
**An-Najah National University**
**Faculty of Graduate studies**

# EXPLORING THE CAPACITY AND PERFORMANCE OF SUPERVISED LEARNING METHODS FOR LABEL CLASSIFICATION IN CAUSAL INFERENCE: A COMPARATIVE STUDY

**By**
**Ola Mohammad Abu Saqer**

**Supervisor**

**Dr. Abdelrahman EID**

**This Thesis is submitted in Partial Fulfillment of the Requirements for the Degree of Master of Mathematics, Faculty of Graduate Studies.**
**An-Najah National University, Nablus, Palestine.**

**2024**

# EXPLORING THE CAPACITY AND PERFORMANCE OF SUPERVISED LEARNING METHODS FOR LABEL CLASSIFICATION IN CAUSAL INFERENCE: A COMPARATIVE STUDY

**By**
**Ola Mohammad Abu Saqer**

This Thesis was Defended Successfully on 8 / 8 /2024 and approved by:

| | |
|---|---|
| Dr. Abdelrahman EID<br>Supervisor |  <br>Signature |
| Prof. Saed Mallak<br>External Examiner |  <br>Signature |
| Dr. Ahmed Awad<br>Internal Examiner |  <br>Signature |

## Dedication

"I dedicate this thesis to:

my grandfather: Abdelkhalq, his soul rest in peace.

my parents: Mohammed & Tharwat.

my brothers: Abdelrahman, Abdelazez, Abdelsalam, Abdelqader and Abdelmalek.

my two beautiful little nephews: Mariam and Ali."

# Acknowledgement

First of all, I would like to express my gratitude to the Almighty for guiding me to this work and for giving me the energy and patience to complete this journey.

None of this work would have been possible without the unwavering guidance and insightful feedback of my supervisor, Dr. Abdelrahman EID. His mentorship has been invaluable, going beyond the traditional role of a supervisor to not only significantly enhance my academic skills, but also instil a strong sense of professionalism and dedication. Dr Abdelrahman's calm demeanour, coupled with words of encouragement and genuine enthusiasm for progress, has kept me motivated throughout this journey. I feel fortunate to have had such a supportive and inspiring supervisor. Most importantly, when I think about the kind of scientist I want to be, I know I want to be just like you - thank you.

I would like to thank Dr. Zaid Hattab for his invaluable comments, discussions and suggestions throughout the completion of chapter three of this thesis, and I would like to thank Dr. Ahmed Awad and Professor Saed Mallak, for taking the time to participate in the committee discussing my thesis. Their comments were extremely valuable and influential.

I owe my worthiness to the goodness instilled in me by my parents. I express my deepest gratitude to my beloved father and mother, whose support has been as vast as the sands of the earth, your guidance has been a guiding light in my darkest moments, illuminating my path like beacons in the night. From my earliest memories to the present day, your unwavering presence has been a source of strength and motivation. The love you have shown me has been the most powerful. Thank you for being the pillars of strength in my life. I am who I am because of you and I will always strive to make you proud.

To my brothers: Abdelrhamn, thank you for always being there to listen and offer words of encouragement whenever I needed them, you have always been my role model for success and perseverance.

Abdelazeez, distance may have separated us physically, but your presence in my life has always felt close and comforting.

Abelsalam, thank you for always being willing to lend a helping hand whenever I needed it.

Abdelqader, my close friend, thank you for always being my source of laughter even when I was in challenges.

And to Abdelmalek, despite being the youngest among us, your boundless energy and motivation have been a constant source of inspiration.

Ola Abu Saqer
August 8, 2024

## Declaration

I, the undersigned, declare that I submitted the thesis entitled:

# EXPLORING THE CAPACITY AND PERFORMANCE OF SUPERVISED LEARNING METHODS FOR LABEL CLASSIFICATION IN CAUSAL INFERENCE: A COMPARATIVE STUDY

I declare that the work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted elsewhere for any other degree or qualification.

**Student's Name:** Ola Mohammad Lutfi Abu Saqer

**Signature:** Ola Abu Saqer

**Date:** August 8, 2024

# List of Contents

# List of Tables

# List of Figures

# List of Appendices

**EXPLORING THE CAPACITY AND PERFORMANCE OF SUPERVISED LEARNING METHODS FOR LABEL CLASSIFICATION IN CAUSAL INFERENCE: A COMPARATIVE STUDY**

**By**

**Ola Mohammad Lutfi Abu Saqer**

**Supervisor**

**Dr. Abdelrahman EID**

## Abstract

In fact, discussions about machine learning are increasingly prevalent due to its accuracy in prediction and its ability to handle vast amounts of data. Furthermore, many relationships in life are causal, which motivates the efforts to comprehend the cause-and-effect relationships among variables. For instance, understanding the extent of the effect of a particular medicine on an individual with an illness becomes crucial. While it might seem straightforward at first glance, a deeper examination tell the complexity inherent in such endeavors when using machine learning in causality. Machine learning methods have made a valuable contribution to the field of causal inference because unlike traditional approaches, machine learning methods offer greater flexibility in estimating causal effects, since machine learning techniques do not require modelling hypotheses, yet there is still a research in estimation causal effect when both treatment and outcome are binary variables, because machine learning has proven its ability to predict, and prediction does not mean causality. Perhaps this is the challenge for machine learning in obtaining more accurate and less biased estimates of causal effects.

This study conducts a comparative analysis of supervised learning methods for label classification in causal inference. We evaluate the performance and capacity of four techniques: Causal Forest (CF), Support Vector Machine (SVM), Generalized Linear Models (GLM), and Linear Probability Models (LPM) in estimating the causal effects for categorical response variable. In a randomized controlled trial simulation and real experiments were performed to evaluate the methods' performance under varying conditions, by

changing the main characteristics of the data including the sample size, and the number of the explanatory variables.

We have focused on these four methods because of their specific advantages: Causal Forests are particularly adept at making causal inferences easily; Support Vector Machines are recognised for their effectiveness in binary classification tasks; Generalised Linear Models are well established as optimal for modelling the binary response variable; and Linear Probability Models are used for their ability to provide predictions as probabilities.

The results provide valuable insights into the strengths and limitations of each method in each scenario in the causal effects simulation study. Furthermore, the methods are able to detect heterogeneity in the real data results, and it was expected that SVM, GLM and LPM would detect more heterogeneity than Causal Forest. This thesis helps us to improve our knowledge of machine learning techniques in causal inference and emphasizes the importance of carefully evaluating their performance in real-world applications.

**Keywords**: Causal Forest, LPM, GLM, SVM, Causal effect.

## Introduction

Causality is a fascinating topic of research, but after thousands of years of work on causality, causality is still an unsolved problem. We study causality because we need to make sense of data, to guide action and policy, and to understand how and why causes influence their effects (Pearl et al., 2015). It's hard to think of a field where there is no interest in or need for causes (Kleinberg, 2015). Using causes responsibly involves assessing not just the statistical and methodological validity of results, but also considering their ethical foundation and repercussions.

Causal inference is a valuable tool for uncovering important insights about how policies, interventions, or treatments affect outcomes (Bojinov et al., 2020). It can help us make sense of complex processes by analysing different scenarios and determining the most effective approach. By predicting how a specific intervention may influence outcomes, we can make informed decisions based on those predictions. This is a key focus of many research studies. Additionally, causal inference methods can identify which groups are more likely to benefit from a particular intervention or treatment.

In this thesis, we focus on machine learning for causal effect estimation, a class of classification and regression techniques. A model is built and used to predict outcomes for each treatment and control group. In particular, we focus on supervised machine learning algorithms for causal inference. This field is about more than just making predictions, it is about building models by understanding the causal relationship between variables.

Investigating causality using machine learning algorithms helps us answer two important questions for a given dataset: What is the effect of changing one of the variables (treatment) on the outcome? and, more interestingly, how can this effect (if any) be modified when more variables are considered?

Due to the high accuracy of such methods compared to traditional statistical methods and the great emphasis on estimating the Average treatment effect (ATE), there was a need to develop these methods to answer causal questions. The strength of these methods is their ability to estimate the conditional causal effect, which allows us to make decisions about

subgroups of society.

The thesis is divided into four chapters, with the first chapter is a general introduction to causal inference, supervised machine learning methods, traditional methods for estimating causal influence, in addition to the challenges and problems in the use of supervised machine learning in causal influence, which led us to have prompted us to find appropriate solutions to these challenges, then we review the literature on the methods used to estimate causal inference, as this section provides an overview of the different methodologies that have been developed and used in the field of causal inference and supervised machine learning.

In **chapter 2**, we first examine the use of randomized controlled trials (RCTs). We then present the assumptions that have been formulated to use the Neyman-Rubin causal model to determine causality, and we present the methodology used in each of the four methods to estimate the treatment effect which are Causal Forest (CF), Support Vector Machine (SVM), Generalized Linear Models (GLM), and Linear Probability Models (LPM).

In **chapter 3**, we present the results of simulation studies and apply the methods to a real data set. This chapter is divided into two main sections in order to demonstrate the effectiveness and practical application of the causal inference methods discussed in the previous chapters. First, we present the results of simulation studies. Then we apply these methods to a real data set to demonstrate their practical use.

**chapter 4** is dedicated to the conclusion of this thesis and summary of the results we found during the estimation of the causal effect using machine learning.

# Chapter One
# Preliminaries

This introductory chapter provides an overview of the main concepts and techniques used in subsequent chapters, as well as presenting some results of previous studies related to the research problem in this thesis.

## 1.1 Causal Inference

### 1.1.1 Correlation and Causality

Many studies involve multiple response variables, and the dependencies between them are often of great interest (Altman and Krzywinski, 2015). When we want to describe the relationship between two variables, both correlation and causation can do this, but there is certainly a difference between them in terms of direction (Zhao, 2018).

When describing research results, Gershman and Ullman (2023) confirmed that scientists are careful not to confuse causation and correlation, because causation implies correlation, but correlation does not imply causation (Liang and Yang, 2021). We say that two variables are correlated when they show an increasing or decreasing trend (Altman and Krzywinski, 2015), in other words correlation measures how strongly two variables are related. If we know one of them, we can predict the other if they are strongly correlated. There are two main measures of correlation: Pearson's correlation, which measures a linear relationship, and Spearman's correlation, which measures a non-linear relationship. However, causality exists when one variable causes another to change, so the second variable (the effect) is understood to be a consequence of the first variable (the cause) (Dubitzky et al., 2013).

Knowing all the ways in which we can find correlation without causation is important because it helps us to critically evaluate our findings and assumptions, and can prevent ineffective interventions.

### 1.1.2 Definition of Causal Inference

Over the years, there has been an emphasis on three research tasks in data science, namely description, prediction and causal inference. Causal inference has recently become a very popular area because it helps us understand what happened in the past and predict what will happen in the future. People need to answer such questions. "Increasing study hours helps students improve their academic performance", "which leads to high blood sugar levels". Although causes are difficult to define and find, there are three main things that can either only be done with causes, or can be done most successfully with causes: prediction, explanation and intervention.

The term "cause" is commonly defined as a factor that increases the likelihood of an effect, without which the effect would not be possible, or as a factor that has the potential to produce an effect under the right conditions (Kleinberg, 2015).

In the classical definition, causal inference refers to a statistical model for determining causal effects from data by estimating the effect of a cause by comparing the outcome when a unit is exposed to a particular treatment with the outcome when it is not exposed to that treatment.

Causal inference problems typically focus on estimating two types of treatment effect: the individual treatment effect (ITE) and the average treatment effect (ATE), defined as the difference between the population average potential outcome if all units received the treatment $(E[Y(1)])$ and the average potential outcome if all units did not receive the treatment $(E[Y(0)])$.

$$\tau = E[Y(1) - Y(0)] \tag{1.1}$$

(Parikh et al., 2022).

Knowing that a treatment is effective for the population 'on average' does not mean that it is effective for a particular individual, so it may be better to infer the individual treatment effect (ITE). Chernozhukov et al. (2023) defines the difference in outcome for an

individual with and without treatment as the individual treatment effect (ITE).

$$\tau_i = Y_i(1) - Y_i(0) \tag{1.2}$$

### 1.1.3 Importance of Causal Inference

The importance of causal inference appears in many applications in life such as health care, marketing, political science and online advertising. Causal inference aims to explore the causal relationships between variables and to estimate a well-defined causal effect of interest (Hernán and Robins, 2020). The need to understand causality stems from our interest in the meaning of data in the sense that the data cannot express itself, in addition to being aware of how and why causes influence their effects. .

The ability to identify causes is a major advantage in prediction, interpretation and intervention. In prediction, the use of causes is more successful than the use of correlations, and knowledge of causality leads to prediction accuracy and improved decision making. It is important to know why the outcome occurred, but it is even more important to know what interventions affected the outcome, we need to know a particular outcome, or even to prevent an undesirable outcome, so knowing that these interventions do not affect the outcome leads us to reduce the time (Kleinberg, 2015) .

Any researcher interested in causal inference would be able to identify the appropriate population for such a task, identify causal parameters, and conduct sensitivity analyses.

### 1.1.4 Different Causal Frameworks

There are several frameworks for analysing problems of causal inference, i.e. we only observe the actual outcome and never the counterfactual outcomes that might have happened if a different treatment had been chosen. Three common frameworks for formulating causal inference models are: the Neyman-Rubin causal model, structural equation models and the structural causal model. These frameworks allow the researcher to describe the framework of the causal relationship. The first framework is the approach used in this thesis and it is also known as the Rubin causal model, it defines causal effects as comparisons of potential outcomes, structural equation models are the traditional econometric

approach to causal analysis, structural equation models explicitly model the relationships between variables. The structural causal model combines the potential outcome framework, graphical models and structural equation models, so it includes the causal graph and the structural equations (Yao et al., 2021).

The key difference between the above frameworks is that structural equation models and the structural causal model do not require the assumption that selection occurs only on observable variables.

## 1.2 Supervised Machine Learning

### 1.2.1 Machine Learning

Machine learning is a branch of computer science that aims to solve a specific problem by collecting data and then using it to create an algorithmic statistical model without explicit programming, we call this data training data. Machine learning combines two fields: statistics, which is interested in learning and interpreting through data, and computer science, which looks at data from a different angle and puts it to practical use. We can therefore define machine learning as "the science of programming computers to learn from data". The term "learning" refers to using the training data to explore the patterns of that data.

The extraordinary approach of machine learning shines in its ability to solve problems that cannot be solved by traditional methods because there is no algorithm for the problem or the problem contains a large number of data; moreover, the power of machine learning lies in its ability to deal with multidimensional data. It is interesting to know that the first application of machine learning is the spam filter, which helps people to improve their lives.

Special attention was given to data mining as an important concept of machine learning, because our lives produce a large amount of data every day and we want to understand and analyse this data. Therefore, we can define data mining as the extraction of knowledge from data.

Machine learning is divided into three types depending on the outcome:

- Supervised Learning: where the labeled input is known.
- Unsupervised Learning: where the labeled input is not known.
- Semi-Supervised Learning: where we have a mixture of labelled and unlabelled input.

### 1.2.2 Overview about Supervised Machine learning

Supervised machine learning is the most common type of machine learning and performs two types of tasks: classification and regression. The peculiarity of this type is that the training data entered into the model or algorithm has a specific and known output (i.e. labelled examples), which is set by the researcher. To further illustrate the idea, the training data (input, output) can be represented in the form of this collection $\{(x_i, y_i)\}_{i=1}^{N}$, where $x_i$ is called the feature vector and $y_i$ are the labels, which can be a finite set of classes or a set of real numbers or other complex types such as matrices.

Collect data so that each example is categorised for one of the labels, this would be the starting point for the supervised learning work. For example: Suppose you have 1,000 students, each categorised as (pass,fail), then convert these examples (students) into a feature vector. By training data we get a model using a learning algorithm, and then we can use this model to predict the outcome of any input. We need to be aware that algorithms in these areas need to convert labels into numbers, such as a support vector machine that gives the label pass value (+1) and the label fail value (0).

There are a variety of methods in supervised learning which are :

- k-Nearest Neighbors.
- Linear Regression.
- Logistic Regression.
- Support Vector Machines (SVM).
- Decision Trees and Random Forests.
- Neural Networks.

### 1.2.3 Types of Supervised Machine Learning Problems

There are two groups of supervised machine learning problems: classification and regression; the outputs in classification are categories, whereas in regression they are continuous.

There are two main types of classifications

- Binary classification : In this type, the number of classes is two.
- Multi-class classification : In this type, the number of classes is more than two.

Sometimes we need to make classifications in medical fields, for example, classifying patients according to the type of treatment they are receiving. This could be treatment (A), treatment (B) or treatment (C). Or in banking, to determine whether a credit card is fraudulent or legitimate.

Classification can be seen as a two-stage process, the first stage is called learning, which involves building the model by training the data using the classification algorithm, and the second stage is checking the accuracy of the model for use on a new sample of data.

Regression analysis is used to determine the correlation between two or more variables and to make predictions using relationships. The main aim of regression is to construct an efficient model to predict the dependent variable from a set of independent variables.

The types of regression techniques that are explained by Kadam et al. (2020) are:

- Simple linear regression: the independent variable has a linear relationship with the dependent variable.
- Multiple linear regression: more than one independent variable has a linear relationship with one dependent variable.
- Polynomial Regression: we transform the original features into polynomial features and then perform a regression.
- Support Vector Regression: this type identifies a hyperplane with maximum margin and it is similar to the support vector machine classification algorithm.

- Decision Tree Regression: splitting nodes by reducing the standard deviation.
- Random Forest Regression: ensemble the predictions of several decision tree regressions.

## 1.3 Causal Inference with Supervised Machine Learning

### 1.3.1 Appearance of Machine Learning in Causal Inference

A variety of methods can be used to reliably estimate causal inference, but there is great development and promising results for the emergence of machine learning in causal inference because machine learning is able to make a clear and interpretable prediction in the area of causal inference. We also know that regression requires knowledge of the real data generation mechanism, and this is not required when we use machine learning, and machine learning techniques do not require modelling hypotheses.

It is clear that using machine learning to estimate a causal effect has an important advantage, namely that there is an additional explanation that is not found in traditional causal inference tools such as nearest neighbour matching, propensity score matching, which Cui et al. (2020) talked about. In addition, Collischon (2023) provided an overview of instrumental variable (IV) regression, difference-in-differences (DiD), and regression discontinuity design (RDD). Finally, Edwards et al. (2016) presented inverse probability weighting.

It appears that when a researcher aims to estimate the average treatment effect (ATE) from an observational study, the use of machine learning methods allows for better control of confounding variables, leading to more accurate causal inference. Furthermore, for researchers interested in estimating heterogeneous treatment effects, machine learning proves effective in capturing heterogeneity, where heterogeneity occurs when there are interactions between treatment and other covariates, as opposed to cases of homogeneity (Hoogland et al., 2021).

One of the models that combines causal inference and machine learning is uplift modelling. It is considered a causal inference problem and a machine learning problem because the researcher needs to estimate the difference between two different outcomes for

the individual, in terms of machine learning we need more than one model, train them and then choose the model that gives us more accuracy. Uplift modelling is a set of methods used to estimate the causal effect of a treatment based on the Conditional Average Treatment Effect (CATE):

$$\tau(x_i) = E[Y_i(1)|x_i] - E[Y_i(0)|x_i] \tag{1.3}$$

Where researchers are typically interested in estimating this type, these models determine when the treatment is positive, negative or neutral for the individual.

Recently, novel machine learning methods have been applied to the field of causal inference, pushing it to new frontiers. Some methods use the power of machine learning to characterise all the features in the data that influence an outcome of interest (Crown, 2019). While most studies focus on estimating the causal effect associated with discrete variables, particular attention has been paid to an algorithm-free method for estimating the causal effect of continuous variables. This method allows the use of any machine learning algorithm and aims to minimize assumptions on the training data (Kitazawa, 2022). Finally, a particular advantage of machine learning is that we need it to meet the challenge of describing parametric regressions in a true way, since the information for this is incomplete. (Balzer and Petersen, 2021).

Dorie et al. (2022) discuss the problem of missing data that arises from not knowing the potential outcome for individuals if they receive treatment or not at the same time, where many causal inference strategies deal with this problem using machine learning algorithms.

### 1.3.2 Challenges in Using Machine Learning in Causal Inference

Even though machine learning focuses on classification and regression tasks and dose them better than humans, it's a smart idea to use it for causal inference. However, the challenge here is to create a machine learning model that will helps humans see the relationship between cause and effect.

The main objective of this thesis is to provide a comparison between causal machine learning methods and traditional methods using simulated datasets that generate different scenarios small/large sample sizes, and effective/non-effective confounding variables. then apply the algorithm to a real dataset and assess its consistency.

## 1.4 Literature Review

Several research efforts have attempted to address the challenges of causality detection using machine learning. These studies have approached the topic from different perspectives, using different methodologies and techniques.

One notable study by Yu et al. (2016) used support vector machine (SVM) methodology to classify drug-related tweets and their cause-and-effect relationships. Their work focused on using SVM to effectively identify causal relationships in the context of social media data, and the results showed that these classifiers achieved up to 77% accuracy in identifying the cause-effect relationships of drugs in Twitter data.

In another paper, Ratkovic (2014) highlighted the importance of incorporating non-deterministic treatment rules when estimating causal effects using SVM. To address this, the author proposed the use of a modifying hinge loss function within SVM, which provides a framework for estimating causal effects in situations where treatment rules are not predetermined.

Tarr and Imai (2021) focused on situations where there is a desire for approximately equal numbers of treated and control units. They used SVM to estimate causal effects by fitting dual coefficients to the soft-margin SVM classifier as kernel balancing weights. Bayesian adjustment for confounding (BAC) was used to estimate the average causal

effect (ACE) using two sets of generalized linear models (GLMs) - one for exposure and one for outcome.

There are many problems in causal discovery related to smaller sample sizes, focusing on a sparse causal model that does not represent real-world data generation, and claims of causal unrelatedness that have high error rates. Petersen et al. (2022) addressed these problems by proposing a new method based on supervised machine learning (SLdisco), but this method has a crucial limitation in that it does not account for unobserved confounding when considering Gaussian data. The method proposed by Chikahara and Fujino (2018) offers a new approach to causal inference in time series using supervised learning, which uses a classifier instead of regression models in traditional methods. In particular, they solve the problem of determining Granger causality using ternary classification. The paper published in Linden and Yarnold (2016) describes how it is possible to use a machine learning algorithm called optimal discriminant analysis (ODA) in conjunction with GPS-based weighting to estimate treatment effects in multivalued treatment evaluations using permutation P-values that do not require distributional assumptions.

Despite the weakness of the relationship between machine learning and doubly robust estimators, Naimi et al. (2020) proved that using machine learning to estimate the causal effect is associated with using doubly robust estimators, and his inability to use single robust techniques even with machine learning, the idea is that confounding variables in the outcome or treatment model can be neglected in doubly robust estimators. Although sometimes not to the liking of analysts, in certain circumstances the estimate remains unbiased even when machine learning is used in the outcome and treatment models, thus solving the dimensionality problem.

Lee et al. (2020) went on to isolate the effects of heterogeneous treatment in order to be able to use machine learning alongside experimental designs to find the causal effect of leadership. From his point of view, machine learning is used for a deeper understanding of causal inference when combined with experimental designs.

The description of uplift modeling relies heavily on the concept of a causal inference and a machine learning, the work of Gutierrez and Gérardy (2017) gives a fuller discussion

of the concept of uplift modeling, he presents three approaches to this type of modeling, mainly based on the existence of a common framework for causal inference, these approaches are two-model approach, the class transformation approach and modeling uplift directly. The proposal that is supported by evidence through which Zaniewicz and Jaroszewicz (2013) use support vector machine in uplift modeling, which is called Uplift Support Vector Machines (USVMs), this method divides the training data into two parts, the treatment group and the control group and models the differences in class behaviour between these sets, a difference of this method is that it allows us to minimize the negative effects of treatment because it identifies the negatively affected group.

This issue has been addressed by the recent work in which the Song et al. (2023) relied on the Augmented Synthetic Control Method (ASCM) Where used as a second approach to the random approach forest based machine learning model using quasi-natural experimental designs to estimate the causal effect of winter heating on air quality in northern China.

Samii et al. (2016) was able to break through important limitations of traditional methods used in retrospective causal inference, namely the lack of clarity in the counterfactual and the presence of a large number of covariates that are ineffective in causal inference, with the help of machine learning and the use of the conditional independence hypothesis. Machine learning is used here to obtain the largest possible number of control variables to determine causality.

In a realistic scenario, Le Borgne et al. (2021) concluded that machine learning methods, in particular SVM, give a good result in estimating the causal effect of binary exposure statuses on binary outcomes, provided there is a small sample size, n = 100.

The method introduced by Clarke and Windmeijer (2010) has the advantage of estimating causal effects on binary outcomes with non-compliance using structural mean models (SMMs), but there are some non-standard assumptions required by the author here.

One method for individualized treatment effects is conditional generative adversarial networks for inference of individualized treatment effects (GANITE), which can deal with

binary treatment. Ge et al. (2020) developed this method, a modification of the Conditional Generative Adversarial Network, which showed superiority over Random Forest Classification [RF (C)], Random Forest Regression [RF(R)], and Support Vector Machines (SVM) to estimate individual effects of all types of treatments, including binary, categorical and continue treatment.

Several authors have acknowledged the importance of using random forests in causal inference, one of them Venkatasubramaniam et al. (2022), who used penalised regression along with causal forest to make a comparison between them, when he found that regression outperformed causal forests, he emphasised that it is necessary to have regression when detecting treatment effects and not to rely on causal forest algorithms or other similar machine learning algorithms alone. Hattab et al. (2024) did not compare causal forests with other methods, and determined the heterogeneous causal effect of a continuous response variable using honest inference for treatment effects.

Imai and Ratkovic (2013) considered the determination of treatment effect heterogeneity as a variable selection problem using a support vector machine (SVM), with separate constraints on the pretreatment and causal heterogeneity parameters of interest.

However, despite the above theoretical inferences, our main aim is to compare the strength of the four methods mentioned in terms of causal effect. Therefore, in a series of studies we want to explore the valid way to use these methods.

# Chapter Two

# Methodology

## 2.1 Randomized control trials

The randomized controlled trial is a simple but highly effective research method. Essentially, it involves randomly assigning people to different interventions for the purposes of the study.

The strength of a randomized controlled trial lies in the randomization process. Randomization ensures that each person has an equal chance of being assigned to one of the groups. This helps to make the characteristics of the participants in each group as similar as possible at the start of the comparison (Stolberg et al., 2004). The importance of Randomized Control Trials (RCTs) lies in their ability to effectively manage the confounding variables that influence outcome and treatment. Randomized groups in RCTs typically have a balanced or equal distribution of confounding factors (Stoltzfus, 2011).

An RCT is a common research method used in experiments to test the effectiveness of interventions. It is particularly useful for determining whether there is a cause and effect relationship between a treatment and an outcome. The results are high-quality evidence, unlike observational data, which can be biased. To remove bias, people should be randomly assigned to treatments. Differences between groups in outcomes may be due to systematic group differences as well as the treatment, which can be seen in observational studies as Bhide et al. (2018) explained. So Randomized controlled trials (RCTs) are considered the best way to measure causal effects because randomization helps reduce bias.

Our data meet the concept of RCTs if we find no significant differences in the characteristics between two groups using the p-value of the t-test. This is shown in Table 7 in Chapter 3.

## 2.2 Neyman-Rubin Causal Model

Neyman-Rubin causal modelling, also known as the potential outcomes framework, was originally developed to understand the causal impact of a treatment on a specific outcome (Lara, 2024). Over the past five decades, the potential outcomes framework has gained considerable popularity as a widely used approach to defining, identifying and estimating causal effects (Keller and Branson, 2024).

This framework has received considerable attention in various fields, including statistics, medicine, economics, and political science, as highlighted by (Sekhon, 2008). In particular, Pearl (2007) showed that by using this model, statisticians have achieved a definitive understanding of causality, which is a remarkable advance in the field.

All authors agree that this framework has two fundamental components: the first is to determine causality by potential outcomes, while the second is to make assumptions about the "assignment mechanism", which refers to how subjects receive the interventions under study.

Let D represent a binary treatment status, such that D=0 indicates no treatment and D=1 indicates treatment, and from here the unit follows either the first group, which is the control group, or the second group, which is the treatment group, in a random manner, where random assignment of treatment greatly simplifies the process of identifying and estimating causal effects.

Let $Y_i$ be the observed outcome, so that each unit has two potential outcomes $Y_i(0)$ and $Y_i(1)$, one if the unit is untreated and the other if it is treated, but only one of the two potential outcomes is observed, so causal inference is a missing data problem because half of the potential outcomes are missing, this problem is referred to as the fundamental problem of causal inference.

In mathematical terms, we can define the observed outcome as

$$Y_i = D_i Y_i(1) + (D_i - 1) Y_i(0) \tag{2.1}$$

Specifically, if D=1, then $Y_i = Y_i(1)$, so we cannot determine $Y_i(0)$ by mere observation. In this case $Y_i = Y_i(1)$ is called the factual outcome, while $Y_i = Y_i(0)$ is called the counterfactual outcome.

and so on to the treatment effect:

$$\tau_i = Y_i(1) - Y_i(0) \tag{2.2}$$

These differences are called individual treatment effects. They may be positive for some subjects and negative or zero for others.

Note that the potential outcomes $Y_i(1)$ and $Y_i(0)$ are random variables, and that these potential outcomes, as well as the treatment, may depend on some covariates X, which are other important variables measured prior to the implementation of the intervention. We refer to the random vector (D, X, $Y_i(0), Y_i(1)$) as the Neyman-Rubin causal model.

To make sense of the causal relationship between treatment and outcome in this context, we need to ask a hypothetical question such as "What would the value of Y have been if D had been 1 instead of 0 for a given individual (with X(i) = x)?".

We will focus on estimating the average treatment effect, in particular the conditional average treatment effect treatment effect. The two main quantities of interest are the average treatment effect $\tau$ and the average treatment effect for the treated, which is an example of a conditional average treatment effect (CATE), defined as

$$\tau = E[Y(1)_i - Y(0)_i] = E[Y(1)_i] - E[Y(0)_i] \tag{2.3}$$

$$\tau_T = E[Y(1)_i - Y(0)_i | D = 1] = E[Y(1)_i | D = 1] - E[Y(0)_i | D = 1] \tag{2.4}$$

In a fully randomised experiment, $\tau$ and $\tau_T$ are equivalent because D is randomly assigned and therefore independent of the potential outcomes.

Researchers are very interested in CATEs because they help answer questions such as "What works and for whom? For example, CATEs can be determined by conditioning on

any combination of baseline covariates.

Occasionally, it is observed that the treated unit produces a lower result than the untreated unit. For example, in the field of education, it is often observed that a weak student is more inclined to face additional study hours (treatment). However, it is possible to observe a lower level of academic performance in the treated group compared to the non-treated group due to the student's condition, we would observe $E[Y|D=1] < E[Y|D=0]$ while $E[Y(1)] > E[Y(0)]$ in this situation we call the student's condition a confounders: a variable that affects both the treatment and the outcome.

The framework therefore consists of three basic steps: identifying, recognizing and assessing causal effect

- Define causal estimates of interest, such as conditional or other causal effects, in terms of potential outcomes.
- Identify observable statistical quantities that, under certain assumptions, are equivalent to the causal effect, consistent with the study of the causal effect.
- Estimate the observable quantity in step 2 using machine learning models in this thesis.

Randomized experiments are conducted by the proportion p of units receiving the treatment, the probability being given by $P(D_i = 1)$ = p for all units. Alternatively, units can be divided into two "blocks" based on a covariate variable, and then the treatment is randomized within these blocks. The probability in this case is $e(X_i) = P(D_i = 1|X_i = x)$, where the unit's probability of receiving the treatment depends on its $X_i$ value, $e(X_i)$ is called the propensity score.

### 2.2.1 A assumptions of potential outcome framework

The individualised treatment effect, as defined above, is the difference between two potential outcomes, since we observe one of the potential outcomes for each individual. We have to make some assumptions. The methods will be applied with a number of assumptions: The Stable Unit Treatment Value Assumption (SUTVA), positivity, consistency, unconfoundedness and the no interference assumption.

- **The Stable Unit Treatment Value Assumption (SUTVA)**

  SUTVA is a fundamental assumption commonly utilized in causal inference. This assumption highlights the presence of potential outcomes corresponding to each possible value of the treatment variable. In this thesis, we focused on emphasizing only two potential outcomes. Furthermore, one of the binary outcomes observed, represented as $Y_i \in Y = \{0,1\}$ (Laffers, 2020).

- **Positivity**

  It is a common assumption in randomized controlled trials and states that each participant has a certain probability of receiving each treatment, i.e. $0 < P(D = 1|X = x) < 1$. In other words, $0 < e(X_i) < 1$, where $e(X_i) = P(D = 1|X_i)$.

- **Consistency**

  This assumption state that the observed outcome is the same as the potential outcome for an individual being treated. Put simply, the outcome you observe is exactly what you expect. This notion is valid in the context of the experiments conducted in this research due to the presence of medical treatments, as we can hypothetically know an individual's treatment status. This assumption holds if the treatment effect is well defined, otherwise we cannot clearly define $Y_i(1) - Y_i(0)$ (Cole and Frangakis, 2009).

  Consistency is stated here as equation 2.1 or

$$Y_i^{obs} = Y_i(D) \tag{2.5}$$

  where D is 1 or 0

- **Unconfoundedness**

  Unconfoundedness, or what may be called by other names such as ignorability and causal interchangeability, is important in addressing the core challenge of causal inference; the efficacy of this assumption emerges as a solution to this problem, as it posits that the potential outcomes $Y_i(0)$ and $Y_i(1)$ remain unaffected (independent) by the treatment $(D_i)$ assigned to each person. We can express this as $D_i \perp (Y_i(0), Y_i(1))|X_i$ (Zhao et al., 2017). Randomized experiments guarantee unconfoundedness because randomization ensures that $D_i$ is not influenced by other variables.

An immediate consequence of unconfoundedness is

$$\tau(x) = E[Y_i(\frac{D_i}{e(x)} - \frac{1-D_i}{1-e(x_i)})|X_i = x] \tag{2.6}$$

- **No interference**

  It is important for our study that the effect of treatment on the outcome of individual i is not altered by the exposure or non-exposure of other individuals, i.e. no interference assumption in causal inference. If this assumption is not met, it becomes more difficult to determine the potential outcome for each individual, as it includes not only the specific treatment for the individual, but also the treatment assigned to the rest of the individuals (Goetghebeur et al., 2020).

So, the Conditional Average Treatment Effect (CATE) equation,

$$\begin{aligned}
\tau &= E[Y_i(1)|X = x_i] - E[Y_i(0)|X = x_i] \\
&= E[Y_i(1)|D = 1, X = x_i] - E[Y_i(0)|D = 0, X = x_i] \quad \text{(by exchangeability)} \\
&= E[Y^{obs}|D = 1, X = x_i] - E[Y^{obs}|D = 0, X = x_i] \quad \text{(by Consistency)}
\end{aligned} \tag{2.7}$$

### 2.3 Heterogeneous treatment effect

Predicting the effect of treatment is an important area of research in many fields. It is used to assess the effectiveness of new drugs, to evaluate educational programmers and to apply social policies. Each individual may respond differently to treatment - positively, negatively or not at all. This means that there is variation in how interventions affect people, which may be due to observed or unobserved covariates, and this type of variation is called heterogeneous treatment effect. If the treatment has the same effect on everyone, regardless of other factors, it is said to have a constant or homogeneous treatment effect. Alternatively, if the treatment has no effect on anyone, it is said to have no effect (Keller and Branson, 2024).

There is a growing interest in various feilds to study heterogeneous treatment effects, for example, patients do not show a homogeneous response to the same treatment or medica-

tion, and researchers are increasingly using supervised learning techniques to study these effects (Athey and Imbens, 2016). However, (Wager and Athey, 2018) explained that there is a concern among researchers when estimating such effects. They often tend to focus on subgroups with significant treatment effects, and only report results for subgroups with extreme effects.

(Bénard and Josse, 2023) defined that the treatment effect $\tau$ is said to be heterogeneous with respect to $X$ if there exist $x, x' \in \mathbb{R}^p$ such that $\tau(x) \neq \tau(x')$.

**Sources of Heterogeneous treatment effect**

Understanding the source of treatment heterogeneity is critical to understanding the outcome. There are two primary sources: heterogeneity in the basis of treatment, which refers to variation in the treatment given to individuals. The second origin arises from different responses to uniform treatments, where the difference lies in the value of tau (Smith, 2022).

Statistically, Greenfield et al. (2007) said that HTE occurs when there is an interaction between unit characteristics and treatment, where these interactions reduce the treatment effect.

**Why Heterogeneous treatment effect important**

Treatment heterogeneity is important for several reasons. First, it provides insights into the mechanisms by which an intervention produces effects. In addition, awareness of treatment heterogeneity encourages researchers to screen participants for the trial, particularly in cases where the causal effect of the treatment is poor. Particularly in the medical field, researchers and clinicians can identify the groups that would benefit the most with the least risk from treatments by understanding HTE, and they can better prepare for the challenges that come with scaling up interventions by gaining a thorough understanding of treatment effect heterogeneity (Asher, 2021).

In order to estimate the conditional average treatment effect at a specific test point x, as stated in equation 2.7, we can make an assumption and then estimate it using the provided

equation.

$$\tau(x) = E[Y^*|X_i = x] \tag{2.8}$$

where $Y^*$ is a transformed outcome defined by Zhao (2018) which matches the equation 2.6.

$$Y^* = Y_i \cdot \frac{D_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \tag{2.9}$$

## 2.4 Logistic Regression

Logistic regression is part of a larger family of generalized linear models. It is a widely used statistical technique in research and is considered an essential tool in various fields including health analysis, medical statistics, ecology, social statistics, econometrics and others. However, while linear regression is often used to analyse continuous outcomes, it is not suitable for binary outcomes. In such cases, logistic regression is more appropriate as it converts the binary outcome into a continuous variable (Stoltzfus, 2011).

A logistic regression models the logarithm of the odd ratio of the probability of the outcome.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{2.10}$$

The probability ranges from zero to one, while the odds range from zero to infinity and the natural logarithm ranges from negative infinity to positive infinity.

Solving 2.10 in terms of p we get

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \tag{2.11}$$

We are going to modify this equation slightly so that we can use it

$$\text{Probability of x} = \frac{1}{1 + e^{-x}} \tag{2.12}$$

The equation 2.12 was influential as we adjusted the binary outcome variable for the models. Therefore, the new Y is defined as the sum of the probability of tau multiplying by D and the probability of Y(0), both derived from 2.12.

The following system of equations describes the relationship between the original variable $Y$ and the transformed variable new $Y$ in simulation studies, taking into account the application of probability transformations to the variables $Y(0)$ and $\tau$:

$$\begin{aligned} Y &= Y(0) + D \cdot \tau \\ \text{new } Y &= \frac{1}{1+e^{-Y(0)}} + D \cdot \frac{1}{1+e^{-\tau}} \end{aligned} \tag{2.13}$$

where:

$$\begin{aligned} \text{probability } Y(0) &= \frac{1}{1+e^{-Y(0)}} \\ \text{probability } \tau &= \frac{1}{1+e^{-\tau}} \end{aligned} \tag{2.14}$$

These equations capture the essence of the transformation process, reflecting the probabilistic nature of $Y(0)$ and $\tau$ in determining the resulting new $Y$.

In direct application to real data, we will use 2.12 and convert Y from binary to continuous by the following equation.

$$\text{new}(Y) = \frac{1}{1+e^{-Y}}. \tag{2.15}$$

We can write the model of heterogeneous treatment effect

$$logit(P(Y^{obs} \,|\, D = d_i, X = x_i)) = \beta_0 + \beta_t d_i + \beta_m^T x_i + \beta_z^T z_i d_i \tag{2.16}$$

where $z_i$ is a subset of $x_i$, $\beta_m$ includes the coefficients for the effects of $x_i$, and $\beta_z$ includes the coefficients for treatment-covariate interactions.

After that $\tau_i$ can be estimated by below equation

$$\tau_i = P(Y^{obs}|D = 1, X = x_i) - P(Y^{obs}|D = 0, X = x_i) \tag{2.17}$$

Logistic regression creates a simple model for predicting the specific treatment effect for each individual by considering their characteristics, which can result in significant differences in treatment effectiveness.

## 2.5 Model of supervised machine learning

The purpose of this section is to provide a detailed account of the research design and methods used in this thesis. We will provide a detailed overview of each Causal Forest (CF), Support Vector Machine (SVM), Generalized Linear Models (GLM), and Linear Probability Models (LPM).

### 2.5.1 Causal Forest

In this subsection, we discuss how to predict the individual treatment effect using the causal forest algorithm. We start with a definition of random forests and end by highlighting the success of causal forests in estimating causal effects.

Random forest is a modern machine learning algorithm introduced by Breiman in 2001. It is widely used for classification and regression tasks and is known for its high classification accuracy (Liu et al., 2012).

It works by averaging the predictions of multiple random decision trees. This method is particularly effective when dealing with datasets where the number of variables exceeds the number of observations. In addition, (Biau and Scornet, 2016) say that random forests differ from most other methods in that they produce a partition of the population estimating to covariates (Athey and Imbens, 2016). It is a powerful method for accounting for diverse causal effects in randomized control, particularly causal forests.

**Causal forest**

A causal forest is a form of causal machine learning, specifically part of generalized random forest (GRF). Its power becomes apparent when dealing with high-dimensional data, i.e. when there are a large number of covariates, because it estimates the effect of the treatment on each individual while taking the covariates into account. In this way, we discover heterogeneity in the effect of the treatment and then identify the characteristics present in the individual that are responsible for this difference (Kim, 2023).

**Honest model**

In an honest model, the choice of model structure and the estimation process are not based on the same information. Although the accuracy of estimation is reduced due to sample splitting, there is an advantage in reducing bias. This approach modifies the traditional classification and regression tree (CART) approach by emphasising the estimation of conditional treatment effects rather than outcome prediction. It also uses different data to construct partitions and estimate leaf effects. Instead of minimizing the sum of squared errors for the predicted outcome (MSE), the focus is on minimizing the expected mean squared error (EMSE) (Athey and Imbens, 2016).

**Constructing causal trees.**

Typical decision trees for classification or regression are quite different from those for estimating the heterogeneous treatment effect, because ordinary decision trees do not succeed in determining causality and their nodes have no causal interpretation.

The work of causal trees is divided into two steps for causal inference, the first being the splitting step and the second being the estimation step. But a single decision tree may not give us accuracy in estimating the heterogeneous causal effect, so we combine a number of causal trees to obtain a causal forest. This is done by taking B subsamples from the given training data and then running each tree on a subset of the data (Younas et al., 2022).

If you add covariant variables X to the model, the observed result in equation 2.1 can be represented by the following equation

$$Y_i = \mu_i(x) + D_i\dot{\tau}_i(X_i) + error \tag{2.18}$$

with conditional mean function

$$E(Y_i|X_i = x) = \mu_i(x) + \tau_i(x)e_i(x) = m_i(x) \tag{2.19}$$

where $ei(x)$ is the propensity score estimated by regressing D on the covariates, $m_i(x)$ is the marginal mean, and $\mu_i$ is the effect of the covariates X on the outcome.

For the development of causal forests, the equation 2.18 becomes

$$(Y_i|X_i = x) = m_i(x) + \tau_i(x)(D_i - e_i(x)) + error \tag{2.20}$$

To estimate the heterogeneous effect, regression forests are used to estimate both $m_i(x)$ and $e_i(x)$. The treatment effects $\tau(x)$ in the model are then determined by the Dandl et al. (2024) equation

$$(Y_i|X_i = x, D_i) = \hat{m}_i(x) + \tau_i(x)(D_i - \hat{e}_i(x)) + error \tag{2.21}$$

with minimizes the locally centered loss function

$$loss = \frac{1}{2}[Y_i - \hat{m}_i(x) - \tau_i(x)(D_i - \hat{e}_i(x)))]^2 \tag{2.22}$$

Finally, the causal effect is estimated by constructing a CART regression tree, where instead of predicting a specific value for one of the two classes, the target variable is a numerical value (Timofeev, 2004). Using "recursive partitioning", we can estimate a value of the conditional mean function $E(Y_i|X_i = x)$ by using any leaves containing the test point x

$$E(Y_i|X_i = x) = \frac{1}{|\{i : X_i \in L(x)\}|} \sum_{i:X_i \in L(x)} Y_i^{obs} \tag{2.23}$$

then, to estimate the treatment effect for each $x \in L$, we have

$$\hat{\tau}_i = \frac{1}{|\{i : D_i = 1, X_i \in L(x)\}|} \sum_{i:X_i \in L(x)} Y_i^{obs} - \frac{1}{|\{i : D_i = 0, X_i \in L(x)\}|} \sum_{i:X_i \in L(x)} Y_i^{obs} \tag{2.24}$$

To construct the causal forest, a set of k causal trees is generated. Rather than looking for the tree that gives the optimal estimate, we get $\hat{\tau}_k(x)$, then the final estimate derived from the causal forest is the mean of all the estimates from the causal trees

$$\hat{\tau}(x) = \frac{\sum_1^k \hat{\tau}_k(x)}{k} \tag{2.25}$$

In simulation studies, we use the causal forest algorithm to implement 1000 trees for each outcome. However, due to the sample size of the real data, we implement 150000 trees.

### 2.5.2 Support Vector Machine

Support vector machine (SVM) is widely recognised as an important supervised machine learning technique, especially in the context of handling large datasets (Suthaharan, 2016). Its remarkable ability to detect and understand data classification patterns has driven its popularity in the field of data science (Pisner and Schnyer, 2020). By selecting an appropriate kernel function and its parameters in a research problem for classification or regression, one can increase the accuracy to either exceed or match the level of accuracy achieved by other classifiers. This makes the approach one of many theoretically sound methods.

SVM has a higher quality than other ML methods, because if linear spearing fails in the feature space (the space where SVM tries to find the best hyperplane), it maps the data to a higher dimensional space and then tries to make a linear separation.

Before we discuss causal effect estimation using SVM estimates, we need to understand some definitions of SVM components, which are defined by

- *Hyperplane*:
  The subspace has a dimension that is one less than the dimension of the feature space, SVM is called ideal SVM when the hyperplane is able to completely separate the data points into two different classes.
- *Support vectors (SVs)*:
  These are the points that are closest to the other class.

- *Margin*:

  The distance between the support vectors in two classes.
- *Maximum Margin Classifier*:

  The optimal separating hyperplane.

Support Vector Machines (SVM) use a data set to train a learning algorithm. The data set consists of positive labels with a numerical value of +1 and negative labels with a numerical value of 0, representing the desired output. To process the input, it is converted into feature vectors, ensuring that the vectors have the same number of inputs. The SVM then interprets each feature vector as a point in a high-dimensional space where the dimensionality matches that of the feature vectors. Finally, the SVM constructs a hyperplane that effectively separates the positive labels from the negative labels. In order to improve the generalization ability and reduce the classification errors of the training data, it is imperative to maximize the margin (Cervantes et al., 2020).

The two-parameter equation of this hyperplane is given by

$$wx - b = 0, \tag{2.26}$$

The first part of this equation can be rewritten by this formula

$$w_1x_1 + \ldots + w_ix_d$$

where the feature vector consists of "d" dimensions and w is a vector perpendicular to the hyperplane, and the second part (b) is a real number representing the bias.

To predict the label $\{0,+1\}$, that's what we called the output, we use the above information.

$$Y = sign(wx - b), \tag{2.27}$$

then by finding the optimal parameters $w^*$ and $b^*$, 2.27 becomes

$$f(x) = sign(w^*x - b^*), \tag{2.28}$$

with two constraints

- if $Y_i = +1$ then, $wx_i - b \geq +1$
- if $Y_i = -1$ then $wx_i - b \leq -1$

which is equivalent to $Y_i(wx_i - b) \geq 1$ (Burkov, 2020).

There are two types of SVMs, linear SVMs and non-linear SVMs.

**Linear Separable Case** (Simple SVM)
Linear Support Vector Machines solve the problem of binary classification of outcome data points by dividing them into two classes or categories in the feature space.

**Non linearly separable case** (Kernel SVM)
Nonlinear SVM aims to find the best hyperplane in a high-dimensional feature space. However, instead of computing inner products in the feature space, which can be costly due to its high dimensionality, nonlinear SVM uses a kernel function in the input space (Al-Mejibli et al., 2020b).

There are many components in SVM that affect the performance of SVM in classification or regression with sufficient accuracy, one of them is selected the kernel. Since the kernel solves non-linear classification, there is no way to choose the type of kernel. It can be said that the choice has the greatest impact on the quality of SVM work (Roman et al., 2021).

The purpose of the kernel function is to transform the non-linear separation into a linear separation by converting the original space into a high-dimensional one, and there are many kernel functions available for this task

- Linear kernel

$$K(x_i, x_j) = x_i \cdot x_j \tag{2.29}$$

- Polynomial kernel

$$K(x_i, x_j) = (x_i \cdot x_j + c)^d \tag{2.30}$$

- Radial basis function kernel

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}),\ \sigma^2 > 0 \tag{2.31}$$

- Sigmoid function kernel

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \tag{2.32}$$

- The Mahalanobis kernel

$$K(x_i, x_j) = \left(-\frac{\delta}{m}(x_i - x_j)^T Q^{-1}(x_i - x_j)\right) \tag{2.33}$$

Soft margin SVM, in general, is formulated as a minimization problem:

Minimize:

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\zeta_i$$

subject to restrictions:

$$y_i(\mathbf{w}\cdot\phi(x_i) + b) \geq 1 - \zeta_i$$

for $i = 1, 2, ..., n$

and the dual form of the SVM minimization is

$$min\ \frac{1}{2}\sum_{ij} y_i \cdot \alpha_i \cdot y_j \cdot \alpha_j \cdot K(x_i, x_j) - \sum_i \alpha_i$$

subject to

$$\sum_{ij} y_i \cdot \alpha_i = 0$$

with

$$0 \leq \alpha_i \leq C$$

**Radial Basis Function (RBF)**

The Gaussian Radial Basis Function (RBF) kernel is the kernel function used in this thesis. SVM with an RBF kernel is usually one of the best classification algorithms for most datasets. It is defined as

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma^2}),\ \sigma^2 > 0 \tag{2.34}$$

where squared euclidean distance between two feature vectors $x_i$ and $x_j$ is

$$d(x_i, x_j)^2 = ||x_i - x_j||^2$$

and $\sigma^2$ is the kernel parameter called the bandwidth or the width of the Gaussian kernel.

The efficiency of the RBF kernel depends on the choice of $\sigma^2$. The larger $\sigma^2$, the more influence $x_j$ will have on $x_i$. More precisely, if $x_j$ is a support vector, a large $\sigma^2$ implies that the class of this support vector will have an influence on deciding the class of the vector $x_i$, even if the distance between them is large. On the other hand, if $\sigma^2$ is small, then the support vector does not have such a widespread influence (Eid and Wicker, 2023).

In our work there is another parameter to consider for this kernel C, C is how much error you want to avoid in classifying the training data (regularisation term or cost factor), lower C values create a larger margin, making the model simpler but with poorer predictions on the training set, while higher C values create a smaller margin and better prediction performance on the training data, the cost factor C accepts values from 0.001 to 1000.

**Support vector regression**

The concept of Support Vector Machines (SVMs) originated from binary classification tasks and has since evolved to include Support Vector Regression (SVR) for predicting numerical values. The basic idea is the same as classification, but instead of dividing the data into classes, we fit a regression function (SVR). In SVR, the $\varepsilon$-insensitive tube serves a similar purpose to the margin in SVM classification, indicating acceptable deviations in numerical value predictions. Data points that fall outside the $\varepsilon$-insensitive tube are considered support vectors in SVR. The goal in SVR is to minimize the difference between

observed and predicted numerical values (Rodríguez and Bajorath, 2022).

**Hyperparameter Tuning**

Before starting the SVM estimation process, we need to select the value of two hyperparameters, C and $\sigma$. There are different approaches to SVM hyperparameter search, the method that was used is K-fold cross-validation, in general this method applies a grid search to $\sigma$ and C with k-fold cross-validation using the "trainControl" function from the caret package with $k = 10$.

### 2.5.3 Generalized Linear Models

In the realm of machine learning, generalized linear models are gaining prominence as a flexible tool in comparison to linear or multiple regression models. Unlike their counterparts, generalized linear models aim to capture a wider range of outcome distributions by estimating the link function for the expected value of the outcome. This distinction can be seen mathematically in this equation:

$$E(Y) = \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{2.35}$$

The primary purpose of generalized linear models is to make predictions rather than provide causal explanations so it is consider as a limitation of using the generalized linear models for causal inference lies in their lack of assumptions about causality, which are necessary for making causal inferences (Arnold et al., 2020).

The first step in understanding how generalized linear models work is to check for the presence of three primary components:

- Distribution of Y ( random component)

  The response variable should follow the distribution of the exponential family, which includes the normal, poisson, binomial and other popular distributions; this applies to the response variable in this thesis, which is from the binomial family.
- Covariates (Systematic Component)

  The set of covariates $x_1, \ldots, x_n$ that relate to the expected value of the response variable, as in the classical linear model, we can write this relationship as $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 +$

$\ldots + \beta_k x_k$ where $\eta$ is called a linear predictor and it can be a quadratic, cubic or higher order polynomial.

- link function

  Instead of saying that $\mu = E(Y)$ is a linear predictor, we assume that this linear predictor is equal to a function of $\mu$ or E(Y), called a link function, as written in the equation below (Faraway, 2016),

$$f(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{2.36}$$

  Its purpose can be perceived as revealing the relationship between the mean of the response variable and a linear combination of the predictors, and this relationship can be non-linear, and this combination must be differentiable, since the mean here will be the inverse transformation of f,

$$\mu = f^{-1}(\eta) = f^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \tag{2.37}$$

The reason for linking the mean of the response variable rather than the response itself is that it requires a distribution that describes the population distribution of the data, which makes the process more complex.

**Logit link function**

Binomial response models are used in thesis. Because of their importance in this section, we will highlight some relevant characteristics of these models.

If Y is the response variable and there are k explanatory variables $x_1, x_2, \ldots, x_k$, the distribution of Y is determined to be binomial with $n_i$ independent Bernoulli trials and probability of success p on each trial, so Y $\sim$ Binomial$(n_i, p_i)$. To model response variables with GLM, it is necessary to use the link function called "logit", defined as

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

in this case, the linear predictor will be equal to the following

$$\eta_i = f(p) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \tag{2.38}$$

This model may be written as

$$\eta = f(p) = X\beta.$$

If our aim is to express equation 2.38 in terms of an odd ratio of response, then we have

$$\frac{p_i}{1-p_i} = e_i^{\eta} \tag{2.39}$$

finally, the probability of the response can be determined by

$$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} \tag{2.40}$$

**Assumptions of a GLM**

To define a GLM, we need the following assumption:

- The n response variables must be independent.
- The relationship between the mean of response transformed by the link function and the explanatory variables.
- Maximum likelihood (ML) or other methods are used to estimate the parameter, rather than ordinary least squares (OLS).
- Errors are independent but not normally distributed.

**Estimation in GLM Model**

MLE is often used to fit GLMs with binomial response variables instead of OLS because of the non-linearity in the parameters.

In Maximum Likelihood Estimation (MLE), our goal is to estimate the parameter by choosing it in a way that maximizes the likelihood of the observed sample.

The glm() function uses an algorithm called *iteratively reweighted least squares*. (Hanck et al., 2021).

### 2.5.4 Linear Probability Models

The Linear Probability Model (LPM) is a type of linear regression model commonly used to analyze binary outcomes. It is widely used in various fields due to its simplicity of interpretation and speed of computation (Al-Mejibli et al., 2020a). Despite the fact that the estimation results may fall outside the interval [0, 1] and the significance of linearity may not apply, the LPM is still considered appropriate for modelling binary outcomes. It is an alternative to logistic regression or probit regression and there are rarely large differences in the results between the three models.

If we have N number of observations or sample size and k independent variables, then the linear regression model is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \mu_i \\ &= X\beta + \mu, \end{aligned} \tag{2.41}$$

If $Y_i$ is binary, then we call this model a linear probability model. In the linear probability model

$$P(Y = 1|x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k. \tag{2.42}$$

where

$$E(Y|x_1, x_2, \ldots, x_k) = P(Y = 1|x_1, x_2, \ldots, x_k)$$

**Assumptions of Linear Probability Model**

- A linear relationship between the dependent and independent variables.
- Multivariate normality.

**OLS**

Each $\beta_i$ can be estimated using Ordinary Least Squares (OLS). The aim of OLS is to determine the regression coefficients that are close between the fitted regression line and the observed points. This closeness is measured by calculating the sum of the squared differences between the observed value of Y and it's predicted value.

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2$$

where $b_0, \ldots, b_k$ are the estimators of $\beta_1, \ldots, \beta_k$.

In LPM, $\beta_j$ can be interpreted as the change in the probability that $Y_i = 1$, holding the other $k-1$ predictors constant.

### 2.5.5 Recycled Predictions

A margin of response or recycled prediction refers to a statistic based on a fitted model where some or all of the covariates are fixed to be different from their actual values. In simple terms, it quantifies the change in the response variable resulting from a change in a given covariate.

For all methods except causal forest, we use a recycled prediction approach to determine the individual treatment effect and then estimate the average treatment effect. By regressing the outcome on the treatment, covariate, and the interaction between treatment and covariate, we predict two values for each individual. The first value corresponds to individuals who received the treatment (D=1) while the second value corresponds to individuals who did not receive the treatment (D=0). Individual treatment effects are then estimated by subtracting between these two corresponding predictions for each individual.

In the recycled prediction and subsequent calculation, the recycled prediction takes into account not only the direct effect of (D=0) or (D=1), but also the indirect effects resulting from the interaction of treatment with covariates.

The response being margined here is a statistic produced predict and the difference between two prediction values is the difference due to treatment holding other covarintes constant (Stata Corporation, 2005).

### 2.5.6 Principal Component Analysis (PCA)

In this thesis we have used unsupervised machine learning, which is the second type of machine learning. Despite its complexity, it performs several statistical tasks. In Chapter 3, we will demonstrate the effectiveness of a set of supervised machine learning tools for detecting causal effects. We will then use principal component analysis to examine its impact on the results.

#### Data Preprocessing

Real-world data tends to contain missing, noisy or inconsistent data because it is collected from different sources or because it is large in size, and this is where data preprocessing comes in to improve data quality and thus get accurate results. This can take several forms: data cleaning to remove noisy data, data integration, data reduction such as removing redundant features, and data transformation such as normalization.

#### The importance of dimensionality reduction (DR) techniques

The dimension reduction (DR) algorithm involves transforming the predictors $x_1, x_2, \ldots, x_p$ and then fitting a model using these transformed variables. As a result, (Bharadiya, 2023) mention the following benefits:

- The number of dimensions is reduced.
- We do not need additional computing time.
- Eliminate irrelevant, noisy and redundant data.
- Optimize the quality of the data.
- Using the reduced data in an algorithm makes it more efficient and improves accuracy.
- The classification process becomes easier.

#### What is principal component analysis?

Large data sets are becoming increasingly common and can be difficult to understand. Principal component analysis (PCA) is a method used to simplify data sets by reducing their dimensions. This helps to make the data easier to interpret while minimizing the loss of information. PCA achieves this by generating new variables that are uncorrelated and maximize the variance in the data (Jolliffe and Cadima, 2016). PCA is one of the

simplest and most robust ways to reduce dimensionality. It is also one of the oldest, and it often uncovers associations that were previously unrecognised, allowing interpretations that would not normally occur.

Suppose we have data with n features. Principal component analysis searches for k orthogonal vectors that best represent the data, where $k \leq n$.

The primary procedures are outlined as Han et al. (2012) explained:

- Data normalization to ensure uniformity of range across all attributes.
- Calculation of k orthonormal vectors, serving as the basis for the normalized input data. These vectors are called principal components and the input data is a linear function of the principal components.
- Ordering of the principal components based on the variance explained by each component.
- Reducing the size of the data by eliminating components with low variance.

# Chapter Three

# Application

## 3.1 Simulated Application

Simulation studies play a crucial role in guiding modelling decisions, especially when dealing with cases with unknown ground truth such as estimating treatment effect. The presence of confounding variables and insufficient sample size pose challenges in accurately measuring treatment effects. In this thesis, simulation studies were instrumental in determining the most appropriate machine learning method for estimating causal effects, and a comparison was made between these methods. As the true effect of treatment is unknown, observational data cannot be used with these models. This is where simulation comes in, particularly in the context of novel methods for causal inference in binary outcomes. The lack of guidance in the literature on the use of large sample sizes further emphasises the importance of simulation studies (Berrie, 2019). In essence, the aim was to evaluate and compare the performance of causal forests, support vector machines, generalized linear models and linear probability models on simulated datasets, and to evaluate the method based on the level of bias in each estimate attributed to the model.

We will perform the following tasks in our simulation study:

- The performance of the four machine learning techniques in estimating causal effects will be evaluated.
- Create different scenarios with different sample sizes to compare these techniques.
- The most and least effective methods will be identified.
- Principal component analysis will be used to examine the impact of the number of dimensions on our estimation.

### 3.1.1 Data generating

When thinking about a research question related to causal inference, it is important to think about the source of the data and then to think about the causal relationship between the variables.

In the thesis we conducted a simulation study by generating data for individuals who smoke in order to simulate a realistic scenario. This simulation was carried out using R programming, specifically version 4.3.2, through the 'simstudy' package.

First, we generated continuous covariates using the normal distribution and binary covariates using the bernoulli distribution for variables $X_1$ to $X_{15}$. We also simulated the treatment variable "smoking" and the outcome variable "lung cancer" using bernoulli distributions.

Next, we examined two studies : one with seven covariates and another with fifteen covariates. Furthermore, we created a third study by applying PCA on the second study. To conduct our study, we analyzed various sample sizes: n = 500, 1500, and 3000.

We considered five main scenarios. Scenario A assumes no treatment effects, so that $\tau_i = 0$ for all i. Scenario B assumes homogeneous or constant treatment effects, so that $\tau_i = 0.5$ for all i. Scenario C includes heterogeneous treatment effects, where some covariates form the $\tau$ value in a linear form. Scenario D includes heterogeneous treatment effects where some covariates form the $\tau$ value in a linear form with an interaction of these covariates. Scenario E also includes heterogeneous treatment effects where all covariates form the $\tau$ value in a linear form.

### 3.1.2 Model Development

#### Causal Forest

To build a Causal forest to estimate the causal effect, the following algorithm is used

**Algorithm 1** Causal Forest

**Input:** $Y, X, D, \tau, y\_(0)$

- **Convert:** Transform the value of $Y$ using the formula:

$$\text{new}(Y) = (\text{probabilities of } y\_(0) + (\text{probabilities of } \tau) \times D) \times 100$$

- **Fit the model:** Train the Causal forest using:

$$forest = causal\_forest(\text{new}(Y), D, X, \text{num.tree} = 1000)$$

- **Return:** Trained Causal Forest model
- **Prediction:**Predict the Individual Treatment Effect (ITE) using forest$predictions
- **Calculation:** Compute the Average Treatment Effect (ATE) as the mean of ITE using $average\_treatment\_effect(forest)$
- **Convert:** Adjust ATE by dividing by 100

**Support Vector Machine**

To build a SVM model to estimate the causal effect, the following algorithm is used

---

**Algorithm 2** Support Vector Machine (SVM)

---

**Input:** $Y, X, D, X * D, \tau, y\_(0)$

- **Convert:** Transform the value of $Y$ using the formula:

$$\text{new}(Y) = (\text{probabilities of } y(0) + (\text{probabilities of } \tau) \times D) \times 100$$

- **Hyperparameter Tuning:** Tune the sigma and $C$ parameters before estimation
- **Fit the model:** Train the SVM model using:

$$\begin{aligned} \text{svm}( \quad & \text{new}(Y), X, D, X * D, \text{data}, \\ & \text{kernel} = \text{'radial'}, \\ & \text{type} = \text{'eps-regression'}, \\ & \text{optimal\_sigma}, \text{optimal\_C}) \end{aligned}$$

- **Return:** Trained SVM model
- **Prediction:** Predict $Y(1)$ from the model using data where $D = 1$
- **Prediction:** Predict $Y(0)$ from the model using data where $D = 0$
- **Calculation:** Compute the Individual Treatment Effect (ITE) as $Y(1) - Y(0)$
- **Calculation:** Compute the Average Treatment Effect (ATE) as the mean of ITE
- **Convert:** Adjust ATE by dividing by 100

---

**Generalized Linear Models**

In order for Generalized Linear Models to effectively estimate the causal effect, it was necessary to change the probability in $y(0)$ to match the tau value we assumed, converting all Y values greater than 0.5 to 1 and those less than 0.5 to 0. This step can be directly understood because as the tau value increases, the probability of infected individuals appearing without treatment must decrease. This can show the effect of treatment on individuals.

To build a GLM model to estimate the causal effect, the following algorithm is used

**Algorithm 3** Generalized Linear Model (GLM)

**Input:** $Y, X, D, X * D, \tau, y\_(0)$

- **Build:** structure the value of $Y$ using the formula:

$$\text{probabilities of } Y = (\text{probabilities of } y(0) + (\text{probabilities of } \tau) \times D) \times 100$$

- **Convert:** Transform the value of $Y$ using the formula:

$$\text{new}(Y) = -log((1/(\text{probabilities of } Y)) - 1)$$

- **Convert:** By imposing a decision rule, convert new(Y) to 1 if new(Y) $> 0.5$ and to 0 if new(Y) $< 0.5$
- **Fit the model:** Train the GLM using:

$$\text{glm}(\text{new}(Y), X, D, X * D, \text{family = binomial(link = "logit")})$$

- **Return:** Trained GLM model
- **Prediction:** Predict $Y(1)$ from the model using data with $D = 1$
- **Prediction:** Predict $Y(0)$ from the model using data with $D = 0$
- **Calculation:** Compute the Individual Treatment Effect (ITE) as $Y(1) - Y(0)$
- **Calculation:** Compute the Average Treatment Effect (ATE) as the mean of ITE

**Linear Probability Model**

The LPM model is designed to accept only 0 and 1 as dependent variables. However, we decided to deviate from this standard by including values greater than 1 in our model.

To build a LPM model to estimate the causal effect, the following algorithm is used

**Algorithm 4** Recycled Prediction Linear Probability Model (RPLPM)

**Input:** $Y, X, D, X * D, \tau, y\_(0)$

- **Convert:** Transform the value of $Y$ using the formula:

$$\text{new}(Y) = (\text{probabilities of } y(0) + (\text{probabilities of } \tau) \times D) \times 100$$

- **Fit the model:** Train the linear model using:

$$\text{lm}(\text{new}(Y), X, D, X * D)$$

- **Return:** Trained linear model
- **Prediction:** Predict $Y(1)$ from the model using data where $D = 1$
- **Prediction:** Predict $Y(0)$ from the model using data where $D = 0$
- **Calculation:** Compute the Individual Treatment Effect (ITE) as $Y(1) - Y(0)$
- **Calculation:** Compute the Average Treatment Effect (ATE) as the mean of ITE
- **Convert:** Adjust ATE by dividing by 100

### 3.1.3 Performance criteria

We estimated the assumed conditional average treatment effect using the average of the individual estimates obtained from the ML models fitted on the data sets simulated as above. We reported the following criteria: bias, mean square error (MSE), $R^2$ and $\sigma^2(ITE)$.

An important aspect of simulation studies is that the assessment metrics themselves are subject to error, which led us to estimate using Monte Carlo and to carefully choose the number of iterations, which is 100.

### 3.1.4 Result

Unlike previous research on predicting the average treatment effect, our focus goes beyond identifying the optimal approach. We also aim to provide a comprehensive ranking of these methods by evaluating them against a specific metric.

**No effect scenario**

**Table 3.1**
*Performance evaluation in estimating the average treatment effect for the first scenario over different sample sizes* $(N = 500, 1500, 3000)$ *with* $\tau = 0$ *in simulation study A.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)} - \sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.50000 | 0.49847 | 0.00153 | 0.00028 | 0.49998 | 0 | 0.00015 | 0.00015 |
| SVM | 0.50000 | 0.49283 | 0.00717 | 0.00842 | 0.50005 | 0 | 0.00789 | 0.00789 |
| GLM | 0.50000 | 0.49818 | 0.00182 | 0.00807 | 0.50007 | 0 | 0.00700 | 0.00700 |
| RPLPM | 0.50000 | 0.49883 | 0.00117 | 0.00100 | 0.49999 | 0 | 0.00087 | 0.00087 |
| CF | 0.50000 | 0.49948 | 0.00052 | 0.00017 | 0.50001 | 0 | 0.00012 | 0.00012 |
| SVM | 0.50000 | 0.49576 | 0.00424 | 0.00629 | 0.50001 | 0 | 0.00602 | 0.00602 |
| GLM | 0.50000 | 0.50216 | -0.00216 | 0.00274 | 0.50003 | 0 | 0.00238 | 0.00238 |
| RPLPM | 0.50000 | 0.50012 | -0.00012 | 0.00032 | 0.50000 | 0 | 0.00029 | 0.00029 |
| CF | 0.50000 | 0.49999 | 0.00001 | 0.00014 | 0.50000 | 0 | 0.00011 | 0.00011 |
| SVM | 0.50000 | 0.49818 | 0.00182 | 0.00508 | 0.50000 | 0 | 0.00494 | 0.00494 |
| GLM | 0.50000 | 0.50240 | -0.00240 | 0.00128 | 0.50000 | 0 | 0.00109 | 0.00109 |
| RPLPM | 0.50000 | 0.50073 | -0.00073 | 0.00015 | 0.50000 | 0 | 0.00013 | 0.00013 |

- **Bias**

  Lower bias values indicate more accurate estimates of the average treatment effect. CF followed by RPLPM have the lowest bias at sample size 3000 and 1500, indicating higher accuracy compared to SVM at 3000 and GLM at 500.

- **Mean Squared Error (MSE):**

  CF and RPLPM achieve the lowest MSE values, indicating that they provide more accurate estimates compared to GLM and SVM at sample size for all at 3000.

- $R^2$**:**

  All methods at all sample sizes achieve a result close to .5, which is normal as the causal effect here is zero, meaning there is no variation in the true treatment effect values.

- $\sigma^2_{pr(\tau)}$

  It is very natural that the variance $\sigma^2_{pr(\tau)}$ is zero, since the tau value is zero for all individuals.

- $\sigma^2_{ITE}$

  On the other hand, $\sigma^2_{ITE}$ varies between methods, with CF and RPLPM showing the lowest variance, then GLM and finally SVM at sample size 3000.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  This metric reflects the difference between the variances, with CF and RPLPM showing the smallest differences, then GLM and finally SVM at sample size 3000.

**Table 3.2**

*Performance evaluation in estimating the average treatment effect for the first scenario over different sample sizes* $(N = 500, 1500, 3000)$ *with* $\tau = 0$ *in simulation study B.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.50000 | 0.49892 | 0.00108 | 0.00021 | 0.50002 | 0 | 0.00010 | 0.00010 |
| SVM | 0.50000 | 0.49088 | 0.00912 | 0.00934 | 0.49999 | 0 | 0.00905 | 0.00905 |
| GLM | 0.50000 | 0.50166 | -0.00166 | 0.01616 | 0.50008 | 0 | 0.01508 | 0.01508 |
| RPLPM | 0.50000 | 0.50048 | -0.00048 | 0.00187 | 0.50000 | 0 | 0.00179 | -0.00179 |
| CF | 0.50000 | 0.49859 | 0.00141 | 0.00012 | 0.50001 | 0 | 0.00008 | 0.00008 |
| SVM | 0.50000 | 0.49634 | 0.00366 | 0.00778 | 0.50001 | 0 | 0.00766 | 0.00766 |
| GLM | 0.50000 | 0.50400 | -0.00400 | 0.00537 | 0.50003 | 0 | 0.00496 | 0.00496 |
| RPLPM | 0.50000 | 0.49995 | 0.00005 | 0.00061 | 0.50001 | 0 | 0.00006 | 0.00006 |
| CF | 0.50000 | 0.49863 | 0.00137 | 0.00010 | 0.50000 | 0 | 0.00007 | 0.00007 |
| SVM | 0.50000 | 0.49752 | 0.00248 | 0.00688 | 0.50000 | 0 | 0.00680 | 0.00680 |
| GLM | 0.50000 | 0.50026 | -0.00026 | 0.00259 | 0.50000 | 0 | 0.00247 | 0.00247 |
| RPLPM | 0.50000 | 0.49984 | 0.00016 | 0.00029 | 0.50000 | 0 | 0.00027 | 0.00027 |

- **Bias**

  RPLPM performed best with a sample size of 1500, followed by GLM with a sample size of 3000. Causal forests performed best with a sample size of 500, and SVM

performed as expected with a sample size of 3000.

- **Mean Squared Error (MSE):**

  Consistent with the first simulation study, the models performed best at a sample size of 3000, in the following order: CF, RPLPM, GLM, and SVM.

- $R^2$**:**

  The value of $R^2$ did not change in this study and consequently the performance of the models did not change in term of this value.

- $\sigma^2_{pr(\tau)}$

  It is very natural that the variance $\sigma^2_{pr(\tau)}$ is zero, since its value is zero for all individuals.

- $\sigma^2_{ITE}$

  As for variance of $\sigma^2_{ITE}$, the least variance appeared from CF, RPLPM, then GLM, and finally SVM, and all the models showed this at a sample size of 3000 except RPLPM at a sample size of 1500.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  Since the variance of $\sigma^2_{pr(\tau)}$ is zero, and the order of the models remains the same for $\sigma^2_{ITE}$

**Table 3.3**
*Performance evaluation in estimating the average treatment effect for the first scenario over different sample sizes* $(N = 500, 1500, 3000)$ *with* $\tau$ *= 0 in simulation study C.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.50000 | 0.49951 | 0.00049 | 0.00019 | 0.50005 | 0 | 0.00008 | 0.00008 |
| SVM | 0.50000 | 0.48073 | 0.01927 | 0.00926 | 0.49999 | 0 | 0.00873 | 0.00893 |
| GLM | 0.50000 | 0.50096 | -0.00096 | 0.01131 | 0.50009 | 0 | 0.01024 | 0.01024 |
| RPLPM | 0.50000 | 0.49969 | 0.00031 | 0.00135 | 0.49999 | 0 | 0.00122 | 0.00122 |
| CF | 0.50000 | 0.49858 | 0.00142 | 0.00010 | 0.50001 | 0 | 0.00007 | 0.00007 |
| SVM | 0.50000 | 0.49039 | 0.00961 | 0.00797 | 0.50000 | 0 | 0.00777 | 0.00777 |
| GLM | 0.50000 | 0.50422 | -0.00422 | 0.00383 | 0.50003 | 0 | 0.00343 | 0.00343 |
| RPLPM | 0.5000 | 0.49990 | 0.00010 | 0.00042 | 0.50000 | 0 | 0.00039 | 0.00039 |
| CF | 0.50000 | 0.49960 | 0.00040 | 0.00007 | 0.50000 | 0 | 0.00508 | 0.00508 |
| SVM | 0.50000 | 0.49586 | 0.00414 | 0.00719 | 0.50000 | 0 | 0.00711 | 0.00711 |
| GLM | 0.50000 | 0.50218 | -0.00218 | 0.00179 | 0.50000 | 0 | 0.00163 | 0.00163 |
| RPLPM | 0.50000 | 0.50069 | -0.00069 | 0.00022 | 0.50000 | 0 | 0.00020 | 0.00020 |

- **Bias**

  After applying PCA, RPLPM remained the best performer with the same sample size as in the previous studies. Causal forests performed well with a sample size of 3000, while GLM performed best with a sample size of 500 and SVM at 3000.

- **Mean Squared Error (MSE):**

  Consistent with the first and second simulation study, the models performed best at a sample size of 3000, in the following order: CF, RPLPM, GLM, and SVM.

- $R^2$**:**

  Despite the application of PCA, nothing changed in the performance of the models, and

consequently the performance of the models did not change.

- $\sigma^2_{pr(\tau)}$

  It is logical that its value remains zero even after application PCA.

- $\sigma^2_{ITE}$

  As for variance of $\sigma^2_{ITE}$, the least variance appeared from CF, RPLPM, then GLM, and finally SVM, and all the models showed this at a sample size of 3000 except CF at a sample size of 1500.

- **($|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|$):**

  Since the variance of $\sigma^2_{pr(\tau)}$ is zero, and the order of the models remains the same for $\sigma^2_{ITE}$

**Constant scenario**

**Table 3.4**
*Performance evaluation in estimating the average treatment effect for the second scenario over different sample sizes* $(N = 500, 1500, 3000)$ *with* $\tau = 0.5$ *in simulation study A.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.62246 | 0.62078 | 0.00168 | 0.00028 | 0.49999 | 0 | 0.00015 | 0.00015 |
| SVM | 0.62246 | 0.61507 | 0.00739 | 0.00863 | 0.50000 | 0 | 0.00818 | -0.00818 |
| GLM | 0.62246 | 0.61862 | 0.00384 | 0.00770 | 0.50003 | 0 | 0.00675 | 0.00675 |
| RPLPM | 0.62246 | 0.62129 | 0.00117 | 0.00100 | 0.49998 | 0 | 0.00087 | 0.00087 |
| CF | 0.62246 | 0.62184 | 0.00062 | 0.00017 | 0.50000 | 0 | 0.00012 | 0.00012 |
| SVM | 0.62246 | 0.61846 | 0.00400 | 0.00614 | 0.50000 | 0 | 0.00592 | 0.00592 |
| GLM | 0.62246 | 0.62031 | 0.00215 | 0.00252 | 0.50000 | 0 | 0.0021 | 0.0021 |
| RPLPM | 0.62246 | 0.62258 | -0.00012 | 0.00032 | 0.50000 | 0 | 0.00029 | 0.00029 |
| CF | 0.62246 | 0.62234 | 0.00012 | 0.00014 | 0.50000 | 0 | 0.00012 | 0.00012 |
| SVM | 0.62246 | 0.62049 | 0.00197 | 0.00490 | 0.50000 | 0 | 0.00478 | 0.00478 |
| GLM | 0.62246 | 0.62173 | 0.00073 | 0.00123 | 0.50000 | 0 | 0.00102 | 0.00102 |
| RPLPM | 0.62246 | 0.62319 | -0.00073 | 0.00015 | 0.50001 | 0 | 0.00013 | 0.00013 |

- **Bias**

  Looking at the results presented in Table3.4, CF and RPLPM give the most accurate

estimates for $\tau_i = 0.5$ with sample sizes of 3000 and 1500 respectively. SVM then follows the GLM approach in terms of accuracy, specifically for a sample size of 3000.

- **Mean Squared Error (MSE):**

  For a sample size of 3000, MSE value was the least for all methods and their performance was arranged in this order: CF, RPLPM, GLM and SVM.

- $R^2$**:**

  There is no difference in the value of $R^2$ from the first scenario, given that the value of the treatment effect does not differ between individuals and $R^2$ is close to the value of 0.5, this appear to the rest of the simulation studies for this scenario.

- $\sigma^2_{pr(\tau)}$

  The value is equal to the variance of the values of the treatment effect in the first scenario, and this applies to the rest of the simulation studies for this scenario.

- $\sigma^2_{ITE}$

  The values of $\sigma^2_{ITE}$ showed that the least variance appeared from CF, RPLPM, then GLM, and finally SVM, and all the models showed this at a sample size of 3000 except CF at a sample size of 1500 and 3000.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  Since the variance of $\sigma^2_{pr(\tau)}$ is zero, and the order of the models remains the same for $\sigma^2_{ITE}$

**Table 3.5**
*Performance evaluation in estimating the average treatment effect for the second scenario over different sample sizes* $(N = 500, 1500, 3000)$ *with* $\tau = 0.5$ *in simulation study B.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.62246 | 0.62117 | 0.00129 | 0.00021 | 0.49999 | 0 | 0.00010 | 0.00010 |
| SVM | 0.62246 | 0.61238 | 0.01008 | 0.00986 | 0.50000 | 0 | 0.00959 | 0.00959 |
| GLM | 0.62246 | 0.62101 | 0.00145 | 0.01533 | 0.50003 | 0 | 0.01432 | 0.01432 |
| RPLPM | 0.62246 | 0.62294 | -0.00048 | 0.00187 | 0.50000 | 0 | 0.00179 | -0.00179 |
| CF | 0.62246 | 0.62086 | 0.00160 | 0.00012 | 0.50000 | 0 | 0.00008 | -0.00008 |
| SVM | 0.62246 | 0.61878 | 0.00368 | 0.00802 | 0.50000 | 0 | 0.00792 | 0.00792 |
| GLM | 0.62242 | 0.62226 | 0.00020 | 0.00519 | 0.50000 | 0 | 0.00489 | 0.00489 |
| RPLPM | 0.62246 | 0.62241 | 0.00005 | 0.00378 | 0.50000 | 0 | 0.05743 | -0.05743 |
| CF | 0.62246 | 0.62100 | 0.00146 | 0.0001 | 0.50000 | 0 | 0.00008 | 0.00008 |
| SVM | 0.62246 | 0.62003 | 0.00242 | 0.00689 | 0.50000 | 0 | 0.00683 | 0.00683 |
| GLM | 0.62244 | 0.62030 | 0.00215 | 0.00253 | 0.50000 | 0 | 0.00239 | 0.00239 |
| RPLPM | 0.62246 | 0.62230 | 0.00016 | 0.00029 | 0.50000 | 0 | 0.00027 | 0.00027 |

- **Bias**

  CF did not give us the most accurate estimate in Table 3.5. However, at a sample size of 500, it secured third place after RPLPM and GLM at a sample size of 1500 for both. Finally, SVM performed well with a sample size of 3000.

- **Mean Squared Error (MSE):**

  For a sample size of 3000, the MSE value was the least for all methods and their performance was arranged in this order: CF, RPLPM, GLM and SVM.

- $\sigma^2_{ITE}$

  The values of $\sigma^2_{ITE}$ showed that the least variance appeared from CF, RPLPM, then GLM, and finally SVM, and all the models showed this at a sample size of 3000.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  Since $\sigma^2_{pr(\tau)}$ is zero, and the order of the models remains the same for $\sigma^2_{ITE}$

**Table 3.6**
*Performance evaluation in estimating the average treatment effect for the second scenario over different sample sizes* $(N = 500, 1500, 3000)$ *with* $\tau = 0.5$ *in simulation study C.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.62246 | 0.62194 | 0.00052 | 0.00019 | 0.50000 | 0 | 0.00008 | -0.00008 |
| SVM | 0.62246 | 0.60202 | 0.02044 | 0.00983 | 0.50000 | 0 | 0.00927 | 0.00927 |
| GLM | 0.62246 | 0.62235 | 0.00011 | 0.01055 | 0.50004 | 0 | 0.00947 | 0.00947 |
| RPLPM | 0.62246 | 0.62215 | 0.00031 | 0.00135 | 0.49999 | 0 | 0.00122 | -0.00122 |
| CF | 0.62246 | 0.62090 | 0.00156 | 0.00011 | 0.50000 | 0 | 0.00007 | 0.00007 |
| SVM | 0.62246 | 0.61256 | 0.00990 | 0.00816 | 0.50000 | 0 | 0.00798 | 0.00798 |
| GLM | 0.62246 | 0.62120 | 0.00126 | 0.00352 | 0.50000 | 0 | 0.00321 | 0.00321 |
| RPLPM | 0.62246 | 0.62236 | 0.0001 | 0.00042 | 0.50000 | 0 | 0.00039 | 0.00039 |
| CF | 0.62246 | 0.62181 | 0.00065 | 0.00008 | 0.50000 | 0 | 0.00006 | -0.00006 |
| SVM | 0.62246 | 0.61807 | 0.00438 | 0.00724 | 0.50000 | 0 | 0.00716 | 0.00716 |
| GLM | 0.62246 | 0.62130 | 0.00116 | 0.00158 | 0.50000 | 0 | 0.00114 | 0.00144 |
| RPLPM | 0.62246 | 0.62315 | -0.00069 | 0.00022 | 0.50000 | 0 | 0.00020 | 0.00020 |

- **Bias**

  In the third simulation study, Table 3.6 shows an unexpected result. The RPLPM and GLM methods appear to be the most superior approaches at sample sizes of 1500 and

500. Subsequently, the CF method performs well at a sample size of 500, while the SVM method performs well at a sample size of 3000.

- **Mean Squared Error (MSE):**

  No change in the value of MSE after applying PCA. For a sample size of 3000, the MSE value was the least for all methods and their performance was arranged in this order: CF, RPLPM, GLM and SVM.

- $\sigma^2_{ITE}$

  The values of $\sigma^2_{ITE}$ showed that the least variance appeared from CF, RPLPM, then GLM, and finally SVM, and all the models showed this at a sample size of 3000.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  Since $\sigma^2_{pr(\tau)}$ is zero, and the order of the models remains the same for $\sigma^2_{ITE}$

**Heterogeneous effect by some covariants.**

For this scenario and the remaining scenarios, all results can be found in the appendices.

**Simulation Study A**

- **Bias**

  As expected, Table A.1 shows that the optimal approach in this scenario would have been CF with a sample size of 3000. We then observe that RPLPM with a sample size of 1500, followed by SVM with a sample size of 3000 and GLM with a sample size of 500.

- **Mean Squared Error (MSE):**

  It seems that all methods give us the best estimate at a sample size of 3000 according to the following order: CF, RPLPM, GLM and finally SVM.

- $R^2$**:**

  It is very natural for the values $R^2$ of to vary in this scenario, as the methods performed their best when the sample size increased to 3000, and the best was taken from CF, RPLPM, SVM, and GLM. This applies to the second simulation study, as even after increasing the number of variables, the performance of the models did not change.

- $\sigma^2_{pr(\tau)}$

  Approximately, the $\sigma^2_{pr(\tau)}$ value was close to 0.003 for all methods for each sample size.

- $\sigma^2_{ITE}$

  As for the value of $\sigma^2_{ITE}$, we benefit greatly from it in detecting the method that did not capture the heterogeneity in the value of tau. We start with GLM at a sample size of 3000, then CF at a sample size of 500, and then RPLPM and SVM at a sample size of 3000.

- **($|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|$):**

  For the difference between the two values above, the lowest value was returned from RPLPM at a sample size of 500, then CF at 3000, GLM at 1500, and finally SVM at 3000.

**Simulation Study B**

- **Bias**

  The introduction of additional variables in simulation study B had a significant impact on the performance of CF. Thus, the most accurate estimate was obtained by RPLPM with a sample size of 1500, followed by CF with a sample size of 500, SVM with a sample size of 3000, and finally GLM with a sample size of 500. This trend continued in simulation study C as shown in Table A.2 and Table A.3.

- **Mean Squared Error (MSE):**

  As in the first simulation study, all methods give us the best estimate at a sample size of 3000, but in the following order: CF, RPLPM, SVM, and finally GLM.

- $\sigma^2_{pr(\tau)}$

  Approximately, the $\sigma^2_{pr(\tau)}$ value was close to 0.009 for all methods for each sample size. The same thing happened in the third simulation study after applying the PCA.

- $\sigma^2_{ITE}$

  The values were more homogeneous at CF, then GLM, RPLPM and finally SVM, all at a sample size of 3000 except CF at 500 and No order changed after PCA.

- **($|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|$):**

  For $|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|$ the lowest value was returned by RPLPM with a sample size of 500,

then GLM with 500, CF and finally SVM with 3000.

**Simulation Study C**

- **Mean Squared Error (MSE):**
  This metric is different in this study to the two previous studies, by all methods give us the best estimate at a sample size of 3000 expect for RPLPM at 1500 and in the following order: RPLPM, CF, SVM and finally GLM.
- $R^2$**:**
  In contrast, the two previous simulation studies gave us the best value for $R^2$ at a sample size of 3000 for CF, but here RPLPM was the best.
- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**
  For $|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|$ the lowest value was returned by RPLPM with a sample size of 500, then GLM with 500, SVM and finally CF with 3000.

**Heterogeneous effect by some covariants with interaction between two variables.**

**Simulation Study A**

- **Bias**
  In the simulation study A, the results of Table A.4 show that the performance of the SVM seems to be the most optimal with a sample size of 3000. CF follows closely with the same sample size. However, with a sample size of 500, RLPLM emerges as the next best option. Finally, GLM is the least superior.
- **Mean Squared Error (MSE):**
  The performance of the models did not change in all simulation studies for this scenario. The rankings were CF, RPLPM, GLM and SVM at a sample size of 3000. However, in the third simulation study, RPLPM outperformed CF.
- $R^2$**:**
  CF achieved an amazing value for $R^2$, followed by RPLPM, SVM and finally GLM at a sample size of 3000. The same was true for the second simulation study, but GLM achieved her best at a sample size of 1500.

- $\sigma^2_{pr(\tau)}$

  Approximately, the $\sigma^2_{pr(\tau)}$ value was close to 0.003 for all methods for each sample size and all simulation studies.

- $\sigma^2_{ITE}$

  As for the value of $\sigma^2_{ITE}$, we benefit greatly from it in detecting the method that did not capture the heterogeneity in the value of tau. We start with GLM at a sample size of 3000, then CF at a sample size of 500, and then RPLPM and SVM at a sample size of 3000.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  It was found that the difference between the two variance values in this scenario was as small as possible when using a RPLPM at a sample size of 500, then CF at a sample size of 3000, GLM at 1500, and finally SVM at 3000.

**Simulation Study B**

- **Bias**

  In simulation study B, we can see from Table A.5 that the ranking was different from what was initially observed. First, CF and RPLPM were ranked first with sample sizes of 500 and 3000 respectively. Then SVM was ranked next with a sample size of 1500, followed by GLM with a sample size of 500

- $\sigma^2_{ITE}$

  The least variation in the individual treatment effect values was CF, GLM, RPLPM and finally SVM, all at a sample size of 3000 except CF at a sample size of 500.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  In this simulation study, the ideal sample size for the following methods was unified: GLM, CF, and SVM, but at a sample size of 1500, RPLPM outperformed them.

**Simulation Study C**

- **BAIS**:

  In the simulation study (C), the results presented in Table A.6 show that RPLPM remains the top performer with a sample size of 1500. It is closely followed by CF with a sample size of 3000, followed by SVM with a sample size of 3000. Finally, GLM

performs best with a sample size of 3000.

- $R^2$**:**

  CF is still the best method here, but at a sample size of 500, followed by RPLPM at a sample size of 3000, then SVM and GLM at a sample size of 1500.

- $\sigma^2_{ITE}$

  As for the value of $\sigma^2_{ITE}$, we benefit greatly from it in detecting the method that did not capture the heterogeneity in the value of tau. We start with CF at a sample size of 500, then RPLPM at a sample size of 1500, and then GLM and SVM at a sample size of 3000.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  After PCA, the ideal sample size for the following methods was unified: CF and SVM at sample size 3000, but at a sample size of 500, RPLPM outperformed them then GLM at 1500.

**Heterogeneous effect by all covariants.**

**Simulation Study A**

- **Bias**

  In all simulation studies, for the last scenario, the estimation of ATE by RPLPM with a sample size of 1500 is considered to be the most accurate estimation. CF is the next best estimator, followed by SVM and GLM with a sample size of 3000, as shown in Tables A.7, A.8 and A.9 below.

  Simulation study B differs from simulation study A in the performance of the CF and GLM. In study B, the CF performs well with a sample size of 500, while the GLM performs well with both 3000 and 1500 sample sizes. However, in simulation study C, the only difference is the sample size of CF and GLM, which is 500.

- **Mean Squared Error (MSE):**

  In this scenario the best estimate for this metric is CF, RPLPM, SVM and GLM at a sample size of 3000.

- $R^2$**:**

  In general, the value may be the highest possible at a sample size of 3000, but we notice that GLM gave a very weak and small result for this metric, while CF had the highest value, followed by RPLPM and SVM.

- $\sigma^2_{pr(\tau)}$

  Approximately, the $\sigma^2_{pr(\tau)}$ value was close to 0.008 for all methods for each sample size.

- $\sigma^2_{ITE}$

  The least variance in the effect of individual treatment was captured by GLM at a sample size of 3000, then CF at a sample size of 500, and then RPLPM and SVM at a sample size of 3000.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**

  At a sample size of 3000, SVM showed the least difference between the two variances, followed by GLM at 500, RPLPM at 1500, and finally CF at 3000.

**Simulation Study B**

- **Mean Squared Error (MSE):**

  After increasing the number of variables, the order changed to become RPLPM, CF, GLM and SVM. The same arrangement was obtained in the third simulation study.

- $R^2$**:**

  RPLPM outperformed CF in this simulation study, and the ranking remained the same as for the other methods, and this was no different in the study of the third simulation.

- $\sigma^2_{pr(\tau)}$

  Approximately, the $\sigma^2_{pr(\tau)}$ value was close to 0.003 for all methods for each sample size. The same thing happened in the third simulation study after applying the PCA.

- $\sigma^2_{ITE}$

  The lowest variance was at 3000 for all methods except CF, which had the lowest variance at a sample size of 500 and outperformed GLM, unlike the first simulation study. Even after applying PCA we got the same analysis.

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**
  The arrangement here was completely different from the first simulation study, as the smallest difference was due to RPLPM and GLM at 1500, then CF and SVM at 3000.

**Simulation Study C**

- $(|\sigma^2_{pr(\tau)} - \sigma^2_{ITE}|)$**:**
  This difference was the only metric that differed from the secound simulation study, as it was less than the difference from RPLPM, then GLM at 500, then CF and SVM at 3000.

## 3.2 Real-world Application

This thesis has focused primarily on data collection in the healthcare sector, recognising its significant role in causal inference over the years and the limited number of methods available for causal inference in healthcare. It is crucial to distinguish between correlation and causation, especially since the ultimate goal of treatment is to improve the well-being of the patient. Therefore, a thorough causal investigation is essential to determine whether or not a treatment has effectively achieved its intended outcomes.

### 3.2.1 Dataset Description

We used a carefully selected dataset called "Cirrhosis Patient Survival Prediction" from the UCI Irvine Machine Learning Repository. This dataset consists of 418 patients and its purpose is to predict patient survival in relation to cirrhosis, with a focus on evaluating the efficacy of the treatment D-penicillamine (Dickson and Langworthy, 2023).

The dataset contains a comprehensive set of information on people with cirrhosis, including 17 variables covering patient characteristics and health status. With a combination of integer, categorical and continuous variables, the dataset offered a wide range of features for analysis. The variable status expresses the outcome of interest, reflecting the survival status of patients.

The Cirrhosis Patient Survival Prediction dataset serves as a strong foundation for our study of supervised learning methods in causal effect estimation. Its composition, size

and consideration of diverse patient demographics contributed to the overall validity and relevance of our research findings, and it is important to emphasise that the inclusion of diverse patient demographics enhanced the generalisability of any causal effects identified.

Table 3.7 shows the baseline characteristics of individuals. The p-value indicates that there is no significant difference in the variables between the treatment and control groups, i.e. they were randomized.

**Table 3.7**
*Baseline characteristics*

| | Control Group N=140 | Treatment Group N=136 | P-value for t-test |
|---|---|---|---|
| N-Days | 2000.30 | 1957.40 | 0.75 |
| Age | 17705.00 | 18688.00 | 0.03 |
| Gender (F) | 90.0% | 85.294% | 0.24 |
| Gender (M) | 10.0% | 14.705% | 0.24 |
| Ascites (Y) | 5.714% | 8.088% | 0.44 |
| Ascites (N) | 94.286% | 91.912% | 0.44 |
| Hepatomegaly (Y) | 55.714% | 47.059% | 0.15 |
| Hepatomegaly (N) | 44.286% | 52.941% | 0.15 |
| Spiders (Y) | 28.571% | 29.412% | 0.88 |
| Spiders (N) | 71.429% | 70.588% | 0.88 |
| Edema (Y) | 12.857% | 17.647% | 0.26 |
| Edema (N) | 87.143% | 82.353% | 0.27 |
| Edema (S) | 7.143% | 11.029% | 0.27 |
| Bilirubin | 37.0% | 29.566% | 0.18 |
| Cholesterol | 376.28 | 366.10 | 0.72 |
| Albumin | 35.382% | 34.948% | 0.38 |
| Copper | 98.26 | 103.35 | 0.63 |
| Alk_Phos | 1977.10 | 2016.70 | 0.88 |
| SGOT | 126.39 | 121.78 | 0.50 |
| Triglycerides | 126.12 | 123.80 | 0.77 |
| Platelets | 265.38 | 258.06 | 0.52 |
| Prothrombin | 10.81 | 10.66 | 0.24 |
| Stage (1) | 2.1430% | 6.6180% | 0.07 |
| Stage (2) | 20.0% | 22.794% | 0.57 |
| Stage (3) | 42.857% | 37.5% | 0.37 |
| Stage (4) | 35.0% | 33.088% | 0.74 |

### 3.2.2 Data Preprocessing

To ensure accurate results and robust analysis, we have developed a series of data preprocessing steps. These steps are implemented before being used in supervised machine learning techniques, which include:

- Data cleaning by eliminating any row or column that contains a missing value.
- Convert the target variable to 0 or 1 by encoding it, and apply the same process to any categorical variable in the dataset.
- Data reduction using Principal Component Analysis (PCA). After obtaining the results, we proceeded with the implementation of this technique to observe the results.

### 3.2.3 Model Selection and Training

We possess a binary drug $D_i$ and a patient characteristics vector denoted by $X_i$ for each patient $i = 1,\ldots,276$. The patient characteristics vector $X_i$ includes the 17 covariates listed in Table 3.7. Additionally, we have a binary outcome $Y_i$ for each patient.

As previously mentioned, our estimation is based on the Neyman-Rubin framework of potential outcomes, which leads us to the fundamental problem of causal inference. Each patient is assigned to either the control group or the treated group. The control group comprises patients who do not receive the treatment or receive a placebo $(D_i = 0)$. On the other hand, if the patient takes the treatment $(D_i = 1)$, they belong to the treatment group.

### 3.2.4 Result

**Table 3.8**
*Comparison of Estimated Average Treatment Effects (ATE) and Confidence Intervals for Different Methods.*

| Method | Pr(ATE) | $\sigma^2_{(ITE)}$ | Confidence Interval |
|---|---|---|---|
| CF | -0.00680 | 0.00001 | (-0.00712 , -0.00647) |
| SVM | -0.00786 | 0.00189 | (-0.01299 , -0.00273) |
| GLM | -0.02921 | 0.02863 | (-0.04918 , -0.00925) |
| RPLPM | -0.00692 | 0.00116 | (-0.01094 , -0.00290) |

Table 3.8 shows almost identical results for the four techniques. Looking at the confidence intervals, we see that the CF method had the narrowest interval, followed by RPLPM, SVM and finally GLM.

**Table 3.9**
*Comparison of Estimated Average Treatment Effects (ATE) and Confidence Intervals for Different Methods After PCA*

| Method | Pr(ATE) | $\sigma^2_{(ITE)}$ | Confidence Intervals |
|---|---|---|---|
| CF | -0.00791 | 0.00001 | (-0.00832 , -0.00749) |
| SVM | -0.00579 | 0.00186 | (-0.01087 , -0.00700) |
| GLM | -0.03375 | 0.01384 | (-0.04763 , -0.01987) |
| RPLPM | -0.00712 | 0.00057 | (-0.00994 , -0.00430) |

We expected that there would be no significant change in the results after the introduction of PCA, as the order of the techniques remained unchanged for CF and GLM, but the order of SVM and RPLPM change, as shown in Table 3.9.

In the real data results, the homogeneity of the causal effect was observed. The four methods gave similar results, and the confidence intervals indicated that the most accu-

rate estimate of the causal effect was obtained from CF, RPLPM then provided the next best estimate, followed by SVM and finally GLM. However, after performing PCA, the ranking is different, being CF, SVM, RPLPM and GLM.

Our results set with those of Jamse (1985), indicating that this treatment leads to higher mortality rates, show a negative effect. In addition, there was homogeneity observed in the result between different individuals. This suggests that the machine learning methods used here performed well in determining causality.
Matloff et al. (1982) also concluded that D-penicillamine, at the dose he used, was not effective in treating primary biliary cirrhosis and was associated with a high incidence of serious side effects.

**Analysis of individual heterogeneity**

CF heterogeneity is expected to be lower than other methods due to the difficulty of these methods in dealing with overfitting and the interaction between treatment and covariates as we can see from Figure 3.1 and Figures 3.2.

**Figure 3.1**
*Histograms of ITE for each method before PCA*

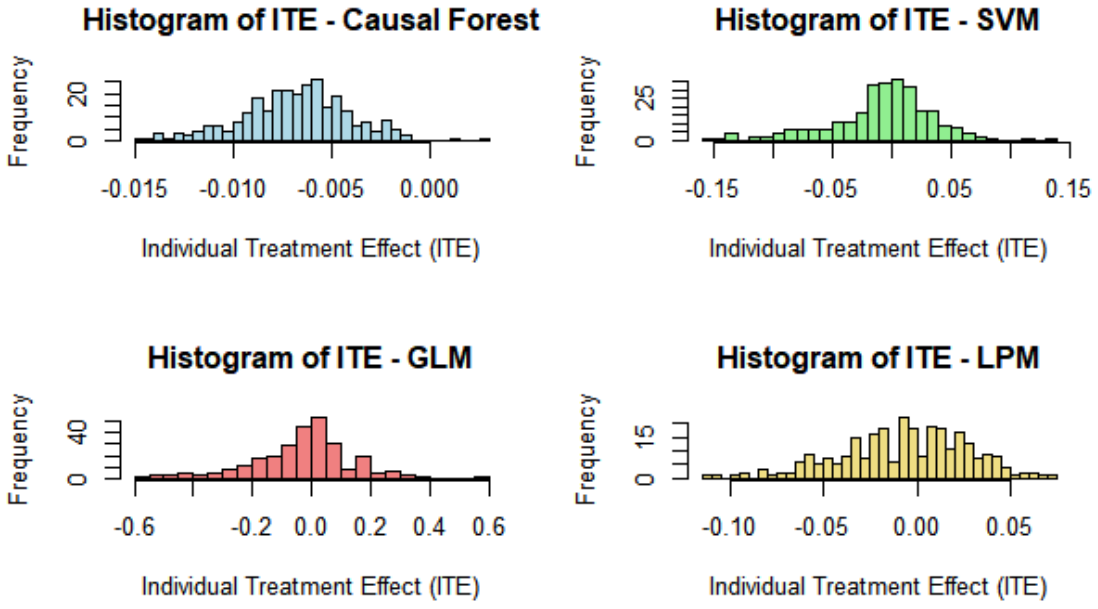
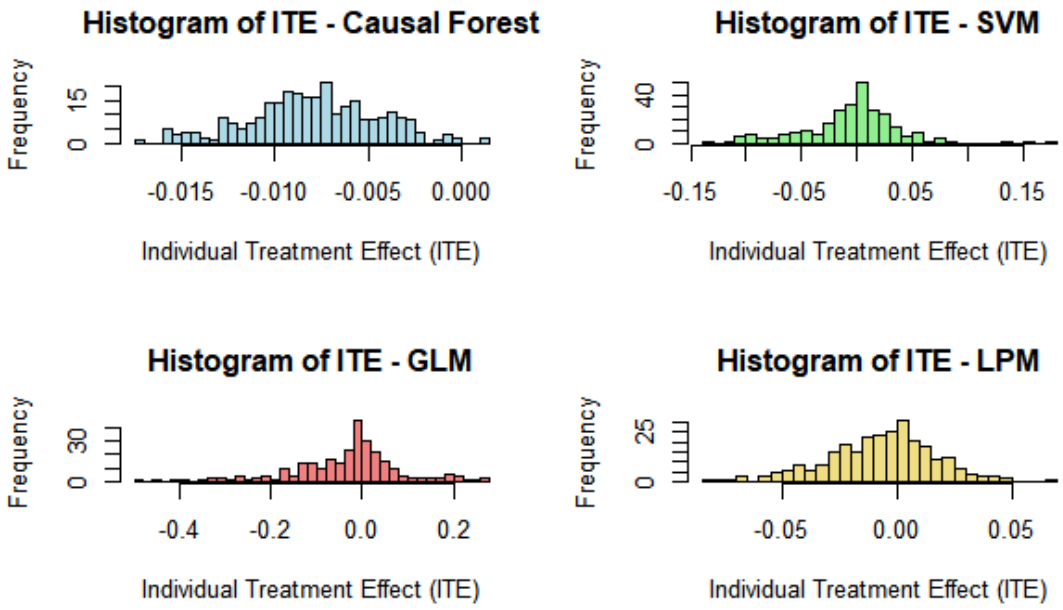


**Figure 3.2**
*Histograms of ITE for each method after PCA*




**Comparison of simulation and real data results**

Based on the variance calculations presented in Tables 3.5 and 3.6 for a sample size of 500, in conjunction with the information presented in Tables 3.8 and 3.9, it is evident that there was no difference in the ranking of the methods for the simulation result and the real result. And we have an indication from the variance that there is no heterogeneity in the causal effect within the causal forests when compared to alternative methods.

# Chapter Four

# Conclusion

This thesis evaluates the effectiveness of four machine learning techniques used in causal inference of binary treatment and outcome variables. The techniques evaluated are Causal Forest (CF), Support Vector Machine (SVM), Generalized Linear Models (GLM) and Linear Probability Model (LPM). In Chapter 2, we introduce the Neyman-Rubin causal model and the assumptions used to apply the algorithms discussed in Chapter 3. We also present the methodology used for each of the four methods. The initial analysis on real data set showed that the first method performed better than the others, followed by LPM, SVM and finally GLM. However, when Principal Component Analysis (PCA) was applied, the Causal Forest method remained the most effective, with SVM replacing LPM as the second best approach. In conclusion, it is recommended to use the causal forest method to estimate the causal effect of binary treatment and outcome variables in real data set.

In chapter 3, we developed four algorithms to estimate causal effect from randomized control trials. The first algorithm is the Causal Forest algorithm (CF), the second is the Support Vector Machine algorithm (SVM), the third is the Generalized Linear Models algorithm (GLM) and finally the Recycled Prediction Linear Probability Models algorithm (RPLPM). In simulation studies the generalized linear models algorithm (GLM) showed better performance only when the probability of y(0) was changed and the findings underscore the distinct strengths and limitations of each method under varying conditions, such as changes in sample size and the number of explanatory variables. Notably, SVM, GLM, and LPM showed a greater capacity for detecting heterogeneity than Causal Forest. In general simulation studies have shown that no single method can be applied to all scenarios and sample sizes in all cases. Instead, each method shows superior performance in certain scenarios and sample sizes as. The results indicate that the performance of these methods on real data and in scenario of constant effect is close. From the results of our simulation studies we concluded the best method that can be used for each of the five scenarios, along with the appropriate sample size for that method. It is noticeable that after running the PCA, RPLPM with a sample size of 1500 became the most appropriate

method to use in each scenario. So we can say that we have not only made a comparison, but we have found the best of one of these methods over its counterparts.

In Table A.10 in the appendices, a complete summary of all the results of the simulation studies is presented, showing us the ranking of the methods in terms of their effectiveness in estimating the causal effect of any of the five scenarios, even when changing the sample size and applying PCA or not.

The importance of selecting precise methods in causal inference is highlighted in this study, resulting in more accurate and unbiased causal estimation in real-world situations. The thesis has effectively achieved its objectives. Our research is significant as it not only tackles present inquiries but also paves the way for future research possibilities, with potential exploration of applying these methods to other types of response variables data and further investigation into additional machine learning techniques.

# List of Abbreviations

| Abbreviations | Meaning |
|---|---|
| ACE | Average Causal Effect |
| ASCM | Augmented Synthetic Control Method |
| ATE | Average Treatment Effect |
| BAC | Bayesian Adjustment for Confounding |
| CART | Classification And Regression Tree |
| CATE | Conditional Average Treatment Effect |
| CF | Causal Forest |
| DID | Difference-In-Differences |
| EMSE | Expected Mean Squared Errors |
| GRF | Generalized Random Forest |
| HTE | Heterogeneous Treatment Effect |
| IV | Instrumental Variable |
| LPM | Linear Probability Model |
| MCGAN | Mean Conditional Generative Adversarial Network |
| MLE | Maximum Likelihood Estimation |
| ML | Machine Learning |
| MSE | Mean Squared Errors |
| ODA | Optimal Discriminant Analysis |
| OLS | Ordinary Least Squares |
| PCA | Principal Component Analysis |
| RBF | Radial Basis Function |
| RCT | Randomized Controlled Trials |
| RDD | Regression Discontinuity Design |
| RF | Random Forest |
| RF(C) | Random Forest Classification |
| RF(R) | Random Forest Regression |
| SMMs | Structural Mean Models |

| Abbreviations | Meaning |
|---|---|
| SVM | Support Vector Machine |
| SV | Support Vector |
| USVMs | Uplift Support Vector Machines |

# Bibliography


Al-Mejibli, I. S., Alwan, J. K., and Abd Dhafar, H. (2020a). Analysis of binary dependent variables using linear probability model and logistic regression: A replication study.

Al-Mejibli, I. S., Alwan, J. K., and Abd Dhafar, H. (2020b). The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering*, 10(5):5497.

Altman, N. and Krzywinski, M. (2015). Points of significance: Association, correlation and causation. *Nature Methods*, 12(10).

Arnold, K. F., Davies, V., de Kamps, M., Tennant, P. W., Mbotwa, J., and Gilthorpe, M. S. (2020). Reflection on modern methods: Generalized linear models for prognosis and intervention—theory, practice and implications for machine learning. *International Journal of Epidemiology*, 49(6):2074–2082.

Asher, C. A. (2021). *Investigating Sources of Treatment Effect Heterogeneity in Intervention Research*. PhD thesis, Harvard University.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Balzer, L. and Petersen, M. (2021). Invited commentary: Machine learning in causal inference—how do i love thee? let me count the ways. *American Journal of Epidemiology*, 190:1483–1487.

Berrie, L. (2019). *Causal Inference Methods and Simulation Approaches in Observational Health Research Within a Geographical Framework*. PhD thesis, University of Leeds.

Bharadiya, J. P. (2023). Tutorial on principal component analysis for dimensionality reduction in machine learning. *International Journal of Innovative Science and Research Technology*, 8(5):2028–2032.

Bhide, A., Shah, P. S., and Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, 97(4):380–387.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *Test*, 25:197–227.

Bojinov, I., Chen, A., and Liu, M. (2020). The importance of being causal. *Harvard Data Science Review*, 2(3):6.

Burkov, A. (2020). *The Hundred Page Machine Learning Book*. Self-published.

Bénard, C. and Josse, J. (2023). Variable importance for causal forests: Breaking down the heterogeneity of treatment effects. *arXiv preprint arXiv:2308.03369*.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., and Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*. doi:10.1016/j.neucom.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2023). Toward personalized inference on individual treatment effects. *Proceedings of the National Academy of Sciences*, 120(7):e2300458120.

Chikahara, Y. and Fujino, A. (2018). Causal inference in time series via supervised learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2042–2048.

Clarke, P. S. and Windmeijer, F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, 11(4):756–770.

Cole, S. R. and Frangakis, C. E. (2009). The consistency statement in causal inference. *Epidemiology*, 20(1):3–5.

Collischon, M. (2023). Methods to estimate causal effects: An overview on iv, did and rdd and a guide on how to apply them in practice. *SozW Soziale Welt*, 73(4):713–735.

Crown, W. H. (2019). Real-world evidence, causal inference, and machine learning. *Value in Health*, 22(5):587–592.

Cui, P., Shen, Z., Li, S., Yao, L., Li, Y., Chu, Z., and Gao, J. (2020). Causal inference meets machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3527–3528.

Dandl, S., Haslinger, C., Hothorn, T., Seibold, H., Sverdrup, E., Wager, S., and Zeileis, A. (2024). What makes forest-based heterogeneous treatment effect estimators work? *The Annals of Applied Statistics*, 18(1):506–528.

Dickson, E., G. P. F. T. F. L. and Langworthy, A. (2023). Cirrhosis Patient Survival Prediction. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5R02G.

Dorie, V., Perrett, G., Hill, J., and Goodrich, B. (2022). Stan and bart for causal inference: Estimating heterogeneous treatment effects using the power of stan and the flexibility of machine learning. *Entropy*, 24(12):1782.

Dubitzky, W., Wolkenhauer, O., Cho, K. H., and Yokota, H. (2013). *Encyclopedia of Systems Biology*, volume 402. Springer, New York, NY, USA.

Edwards, J. K., Cole, S. R., Lesko, C. R., Mathews, W. C., Moore, R. D., Mugavero, M. J., and Westreich, D. (2016). An illustration of inverse probability weighting to estimate policy-relevant causal effects. *American Journal of Epidemiology*, 184(4):336–344.

Eid, A. and Wicker, N. (2023). Kber: A kernel bandwidth estimate using the ricci curvature. *Communications in Statistics-Theory and Methods*, 52(2):398–408.

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC.

Ge, Q., Huang, X., Fang, S., Guo, S., Liu, Y., Lin, W., and Xiong, M. (2020). Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in Genetics*, 11:585804.

Gershman, S. J. and Ullman, T. D. (2023). Causal implicatures from correlational statements. *PLOS ONE*, 18(5):e0286067.

Goetghebeur, E., le Cessie, S., De Stavola, B., Moodie, E. E., Waernbaum, I., and the topic group Causal Inference (TG7) of the STRATOS initiative (2020). Formulating causal questions and principled statistical answers. *Statistics in Medicine*, 39(30):4922–4948.

Greenfield, S., Kravitz, R., Duan, N., and Kaplan, S. H. (2007). Heterogeneity of treatment effects: Implications for guidelines, payment, and quality assessment. *The American Journal of Medicine*, 120(4):S3–S9.

Gutierrez, P. and Gérardy, J. (2017). Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*, pages 1–13.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques, Third Edition*. University of Illinois at Urbana-Champaign.

Hanck, C., Arnold, M., Gerber, A., and Schmelzer, M. (2021). *Introduction to Econometrics with R*. Universität Duisburg-Essen.

Hattab, Z., Doherty, E., Ryan, A. M., and O'Neill, S. (2024). Heterogeneity within the oregon health insurance experiment: An application of causal forests. *PLOS ONE*, 19(1):e0297205.

Hernán, A. and Robins, J. (2020). *Causal inference: What if*. Chapman & Hall/CRC, Boca Raton, FL.

Hoogland, J., IntHout, J., Belias, M., Rovers, M. M., Riley, R. D., Harrell, E. J., and Reitsma, J. B. (2021). A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine*. doi:10.1002/sim.9154.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation.

Jamse, O. F. (1985). D-penicillamine for primary biliary cirrhosis. *Gut*, 26(2):109.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).

Kadam, V. S., Kanhere, S., and Mahindrakar, S. (2020). Regression techniques in machine learning & applications: A review. *International Journal of Research in Applied Sciences and Engineering Technology*, 8:826–830.

Keller, B. and Branson, Z. (2024). Defining, identifying, and estimating effects with the rubin causal model: A review for education research.

Kim, D. (2023). *Estimating Heterogeneous Impacts Using Causal Forest*. PhD thesis, KDI School.

Kitazawa, Y. (2022). Estimating the average causal effect of intervention in continuous variables using machine learning. arXiv preprint.

Kleinberg, S. (2015). *Why: A Guide to Finding and Using Causes*. O'Reilly Media, Inc.

Laffers, L. (2020). Identification of the average treatment effect when sutva is violated. Discussion papers on business and economics, University of Southern Denmark.

Lara (2024). On the (in) compatibility between potential outcomes and structural causal models and its signification in counterfactual inference.

Le Borgne, F., Chatton, A., Léger, M., Lenain, R., and Foucher, Y. (2021). G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. *Scientific Reports*, 11(1):1435.

Lee, A., Inceoglu, I., Hauser, O., and Greene, M. (2020). Determining causal relationships in leadership research using machine learning: The powerful synergy of experiments and data science. *The Leadership Quarterly*.

Liang, X. J. and Yang, X. Q. (2021). A note on causation versus correlation in an extreme situation. *Entropy*, 23(3):316.

Linden, A. and Yarnold, P. (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*, pages 875–885.

Liu, B., Ma, M., and Chang, J., editors (2012). *Information Computing and Applications: Third International Conference*, volume 7473 of *CICA 2012, Chengde, China, September 14-16, 2012, Revised Selected Papers*. Springer.

Matloff, D. S., Alpert, E., Resnick, R. H., and Kaplan, M. M. (1982). A prospective trial of d-penicillamine in primary biliary cirrhosis. *New England Journal of Medicine*, 306(6):319–326.

Naimi, A., Mishler, A., and Kennedy, E. (2020). Challenges in obtaining valid causal effect estimates with machine learning algorithms. arXiv preprint.

Parikh, H., Varjao, C., Xu, L., and Tchetgen, E. (2022). Validating causal inference methods. In *International Conference on Machine Learning*, pages 17346–17358. Proceedings of Machine Learning Research (PMLR).

Pearl, J. (2007). The mathematics of causal inference in statistics. In *2007 JSM Proceedings*, page 337.

Pearl, J., Glymour, M., and Jewell, N. P. (2015). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Petersen, A. H., Ramsey, J., Ekstrøm, C., and Spirtes, P. (2022). Causal discovery for observational sciences using supervised machine learning. arXiv preprint.

Pisner, D. A. and Schnyer, D. M. (2020). Support vector machine. In *Machine Learning*, pages 101–121. Academic Press.

Ratkovic, M. (2014). Balancing within the margin: Causal effect estimation with support vector machines. Department of Politics, Princeton University, Princeton, NJ.

Rodríguez, R. and Bajorath, J. (2022). Evolution of support vector machine and regression modeling in chemoinformatics and drug discovery. *Journal of Computer-Aided Molecular Design*, 36(5):355–362.

Roman, I., Santana, R., Mendiburu, A., and Lozano, J. A. (2021). In-depth analysis of svm kernel learning and its components. *Neural Computing and Applications*, 33(12):6575–6594.

Samii, C., Paler, L., and Daly, S. Z. (2016). Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia. *Political Analysis*, 24(4):434–456.

Sekhon, J. (2008). The neyman—rubin model of causal inference and estimation via matching methods.

Smith, J. (2022). Treatment effect heterogeneity. *Evaluation Review*, 46(5):652–677.

Song, C., Liu, B., Cheng, K., Cole, M. A., Dai, Q., Elliott, R. J., and Shi, Z. (2023). Attribution of air quality benefits to clean winter heating policies in china: Combining machine learning with causal inference. *Environmental Science and Technology*.

Stata Corporation (2005). *Stata Base Reference Manual: Release 9 (Vol. 3)*. Stata Corporation, College Station, TX.

Stolberg, H. O., Norman, G., and Trop, I. (2004). Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544.

Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, 18(10):1099–1104.

Suthaharan, S. (2016). *Support Vector Machine: Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Springer.

Tarr, A. and Imai, K. (2021). Estimating average treatment effects with support vector machines. arXiv preprint.

Timofeev, R. (2004). *Classification and Regression Trees (CART) Theory and Applications*. PhD thesis, Humboldt University, Berlin.

Venkatasubramaniam, A., Mateen, B. A., Shields, B. M., Hattersley, A. T., Jones, A. G., Vollmer, S. J., and Dennis, J. M. (2022). Comparison of causal forest and regression-based approaches to evaluate treatment effect heterogeneity: An application for type 2 diabetes precision medicine. *BMC Medical Informatics and Decision Making*, 23.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

Younas, N., Ali, A., Hina, H., Hamraz, M., Khan, Z., and Aldahmani, S. (2022). Optimal causal decision trees ensemble for improved prediction and causal inference. *IEEE Access*, 10:13000–13011.

Yu, F., Moh, M., and Moh, T. (2016). Towards extracting drug-effect relation from twitter: A supervised learning approach. In *Proceedings of IEEE Big Data Security on Cloud, High Performance and Smart Computing, and Intelligent Data and Security (BigDataSecurity-HPSC-IDS)*, pages 339–344.

Zaniewicz, L. and Jaroszewicz, S. (2013). Support vector machines for uplift modeling. In *IEEE 13th International Conference on Data Mining Workshops*.

Zhao, J., Runfola, D. M., and Kemper, P. (2017). Simulation study in quantifying heterogeneous causal effects. In *2017 Winter Simulation Conference (WSC)*, pages 1925–1936.

Zhao, J. S. (2018). Quantifying and explaining causal effects of world bank aid projects. *Journal of Causal Inference*, 6(2):201–217.

# Appendices

# Appendix A

# Tables

**Table A.1**
*Performance evaluation in estimating the average treatment effect for the third scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study A.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.57348 | 0.57192 | 0.00156 | 0.00132 | 0.66738 | 0.00296 | 0.00121 | 0.00175 |
| SVM | 0.57346 | 0.56646 | 0.007 | 0.00909 | 0.16010 | 0.00298 | 0.00985 | 0.00688 |
| GLM | 0.57340 | 0.57794 | -0.00454 | 0.01082 | 0.10110 | 0.00298 | 0.00710 | 0.00412 |
| RPLPM | 0.57340 | 0.57213 | 0.00128 | 0.00187 | 0.53308 | 0.00298 | 0.00323 | 0.00025 |
| CF | 0.57397 | 0.57328 | 0.00069 | 0.00073 | 0.79597 | 0.00296 | 0.00184 | 0.00112 |
| SVM | 0.57383 | 0.56974 | 0.00409 | 0.00661 | 0.22571 | 0.00296 | 0.00798 | 0.00502 |
| GLM | 0.57369 | 0.58107 | -0.00739 | 0.00577 | 0.10440 | 0.00296 | 0.00227 | 0.00069 |
| RPLPM | 0.57369 | 0.57377 | -0.00008 | 0.0012 | 0.63005 | 0.00296 | 0.00242 | 0.00054 |
| CF | 0.57370 | 0.57366 | 0.00004 | 0.00049 | 0.84993 | 0.00296 | 0.00230 | 0.00067 |
| SVM | 0.57376 | 0.57168 | 0.00208 | 0.00536 | 0.24719 | 0.00297 | 0.00667 | 0.00370 |
| GLM | 0.57381 | 0.58238 | -0.00857 | 0.00436 | 0.10073 | 0.00296 | 0.00105 | 0.00192 |
| RPLPM | 0.57381 | 0.57446 | 0.00064 | 0.00103 | 0.66828 | 0.00296 | 0.00216 | 0.00081 |

**Table A.2**
*Performance evaluation in estimating the average treatment effect for the third scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study B.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.81056 | 0.80915 | 0.00141 | 0.00522 | 0.63034 | 0.00934 | 0.00155 | 0.00779 |
| SVM | 0.81043 | 0.79862 | 0.01181 | 0.01161 | 0.36414 | 0.00927 | 0.01732 | 0.00805 |
| GLM | 0.81028 | 0.80136 | 0.00892 | 0.02086 | 0.05425 | 0.00936 | 0.01094 | 0.00158 |
| RPLPM | 0.81028 | 0.81102 | -0.00073 | 0.00359 | 0.67106 | 0.00936 | 0.00971 | 0.00034 |
| CF | 0.81026 | 0.80843 | 0.00183 | 0.00287 | 0.76879 | 0.00938 | 0.00385 | 0.00553 |
| SVM | 0.81030 | 0.80619 | 0.00411 | 0.00914 | 0.43810 | 0.00933 | 0.01576 | 0.00644 |
| GLM | 0.81022 | 0.79955 | 0.01067 | 0.01332 | 0.05844 | 0.00942 | 0.00353 | 0.00589 |
| RPLPM | 0.81022 | 0.81015 | 0.00007 | 0.00234 | 0.76263 | 0.00942 | 0.00825 | 0.00116 |
| CF | 0.81036 | 0.80844 | 0.00192 | 0.00185 | 0.82457 | 0.00934 | 0.00576 | 0.00358 |
| SVM | 0.81017 | 0.80740 | 0.00277 | 0.00771 | 0.48644 | 0.00940 | 0.01458 | 0.00518 |
| GLM | 0.81044 | 0.80131 | 0.00913 | 0.01127 | 0.06159 | 0.00935 | 0.00163 | 0.00772 |
| RPLPM | 0.81044 | 0.81018 | 0.00025 | 0.00200 | 0.79031 | 0.00935 | 0.00778 | 0.00137 |

**Table A.3**
*Performance evaluation in estimating the average treatment effect for the third scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study C.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.81038 | 0.80996 | 0.00043 | 0.00719 | 0.36858 | 0.00941 | 0.00077 | 0.00864 |
| SVM | 0.81007 | 0.78557 | 0.02450 | 0.01315 | 0.27942 | 0.00939 | 0.01588 | 0.00649 |
| GLM | 0.81027 | 0.80329 | 0.00698 | 0.01634 | 0.06791 | 0.00936 | 0.00667 | 0.00269 |
| RPLPM | 0.81027 | 0.81004 | 0.00024 | 0.00497 | 0.51540 | 0.00936 | 0.00723 | 0.00213 |
| | | | | | | | | |
| CF | 0.81036 | 0.80823 | 0.00213 | 0.00675 | 0.45706 | 0.00942 | 0.00133 | 0.00809 |
| SVM | 0.81010 | 0.79974 | 0.01035 | 0.01047 | 0.33114 | 0.00937 | 0.01414 | 0.00477 |
| GLM | 0.81032 | 0.79927 | 0.01104 | 0.01207 | 0.05690 | 0.00938 | 0.00221 | 0.00717 |
| RPLPM | 0.81032 | 0.81040 | -0.00008 | 0.00410 | 0.57106 | 0.00938 | 0.00615 | 0.00323 |
| | | | | | | | | |
| CF | 0.81048 | 0.80983 | 0.00065 | 0.00545 | 0.49059 | 0.00932 | 0.00181 | 0.00751 |
| SVM | 0.81040 | 0.80563 | 0.00477 | 0.00911 | 0.37897 | 0.00936 | 0.01352 | 0.00417 |
| GLM | 0.81053 | 0.80201 | 0.00852 | 0.01086 | 0.05365 | 0.00934 | 0.00104 | 0.00830 |
| RPLPM | 0.81053 | 0.81111 | -0.00059 | 0.00411 | 0.56377 | 0.00934 | 0.00560 | 0.00374 |

**Table A.4**

*Performance evaluation in estimating the average treatment effect for the fourth scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study A.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.57376 | 0.57191 | 0.00165 | 0.00122 | 0.69034 | 0.00297 | 0.00129 | 0.00168 |
| SVM | 0.57392 | 0.56946 | 0.00446 | 0.00871 | 0.1891 | 0.00299 | 0.00991 | *0.00691* |
| GLM | 0.57354 | 0.57915 | -0.00560 | 0.01041 | 0.11772 | 0.00295 | 0.00637 | 0.00342 |
| RPLPM | 0.57354 | 0.57398 | -0.00044 | 0.00178 | 0.54059 | 0.00295 | 0.00298 | 0.00003 |
| | | | | | | | | |
| CF | 0.56664 | 0.57380 | -0.00716 | 0.00071 | 0.79745 | 0.00298 | 0.00203 | 0.00095 |
| SVM | 0.57354 | 0.57032 | 0.00322 | 0.00612 | 0.23092 | 0.00298 | 0.00737 | 0.00440 |
| GLM | 0.57378 | 0.58344 | -0.00965 | 0.00579 | 0.10805 | 0.00298 | 0.00228 | 0.00070 |
| RPLPM | 0.57379 | 0.57496 | -0.00117 | 0.00118 | 0.63638 | 0.00298 | 0.00225 | 0.00073 |
| | | | | | | | | |
| CF | 0.57370 | 0.57402 | -0.00031 | 0.00047 | 0.85439 | 0.00297 | 0.00234 | 0.00063 |
| SVM | 0.57382 | 0.57383 | -0.00002 | 0.00495 | 0.27061 | 0.00298 | 0.00629 | 0.00331 |
| GLM | 0.57378 | 0.58128 | -0.00750 | 0.00440 | 0.11384 | 0.00299 | 0.00107 | 0.00191 |
| RPLPM | 0.57378 | 0.57450 | -0.00072 | 0.00103 | 0.66926 | 0.00299 | 0.00216 | 0.00083 |

**Table A.5**
*Performance evaluation in estimating the average treatment effect for the fourth scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study B.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.57361 | 0.57383 | -0.00022 | 0.00154 | 0.65209 | 0.00299 | 0.00084 | 0.00214 |
| SVM | 0.57358 | 0.56589 | 0.00770 | 0.01003 | 0.13230 | 0.00297 | 0.01078 | 0.00781 |
| GLM | 0.57389 | 0.57713 | -0.00324 | 0.01761 | 0.04327 | 0.00299 | 0.01347 | 0.01049 |
| RPLPM | 0.57389 | 0.57488 | -0.00099 | 0.00289 | 0.39804 | 0.00299 | 0.00411 | 0.00113 |
| | | | | | | | | |
| CF | 0.57342 | 0.57313 | 0.00030 | 0.00076 | 0.80173 | 0.00298 | 0.00170 | 0.00128 |
| SVM | 0.57377 | 0.57250 | 0.00128 | 0.00840 | 0.15958 | 0.00297 | 0.00949 | 0.00658 |
| GLM | 0.57360 | 0.58066 | -0.00706 | 0.00784 | 0.05682 | 0.00298 | 0.00458 | 0.00161 |
| RPLPM | 0.57360 | 0.57427 | -0.00068 | 0.00149 | 0.56586 | 0.00298 | 0.00275 | 0.00023 |
| | | | | | | | | |
| CF | 0.57376 | 0.57305 | 0.00071 | 0.00048 | 0.85628 | 0.00298 | 0.00225 | 0.00074 |
| SVM | 0.57357 | 0.57191 | 0.00166 | 0.00739 | 0.18452 | 0.00298 | 0.00870 | 0.00571 |
| GLM | 0.57374 | 0.58067 | -0.00693 | 0.00527 | 0.04826 | 0.00298 | 0.00226 | 0.00071 |
| RPLPM | 0.57374 | 0.57397 | -0.00022 | 0.00117 | 0.62758 | 0.00298 | 0.00241 | 0.00056 |

**Table A.6**
*Performance evaluation in estimating the average treatment effect for the fourth scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study C.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.57437 | 0.57361 | 0.00076 | 0.00262 | 0.48258 | 0.00297 | 0.00022 | 0.00275 |
| SVM | 0.57373 | 0.55361 | 0.02012 | 0.01053 | 0.11343 | 0.00298 | 0.01069 | 0.00771 |
| GLM | 0.57406 | 0.58265 | -0.00859 | 0.01400 | 0.04523 | 0.00300 | 0.00999 | 0.00700 |
| RPLPM | 0.57406 | 0.57371 | 0.00035 | 0.00283 | 0.30206 | 0.00300 | 0.00286 | 0.00014 |
| CF | 0.57381 | 0.57232 | 0.00148 | 0.00228 | 0.31098 | 0.00298 | 0.00031 | 0.00267 |
| SVM | 0.57371 | 0.56398 | 0.00975 | 0.00902 | 0.26041 | 0.00298 | 0.00937 | 0.00640 |
| GLM | 0.57376 | 0.58214 | -0.00838 | 0.00661 | 0.04975 | 0.00298 | 0.00330 | 0.00032 |
| RPLPM | 0.57376 | 0.57372 | 0.00004 | 0.00211 | 0.36073 | 0.00298 | 0.00040 | 0.00258 |
| CF | 0.57382 | 0.57346 | 0.00036 | 0.00211 | 0.36072 | 0.00298 | 0.00040 | 0.00258 |
| SVM | 0.57374 | 0.56938 | 0.00436 | 0.00809 | 0.13533 | 0.00298 | 0.00877 | 0.00580 |
| GLM | 0.57389 | 0.58156 | -0.00771 | 0.00474 | 0.04418 | 0.00297 | 0.00152 | 0.00146 |
| RPLPM | 0.57389 | 0.57454 | -0.00065 | 0.00173 | 0.43663 | 0.00297 | 0.00165 | 0.00132 |

**Table A.7**
*Performance evaluation in estimating the average treatment effect for the fifth scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study A.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.76685 | 0.76465 | 0.0022 | 0.00399 | 0.59969 | 0.00782 | 0.00206 | 0.00576 |
| SVM | 0.76756 | 0.75884 | 0.00872 | 0.01046 | 0.29279 | 0.00773 | 0.01321 | 0.00548 |
| GLM | 0.76734 | 0.74707 | 0.02027 | 0.01453 | 0.10587 | 0.00781 | 0.00556 | 0.00225 |
| RPLPM | 0.76734 | 0.76600 | 0.00134 | 0.00386 | 0.55006 | 0.00781 | 0.00578 | 0.00203 |
| CF | 0.76788 | 0.76674 | 0.00115 | 0.00269 | 0.69299 | 0.00774 | 0.00351 | 0.00423 |
| SVM | 0.76735 | 0.76295 | 0.00440 | 0.00741 | 0.38706 | 0.00778 | 0.01088 | -0.00310 |
| GLM | 0.76756 | 0.75125 | 0.01631 | 0.00999 | 0.09997 | 0.00778 | 0.00176 | 0.00603 |
| RPLPM | 0.76756 | 0.76756 | -0.000005 | 0.00317 | 0.60393 | 0.00778 | 0.00528 | 0.00250 |
| CF | 0.76708 | 0.76664 | 0.00044 | 0.00205 | 0.75507 | 0.00778 | 0.00449 | 0.00329 |
| SVM | 0.76754 | 0.76483 | 0.00271 | 0.00599 | 0.43708 | 0.00772 | 0.00947 | 0.00175 |
| GLM | 0.76726 | 0.75118 | 0.01609 | 0.00919 | 0.09437 | 0.00775 | 0.00081 | 0.00694 |
| RPLPM | 0.76726 | 0.76793 | -0.00066 | 0.00297 | 0.62446 | 0.00775 | 0.00506 | 0.00269 |

**Table A.8**
*Performance evaluation in estimating the average treatment effect for the fifth scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study B.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.94063 | 0.93943 | 0.00121 | 0.00249 | 0.17067 | 0.00276 | 0.00021 | 0.00255 |
| SVM | 0.94101 | 0.92642 | 0.01459 | 0.01129 | 0.14348 | 0.00269 | 0.01259 | 0.00990 |
| GLM | 0.94059 | 0.89960 | 0.04098 | 0.01234 | 0.03195 | 0.00274 | 0.00751 | 0.00477 |
| RPLPM | 0.94059 | 0.94087 | -0.00029 | 0.00271 | 0.37902 | 0.00274 | 0.00380 | 0.00106 |
| CF | 0.94290 | 0.93848 | 0.00238 | 0.00209 | 0.30737 | 0.00271 | 0.00029 | 0.00242 |
| SVM | 0.94071 | 0.93573 | 0.00498 | 0.00832 | 0.09152 | 0.00274 | 0.01018 | 0.00744 |
| GLM | 0.94079 | 0.89990 | 0.04089 | 0.00657 | 0.04777 | 0.00274 | 0.00200 | 0.00074 |
| RPLPM | 0.94079 | 0.94064 | 0.00015 | 0.00144 | 0.54599 | 0.00274 | 0.00249 | 0.00025 |
| CF | 0.94082 | 0.93860 | 0.00222 | 0.00187 | 0.41433 | 0.00273 | 0.00035 | 0.00238 |
| SVM | 0.94066 | 0.93779 | 0.00287 | 0.00675 | 0.23025 | 0.00275 | 0.00861 | 0.00586 |
| GLM | 0.94060 | 0.89991 | 0.04089 | 0.00535 | 0.05662 | 0.00274 | 0.00094 | 0.00181 |
| RPLPM | 0.94080 | 0.94053 | 0.00026 | 0.00111 | 0.61846 | 0.00274 | 0.00222 | 0.00052 |

**Table A.9**
*Performance evaluation in estimating the average treatment effect for the fifth scenario over different sample sizes* $(N = 500, 1500, 3000)$ *in simulation study B.*

| METHOD | Pr($\tau$) | Pr(ATE) | BAIS | MSE | $R^2$ | $\sigma^2_{pr(\tau)}$ | $\sigma^2_{ITE}$ | $\lvert\sigma^2_{pr(\tau)}$ - $\sigma^2_{ITE}\rvert$ |
|---|---|---|---|---|---|---|---|---|
| CF | 0.94083 | 0.94014 | 0.00069 | 0.00239 | 0.19589 | 0.00272 | 0.00022 | 0.00250 |
| SVM | 0.94076 | 0.91387 | 0.02688 | 0.01206 | 0.08820 | 0.00274 | 0.01175 | 0.00901 |
| GLM | 0.94079 | 0.90266 | 0.03813 | 0.00865 | 0.04324 | 0.00275 | 0.00400 | 0.00125 |
| RPLPM | 0.94079 | 0.94046 | 0.00034 | 0.00275 | 0.28613 | 0.00275 | 0.00281 | 0.00006 |
| CF | 0.94078 | 0.93863 | 0.00215 | 0.00266 | 0.31434 | 0.00274 | 0.00032 | 0.00241 |
| SVM | 0.94054 | 0.92918 | 0.01136 | 0.00891 | 0.13169 | 0.00278 | 0.00967 | 0.00688 |
| GLM | 0.94087 | 0.89994 | 0.04093 | 0.005834 | 0.03816 | 0.00274 | 0.00130 | 0.00144 |
| RPLPM | 0.94087 | 0.94082 | 0.00005 | 0.00183 | 0.38146 | 0.00274 | 0.00174 | 0.00100 |
| CF | 0.94082 | 0.93971 | 0.00111 | 0.00191 | 0.36846 | 0.00271 | 0.00036 | 0.00235 |
| SVM | 0.94081 | 0.93551 | 0.00520 | 0.00762 | 0.26595 | 0.00272 | 0.00855 | 0.00583 |
| GLM | 0.94085 | 0.90074 | 0.04011 | 0.00503 | 0.04179 | 0.00272 | 0.00061 | 0.00210 |
| RPLPM | 0.94085 | 0.94164 | -0.00079 | 0.00166 | 0.40932 | 0.00272 | 0.00146 | 0.00126 |

**Table A.10**

*Comparison of Method Performance with Different PCA Applied to Different Scenarios and Sample Sizes*

| Sample Size | Scenario | Number of Variables | PCA Applied | CF | SVM | GLM | RPLPM |
|---|---|---|---|---|---|---|---|
| 500 | Scenario 1 | 7 | No | 0.00153 | 0.00717 | 0.00182 | 0.00117 |
| 500 | Scenario 1 | 15 | No | 0.00108 | 0.00912 | -0.00166 | -0.00048 |
| 500 | Scenario 1 | 15 | Yes | 0.00049 | 0.01927 | -0.00096 | -0.00031 |
| 500 | Scenario 2 | 7 | No | 0.00168 | 0.00739 | 0.00384 | 0.00117 |
| 500 | Scenario 2 | 15 | No | 0.00129 | 0.01008 | 0.00145 | -0.00048 |
| 500 | Scenario 2 | 15 | Yes | 0.00052 | 0.02044 | 0.00011 | 0.00031 |
| 500 | Scenario 3 | 7 | No | 0.00156 | 0.00700 | -0.00454 | 0.00128 |
| 500 | Scenario 3 | 15 | No | 0.00141 | 0.01181 | 0.00892 | -0.00073 |
| 500 | Scenario 3 | 15 | Yes | 0.00043 | 0.02450 | 0.00698 | 0.00024 |
| 500 | Scenario 4 | 7 | No | 0.00165 | 0.0046 | -0.00560 | -0.00044 |
| 500 | Scenario 4 | 15 | No | -0.00022 | 0.00770 | -0.00324 | -0.00099 |
| 500 | Scenario 4 | 15 | Yes | 0.00076 | 0.02012 | -0.00859 | 0.00035 |
| 500 | Scenario 5 | 7 | No | 0.00220 | 0.00872 | 0.02027 | 0.00134 |
| 500 | Scenario 5 | 15 | No | 0.00121 | 0.01459 | 0.04098 | -0.00029 |
| 500 | Scenario 5 | 15 | Yes | 0.00069 | 0.02688 | 0.03813 | 0.00034 |
| 1500 | Scenario 1 | 7 | No | 0.00052 | 0.00424 | -0.00216 | -0.00012 |
| 1500 | Scenario 1 | 15 | No | 0.00141 | 0.00366 | -0.00400 | 0.00005 |
| 1500 | Scenario 1 | 15 | Yes | 0.00142 | 0.00961 | -0.00422 | 0.00010 |
| 1500 | Scenario 2 | 7 | No | 0.00062 | 0.00400 | 0.00215 | -0.00012 |
| 1500 | Scenario 2 | 15 | No | 0.00160 | 0.00368 | 0.00020 | 0.00005 |
| 1500 | Scenario 2 | 15 | Yes | 0.00156 | 0.00990 | 0.00126 | 0.00010 |
| 1500 | Scenario 3 | 7 | No | 0.00069 | 0.00409 | -0.00739 | -0.00008 |
| 1500 | Scenario 3 | 15 | No | 0.00183 | 0.00411 | 0.01067 | 0.00007 |
| 1500 | Scenario 3 | 15 | Yes | 0.00213 | 0.01035 | 0.01104 | -0.00008 |
| 1500 | Scenario 4 | 7 | No | -0.00716 | 0.00322 | -0.00965 | -0.00117 |
| 1500 | Scenario 4 | 15 | No | 0.00030 | 0.00128 | -0.00706 | -0.00068 |
| 1500 | Scenario 4 | 15 | Yes | 0.00148 | 0.00975 | -0.00838 | 0.00004 |
| 1500 | Scenario 5 | 7 | No | 0.00115 | 0.00440 | 0.01631 | -0.000005 |
| 1500 | Scenario 5 | 15 | No | 0.00238 | 0.00498 | 0.04089 | 0.00015 |

| Sample Size | Scenario | Number of Variables | PCA Applied | CF | SVM | GLM | RPLPM |
|---|---|---|---|---|---|---|---|
| 1500 | Scenario 5 | 15 | Yes | 0.00215 | 0.01136 | 0.04093 | 0.00005 |
| 3000 | Scenario 1 | 7 | No | 0.00001 | 0.00182 | -0.00240 | -0.00073 |
| 3000 | Scenario 1 | 15 | No | 0.00137 | 0.00248 | -0.00026 | 0.00016 |
| 3000 | Scenario 1 | 15 | Yes | 0.00040 | 0.00414 | -0.00218 | -0.00069 |
| 3000 | Scenario 2 | 7 | No | 0.00012 | 0.00197 | 0.00073 | -0.00073 |
| 3000 | Scenario 2 | 15 | No | 0.00146 | 0.00242 | 0.00215 | 0.00016 |
| 3000 | Scenario 2 | 15 | Yes | 0.00065 | 0.00438 | 0.00116 | -0.00069 |
| 3000 | Scenario 3 | 7 | No | 0.00004 | 0.00208 | -0.00857 | 0.00064 |
| 3000 | Scenario 3 | 15 | No | 0.00192 | 0.00277 | 0.00913 | 0.00025 |
| 3000 | Scenario 3 | 15 | Yes | 0.00065 | 0.00477 | 0.00852 | -0.00059 |
| 3000 | Scenario 4 | 7 | No | -0.00031 | -0.00002 | 0.00750 | -0.00072 |
| 3000 | Scenario 4 | 15 | No | 0.00071 | 0.00166 | -0.00693 | -0.00022 |
| 3000 | Scenario 4 | 15 | Yes | 0.00036 | 0.00436 | -0.00771 | -0.00065 |
| 3000 | Scenario 5 | 7 | No | 0.00044 | 0.00271 | 0.01609 | -0.00066 |
| 3000 | Scenario 5 | 15 | No | 0.00222 | 0.00287 | 0.04089 | 0.00026 |
| 3000 | Scenario 5 | 15 | Yes | 0.00111 | 0.00520 | 0.04011 | -0.00079 |

جامعة النجاح الوطنية
كلية الدراسات العليا

# استكشاف قدرة وأداء طرق التعلم بالاشراف لتصنيف التسميات في الاستدلال السببي: دراسة مقارنة

إعداد
علا محمد لطفي ابو صقر

إشراف
د. عبد الرحمن عيد

قدمت هذه الرسالة استكمالا لمتطلبات الحصول على درجة الماجستير في الرياضيات،
من كلية الدراسات العليا، في جامعة النجاح الوطنية، نابلس، فلسطين.

2024

# استكشاف قدرة وأداء طرق التعلم بالاشراف لتصنيف التسميات في الاستدلال السببي: دراسة مقارنة

**إعداد**

**علا محمد لطفي ابو صقر**

**إشراف**

**د. عبد الرحمن عيد**

## الملخص

تنتشر المناقشات حول التعلم الآلي بشكل متزايد بسبب دقته في التنبؤ وقدرته على التعامل مع كميات هائلة من البيانات. علاوة على ذلك، فإن العديد من العلاقات في الحياة هي علاقات سببية، مما يحفز الجهود المبذولة لفهم علاقات السبب والنتيجة بين المتغيرات. على سبيل المثال، يصبح فهم مدى تأثير دواء معين على فرد مصاب بمرض ما أمرًا بالغ الأهمية. وعلى الرغم من أن الأمر قد يبدو واضحًا للوهلة الأولى، إلا أن الفحص الأعمق يوضح التعقيد الكامن في مثل هذه المساعي عند استخدام التعلم الآلي في السببية. لقد قدمت أساليب التعلم الآلي مساهمة قيمة في مجال الاستدلال السببي، ومع ذلك لا يزال هناك بحث في تقدير التأثير السببي عندما يكون كل من العلاج والنتيجة متغيرين ثنائيين، لأن التعلم الآلي أثبت قدرته على التنبؤ، والتنبؤ لا يعني السببية. ولعل هذا هو التحدي الذي يواجه التعلم الآلي في الحصول على تقديرات أكثر دقة وأقل تحيزًا للآثار السببية.

تُجري هذه الدراسة تحليلاً مقارنًا لطرق التعلّم تحت الإشراف لتصنيف التسميات في الاستدلال السببي. نقوم بتقييم أداء وقدرة أربع تقنيات: الغابة السببية (CF)، وآلة دعم المتجهات (SVM)، والنماذج الخطية المعممة (GLM)، ونماذج الاحتمالات الخطية (LPM) في تقدير التأثيرات السببية لمتغير الاستجابة الفئوية. تم إجراء تجارب محاكاة عشوائية مضبوطة وتجربة حقيقيةلتقييم أداء الأساليب في ظل ظروف متفاوتة، من خلال تغيير الخصائص الرئيسية للبيانات بما في ذلك حجم العينة وعدد المتغيرات التفسيرية.

توفر النتائج رؤى قيمة حول نقاط القوة والقيود الخاصة بكل طريقة في كل سيناريو في دراسة محاكاة التأثير السببي. وعلاوة على ذلك، فإن الأساليب قادرة على اكتشاف عدم التجانس في نتائج البيانات الحقيقية، وكان من المتوقع أن تكتشف (GLM) , (SVM) و ( LPM)عدم التجانس أكثر من الغابة السببية.

لقد ركزنا على هذه الأساليب الأربعة بسبب مزاياها المحددة: تُعد الغابات السببية بارعة بشكل خاص في إجراء استنتاجات سببية بسهولة؛ وآلات المتجهات الداعمة معروفة بفعاليتها في مهام التصنيف الثنائي؛ والنماذج الخطية المعممة راسخة باعتبارها الأمثل لنمذجة متغير الاستجابة الثنائية؛ ونماذج الاحتمالات الخطية لقدرتها على تقديم تنبؤات في صورة احتمالات.

تساعدنا هذه الأطروحة على تحسين معرفتنا بتقنيات التعلم الآلي في الاستدلال السببي وتؤكد على أهمية تقييم أدائها بعناية في التطبيقات الواقعية.

**الكلمات المفتاحية:** الغابة السببية، LPM، GLM، SVM، التأثير السببي.