

Machine Learning for Causal Inference from Observational Data



**Registration Number:**

**Project:**

**Link to GitHub:**

**Table of Contents**

[**Abstract** 2](#_Toc96333259)

[**1.** **Introduction** 2](#_Toc96333260)

[**2.** **Data** 3](#_Toc96333261)

[**3.** **Methodology** 4](#_Toc96333262)

[**3.1** **EconML** 4](#_Toc96333263)

[**3.2** **CATE Estimators** 4](#_Toc96333264)

[**3.3** **StandardScaler** 4](#_Toc96333265)

[**3.4** **Data Processing & Cleaning** 4](#_Toc96333266)

[**4.** **Implementation & Result** 5](#_Toc96333267)

[**5.** **Conclusion** 5](#_Toc96333268)

[**References** 5](#_Toc96333269)

# **Abstract**

The estimation of causal effects based on observations continues to increase in popularity within the epidemiological literature. Inference on causality has many practical applications across a variety of domains like marketing, health care, politics, and online advertising. The estimation of treatment effects, which is a key issue in causal inference has been studied extensively in statistics for a long time. However, conventional treatments for estimation of treatment effects might not be able to deal with large-scale or high-dimensional heterogeneous information. The last 30 years have witnessed numerous conceptual advances that offer part of the solution to the issue of causal inference using observational samples or a mix of experimental and observational data, specifically in the field of graphic causal modeling. A major challenge with causal inference is the lack of data and small sample sizes. The presence of unobserved causes can lead to a reduction in the space of alternative causal hypotheses. In addition, limited background knowledge and experimental interventions can lead to significant biases in the outputs of the model. In such a scenario, thousands of experiments are needed to uncover the underlying mechanisms. In general, this research has great potential to advance the science of predictive analytics. This introduction to the Special Topic on Causality provides an overview of graphical causal modeling. It puts the thesis into the larger context of causal modeling and discusses the difference from causal modeling and normal machine learning problems for classification and prediction.

# **Introduction**

Causal inference in machine learning is the process of identifying the causes of a particular phenomenon. Although this technique helps in some instances, it is not enough to make decisions in many scenarios (Crown, 2019). While this technique is useful for some purposes, it is not always appropriate for others. For example, a causal model may fail to identify the cause of a disease or epidemic. However, it is useful in many situations. For example, it can be used to identify the causes of the spread of coronaviruses. In general, this method works by identifying and categorizing causes and effects. However, there are some problems with this approach. The algorithm struggles to recognize and understand even minor causal relationships. Some researchers have proposed a solution. The Montreal Institute for Learning Algorithms at the Max Planck Institute for Intelligent Systems and Google Research have developed an algorithm for this purpose (Pandya, 2020).

In general, Causal inference attempts to estimate the effect of a treatment on a specific outcome. Among other things, it seeks to resolve the confounding variables that exist between the two types of data. This is a form of correlation, a type of statistical analysis. For example, an intervention can increase or decrease the risk of disease. In addition, a causal relationship may also be found between the two conditions. A causal relationship is the result of a series of events. If a treatment causes a particular outcome, then it will be responsible for the other. For example, a medical researcher may see a positive correlation between two treatments. A person can conclude that a negative correlation is a cause of a disease. A scientific study may also suggest the existence of several other factors that could have caused the problem. When it comes to the causal relationship, a study can be said to be causal if one of the two is unrelated to the original condition (Lecca, 2021).

While correlations do not imply causation, causal inference is an important aspect of research. In many cases, a causal relationship is one of the most important factors in a research study. For instance, if a person is listening to music, the effect of this music is a cause. So, what is a causal relationship? This is a complex question that is often difficult to answer. In science, the causal relationship between two events is an important factor. This type of relationship is often called a "case-effect" and occurs when one object is caused by another. This is an example of a causal relation. If the two conditions are mutually exclusive, the corresponding condition is the cause. The difference between the two causes is the hypothesis. The underlying mechanism. If one factor causes the other, it is the condition of the other.

The first step is to motivate the topic using real-world examples, including the causal issues that arise in the fields of education, marketing healthcare political science, as well as online advertisements. Then specific definitions, as well as implications of a variety of crucial concepts of causal reasoning will be discussed such as counterfactuals, the average effect of treatment (ATE), etc. Principal approaches to causal inference, which include the study of experimental data as well as observational research, will be discussed briefly. The focus will be on the outcome framework that could be used in an observational study and provide definitions of the problem along with the essential assumptions (Keane Lucas, 2020).

# **Data**

In the United States, more than two million infants are born prematurely every year. These low birth weight premature babies have an uncertain future and are at risk of a number of health problems, including epilepsy, diabetes, and a host of other health problems. To help prevent these premature births, the Infant Healthy and Development Program (IHDP) (Anon., n.d.) has been created. This program works to provide a better start for every baby. The study is based on the results of a collaborative multisite randomized trial that examined the efficacy of a comprehensive early intervention for infants. It included care and services from the hospital discharge until 36 months of corrected age. The program was implemented in eight sites between 1985 and 1988 and recruited a sample of 2,000 to 3,000 infants. The study was conducted in English and Spanish and involved home visits and child development centers (Anon., n.d.).

The IHDP Dataset is a large dataset that studies infant development. It contains six continuous 19-binary covariates and an unbalanced treatment/control group. The IHDP data can be used to measure the effects of specialist home visits on cognitive test scores. The IHDP is used to compare the effects of a treatment on children's health outcomes, as well as to evaluate the effectiveness of the treatment. The IHDP Dataset includes data from various research trials, including the Infant Health and Development Program. The IHDP (Hill, 2011) was designed to provide quality childcare for prenatal children. Researchers hoped that early intervention would result in improvements in children's cognitive test scores. The IHDP data are available in the public domain. This dataset is freely available through the National Institutes of Health. However, it is not a good choice for large-scale evaluation of interventions.

This dataset is based on real-world data, which has a high level of statistical precision. It contains 4802 observations, 58 covariates, and 77 datasets. The ACIC data set includes different treatment selection functions, with 100 replications of each dataset. The IHDP dataset incorporates non-linearity and the magnitude of treatment outcomes. It also uses an error correction method called Population CE. This approach is especially useful when using large-scale studies. The IHDP Dataset is a unique database of a large number of variables. One of these is the average weight of the child. This is important because a large number of children in one study can have a wide range of birth weights. The IHDP Dataset has been designed to be as inclusive as possible (Shalit, 2016). This means that the IHDP dataset is a great source for testing the effects of different treatments on children's brains.

Moreover, the IHDP data set includes several variables that can help researchers test their theories. The ACIC dataset is derived from real-world data and contains 4802 observations, 58 covariates, and 77 datasets. Each dataset is created with 100 random replications, and a sample size of a child is used to estimate the effect of a treatment. The IHDP Dataset is also useful for analyzing outcomes.

# **Methodology**

In this part, we will first introduce the EconML Machine learning library and describe the reason why we used this library. After that, we will brief out the Data Processing and cleaning process to execute the whole system.

## **EconML**

EconML (Vasilis Syrgkanis, 2021) can be described as a Python program that harnesses the capabilities of machine learning to calculate individual causal responses from data collected through observation or experimentation. The set of estimation techniques included in EconML is the most recent advancements of causal machine-learning. Through the incorporation of individual machine learning processes into interpretable causal models these techniques increase the accuracy of predictions based on what-if and help make causal analysis faster and easier for a wider range of users (Anon., n.d.).

## **CATE Estimators**

We look at a parameter that is functional called the conditional mean treatment impact (CATE) which is intended to quantify the diversity of a treatment effect over subpopulations, when the assumption of unconfoundedness is applied (Jason Abrevaya†, 2012). Contrary to quantile regressions the subpopulations of concern are identified in terms of potential values of covariates that are continuous instead of the quantiles of possible outcomes distributions. We demonstrate that CATE can be not parametrically defined under confoundedness and suggest inverse probability weighted estimation methods for it. In regularity conditions that are conventional and some new to our literature. We demonstrate (pointwise) the consistency as well as an asymptotic normality for the fully nonparametric as well as a semiparametric estimation tool (Alicia Curth, 2022).

## **StandardScaler**

The StandardScaler function transforms columns to mean 0 and unit variance. It is similar to the R-base function scale, which calculates standard deviation with a single degree of freedom. Mahout is a better option for large data sets, as it does not have parameters. Moreover, it may confuse users with its discrepancy in standard deviation (Anon., n.d.). The StandardScaler requires no parameters, so it is recommended for large data sets. When a feature is normalized, the StandardScaler resizes its distribution so that the mean and standard deviation are the same. The standard deviation of the distribution is a number between -1 and 1. This resizes the feature's range to a unit variance and a unit-standard-summary. The difference between normal and outlier-prone data depends on the type of normalization. The best method to use in this case is a minimax-based linear regression (Anon., n.d.).

## **Data Processing & Cleaning**

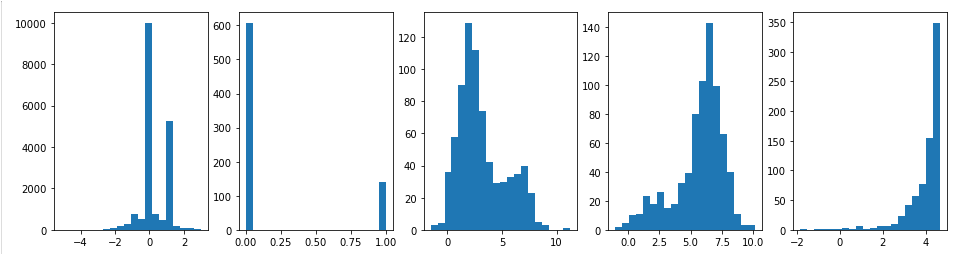
Data processing and cleaning is an important step in any machine learning project. This process is the first step in creating a machine learning model. Clean data will yield better models. For example, the Pandas function can be used to replace specific words in a dataset. Moreover, it will be possible to detect errors in the data before it is processed. This process will allow for a more accurate classification of the data. During the data preparation phase, the dataset should be cleaned (Ilyas, 2016).

The second step in data processing is data cleaning. The task is time-consuming, but it is crucial to the success of a machine learning project. The quality of data is crucial for a successful algorithm. Furthermore, the quality of the data must be maintained even after adding new data. There are a number of methods to clean the data, but the first two stages are critical. To get a clear idea of the process of data cleaning, consider some best practices.

Statistical techniques such as regression and cluster analysis are also used to identify errors. In the case of machine learning, the statistical tools are used to correct errors. The human intervention is required to detect errors in trained data (Venkat N Gudivada, 2017). A systematic approach to data processing and cleaning is a good starting point. The first step is identifying and removing unwanted observations. The second step is identifying anomalies in the data. In both steps, the objective is to improve the quality of machine learning data.

# **Implementation & Result**

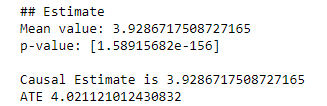
Here we implemented the causal inference on the provided dataset of IHDP and also we derived the actual outcomes and counter factual values based on the given data and determine the causal estimate value. First is the graphical representation of data is displayed below:

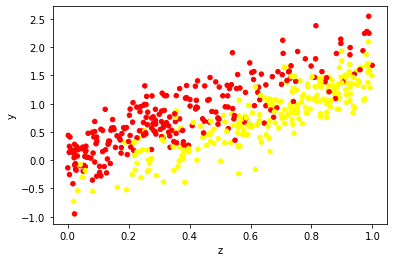


**Fig**. Graphical Representation of Data

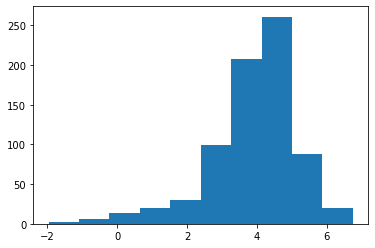
First we determine the standard error from given dataset and After that we find out the P-value and counter factual values using Linear Regression approach.







**Fig.** Estimated Data on Scatter plot



**Fig**. Histogram of CATE Estimator

# **Conclusion**

The purpose of this project was to create ML-based methods and theories to draw causal inferences from data collected through observation and then apply them to the real-time healthcare system (Peng Cui, 2020). In particular, we will devise the following: a) theories and methods for learning causally reliable representations of observations and optimal strategies for decision-making in sequential order and in addition to) strategies (Megan S. Schuler, 2017) and theories to discover multiple pathways the effects of exposures on outcomes from observations and in addition to) methods that integrate observations and experiments to verify and analyze.

This shift in the methodology from causal machine learning to supervised is fundamentally different in algorithmic innovation beyond the choice of an appropriate model class. We will then apply our methodologies to a real-world assessment.

# **References**

1. Alicia Curth, D. S. J. W. M. v. d. S., 2022. *Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation.* s.l., s.n.
2. Anon., n.d. [Online]   
   Available at: https://bookdown.org/phamtrongthang123/notebookCEVAE/exp.html
3. Anon., n.d. [Online]   
   Available at: https://towardsdatascience.com/causal-machine-learning-for-econometrics-causal-forests-5ab3aec825a7
4. Anon., n.d. [Online]   
   Available at: https://www.journaldev.com/45025/standardscaler-function-in-python
5. Anon., n.d. [Online]   
   Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
6. Anon., n.d. *Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight, Premature Infants in the United States, 1985-1988.* [Online]   
   Available at: https://www.icpsr.umich.edu/web/HMCA/studies/9795
7. Crown, W. H., 2019. *Real-World Evidence, Causal Inference, and Machine Learning,* s.l.: s.n.
8. Hill, J. L., 2011. Bayesian Nonparametric Modeling for Causal Inference.
9. Ilyas, I. F., 2016. *Effective Data Cleaning with Continuous Evaluation.* s.l., s.n.
10. Jason Abrevaya†, Y.-C. H. R. P. L., 2012. *Estimating Conditional Average Treatment Effects,* s.l.: s.n.
11. Keane Lucas, B. H. I. S., 2020. Causal Inference.
12. Lecca, P., 2021. *hine Learning for Causal Inference in Biological Networks: Perspectives of This Challenge.* s.l., s.n.
13. Megan S. Schuler, S. R., 2017. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies.
14. Pandya, R., 2020. *From how to why: An overview of causal inference in machine learning.* [Online]   
    Available at: https://www.ericsson.com/en/blog/2020/2/causal-inference-machine-learning
15. Peng Cui, Z. S. S. L. L. Y. Y. L. Z. C. a. J. G., 2020. Causal Inference Meets Machine Learning.
16. Shalit, U. F. D. J. a. D. S., 2016. *Estimating Individual Treatment Effect: Generalization Bounds and Algorithms..* [Online]   
    Available at: http://arxiv.org/abs/1606.03976
17. Vasilis Syrgkanis, G. L. M. O. M. H. K. B. E. D. J. P. Y. W. P. L., 2021. *Causal Inference and Machine Learning in Practice with EconML andCausalML: Industrial Use Cases at Microsoft, TripAdvisor, Uber,* s.l.: s.n.
18. Venkat N Gudivada, J. D., 2017. *Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations,* s.l.: s.n.