0], Y[Treatment == 0])

# Predict potential outcomes for all
mu1 = outcome_model_treated.predict(X)
mu0 = outcome_model_untreated.predict(X)

# Compute AIPW
term1 = Treatment * (Y - mu1) / propensity
term2 = (1 - Treatment) * (Y - mu0) / (1 - propensity)
aipw_estimate = np.mean(term1 - term2 + mu1 - mu0)

print(f"AIPW ATE: {aipw_estimate:.4f}")
```

```
AIPW ATE: 0.4441
```

Fig. 4: Manual Implementation of the AIPW estimate

Robustness Checks Confirm Causal Estimate Validity: Furthemore, we performed several robustness check to confirm the validity of our causal estimate. Here's what we observed:

- **Placebo Test** → AIPW correctly drops to **near zero effect (**$\sim$ 0.01**, p=0.8)** when treatment is randomized.
- **Subset Data Refuter Test** → AIPW remains stable (0.4879 → 0.4264, p=0.3), showing it is **not overly sensitive to sample size**.

AIPW estimate drops from 0.4879 to 0.0105 when a fake treatment is used. p-value = 0.8 → No significant effect remains. we can conclude that AIPW is valid because it detects no causal effect when treatment is randomized. The estimate changes slightly (0.4879 → 0.4264) when using a random subset of the data. p-value = 0.3 → No strong evidence that the effect changed significantly. so we can also conclude that AIPW is relatively stable but shows some sensitivity to sample size.

In the case of IPW, the ATE also remained stable with a high P-value when we estimated the effect on only a subset of the data, showing that the estimated effect isn't just due to outliers or data selection bias. Similarly, When we added a random, fake confounder, the estimated effect did not change at all.It means that adding irrelevant variables doesn't change the result. Conclusively, both causal estimates passed all robustness check thus validating the strength of our analysis.

Effect of Different Outcome Models on AIPW: We tested different combinations of models for propensity score estimation and outcome modeling in AIPW estimation to determine how well the choice of the propensity score influences the outcome.

- **Linear Regression** → Highest AIPW estimate (0.4837), likely overestimating due to linearity assumption.
- **Random Forest** → Lowest AIPW estimate (0.4416), suggesting better handling of non-linearity.
- **Gradient Boosting** → Intermediate AIPW estimate (0.4701), balancing flexibility and stability.

Comparing AIPW Estimates from CausalML vs. EconML:

We also estimated AIPW using two other frameworks(CausalML and EconML) to see how the choice of the package can affect the estimate outcome. Both estimates are relatively close to the estimate computed by doWhy, all are likely capturing the true treatment effect. CausalML's estimate is slightly lower, which could mean that it captures non-linear effects better. EconML assumes linearity, leading to a slightly higher estimate.

6 Conclusion

Conclusion This study provides analysis of causal inference methodologies for estimating the effect of treatment on outcome using synthetic data. We compared Inverse Probability Weighting (IPW) and Augmented Inverse Probability Weighting (AIPW), demonstrating that AIPW yields more reliable results due to its double-robust nature. Robustness checks validate the causal estimates, with placebo tests confirming no significant effect under randomized treatment, and subset data refuter analysis demonstrating stable estimates.

The choice of outcome model impacts AIPW estimates, with Random Forest yielding more conservative effects compared to Linear Regression and Gradient Boosting. These findings underscore the importance of combining propensity score weighting with outcome regression models for more accurate causal effect estimation in healthcare studies. Future research can extend this framework to real-world clinical datasets, incorporating advanced machine learning techniques for more robust and scalable causal inference.

References

1. Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.
2. Christoph F. Kurz. Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167, 2022.
3. Ana Rita Nogueira et al. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, 12(2):e1449, 2022.
4. Judea Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2009.
5. Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.