# CAUSAL INFERENCE USING MODERN MACHINE LEARNING

February 24, 2022

Registration number: 2106373
Project: Causal Inference
Link to GitHub: https://github.com/kkoban/Data-Science-and-Decision-Making/tree/main/Project

| | |
|---|---|
| Executive summary (max. 250 words) | 234 |
| Introduction (max. 600 words) | 114 |
| Data (max. 500 words/dataset) | 590 |
| Methodology (max. 600 words) | 413 |
| Conclusions (max. 500 words) | 68 |
| Total word count | Your word count |

## Contents


**1 Introduction** **2**

**2 Data** **2**
2.1 IHDP . . . . . 2
2.2 JOBS . . . . . 5

**3 Methodology** **5**
3.1 IHDP . . . . . 5
3.2 JOBS . . . . . 7

**4 Conclusions** **9**

**Abstract**

We are interested in understanding how interventions effect a system given only observed (factual) outcomes. We realise that Machine learning focuses more on predicting outcomes rather than understanding causation. Thus, after building a machine learning model, causal inference allows to draw conclusions about the effect of interventions on a system. Simply, the features in our data affect the outcome variables. Causal Inference is the answer to our question of why a particular treatment affects the outcomes and gives us relevant tools to understand the said change. Hence, we can say that Machine Learning and Causal inference can be both mutually beneficial. In the following research we will look at two different datasets with causal effects in them. The Infant health Development Program (IHDP) data is concerned with calculating the cognitive test scores of infants against family support as the causal variable using Metalearners and Machine Learning regressor models as their base. The observed ITE obtained with access to 'alternate reality' (counterfactuals) results is compared against true ITEs for the evaluation of CI estimators. The JOBS data on the other hand is an example of classic binary classification problem where we predict the effect of National Supported Work Program (NSWP) job training on the employment status of people. This data is a combination of experimental and observational outcomes and is used to calculate the risk assessment with the help of ATT or Policy Risk.

# 1 Introduction

Causal inference is study of causation or causal relation in data that help with decision making. It helps identify the causes behind processes taking place in a system. Approximation of the causal effects through observed outcomes as well as finding answers to causal question that are not possible with Randomised Controlled Trials motivates us to use ML techniques in order to approximate accurate from given data. Causal Inference is similar to the classic supervised Machine Learning approach and thus we can say that CI and ML go hand in hand. Thus, the understanding of causality in data help the Modern ML techniques reason out cause and effect relations and build more human-like intelligence.

# 2 Data

## 2.1 IHDP

The Infant health Development Program (IHDP) data introduced in [1] based on a clinical trial [2] and provided by Dr Ana Matran-Fernandez is a simulated dataset used to examine the effect of high-quality childcare and home visits on the future cognitive test score of low-birth-weight of premature infants. There is a total of 747 rows and 29 columns in the dataset out of which 25 of them i.e., x1, x2, ..., x25 are predictor variables (X), both numerical and categorical, representing child measurements such as child-birth weight, head circumference, neonatal health index, etc. and information of the mother at the time of birth such as marital status, education level and if there are any past behavioural records of smoking, drinking or drugs. The treatment variable (t) indicates whether family support was part of the control or not, t = 1 meaning support was given and vice-versa. The factual (yf) and counterfactual (ycf) outcome variables i.e., cognitive test score of infants are simulated from real pre-treatment covariates. The dataset also contains true individualised effects (ite) per each data unit; hence the data is fit for the purpose of performance evaluation. Since, the outcome variable is quantitative and continuous in nature hence we have a regression problem of predicting a quantity at hand.

Plotting boxplots demonstrates that the features vary in scale which can dealt with in the data pre-processing stage, as seen in Figure 1.

The histogram 2 brings forth some important information about the dataset; the features from x1-x6 are numerical whereas x7-x25 are categorical in nature and the ite graph reflects the presence of heterogeneous effect groups.

From the scatterplot 3, x1 and x2 have a strong positive linear relationship. Both x1 and x3 and, x2 and x3 have a moderate negative linear relationship in between them.

The heatmap 4 for the correlation coefficient of the feature values proves that there is a strong positive correlation of 0.85 between x1 and x2 and a strong negative correlation of -0.82 between x4 and x14.

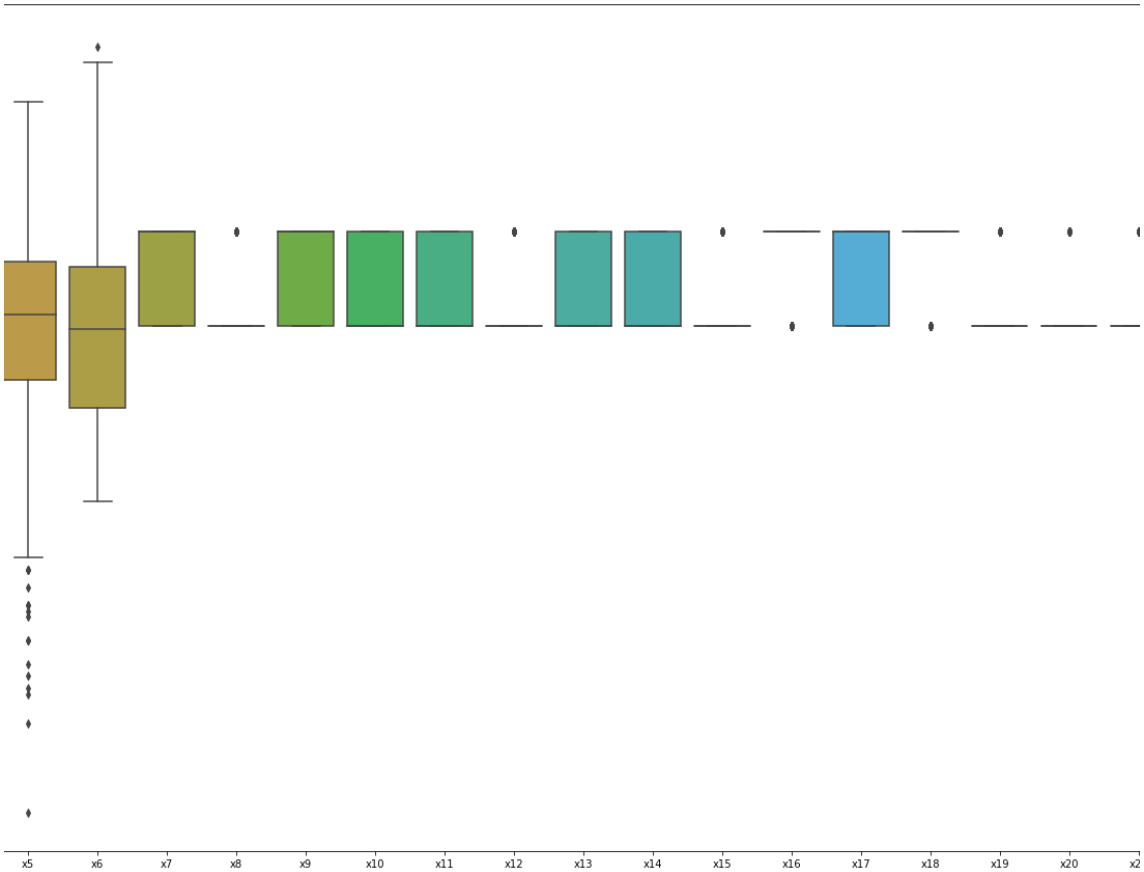


Figure 1: Boxplot

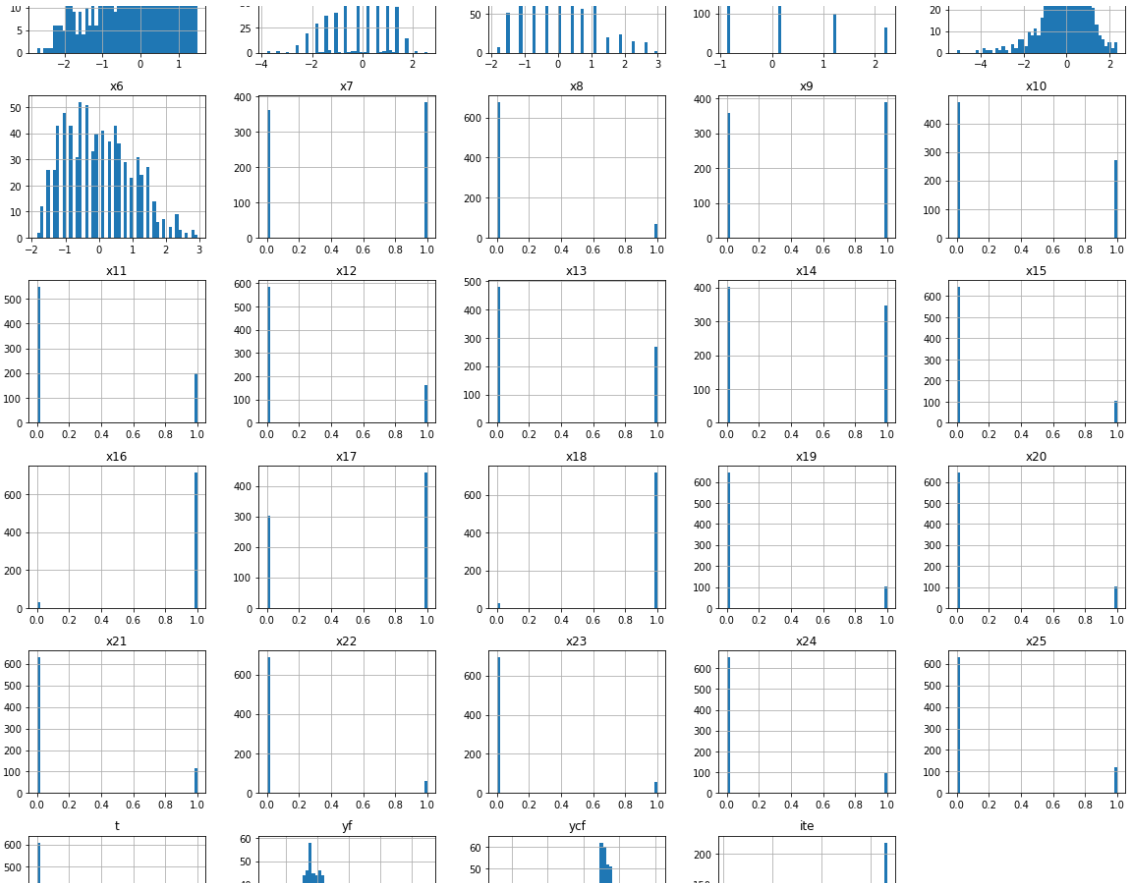


Figure 2: Histogram

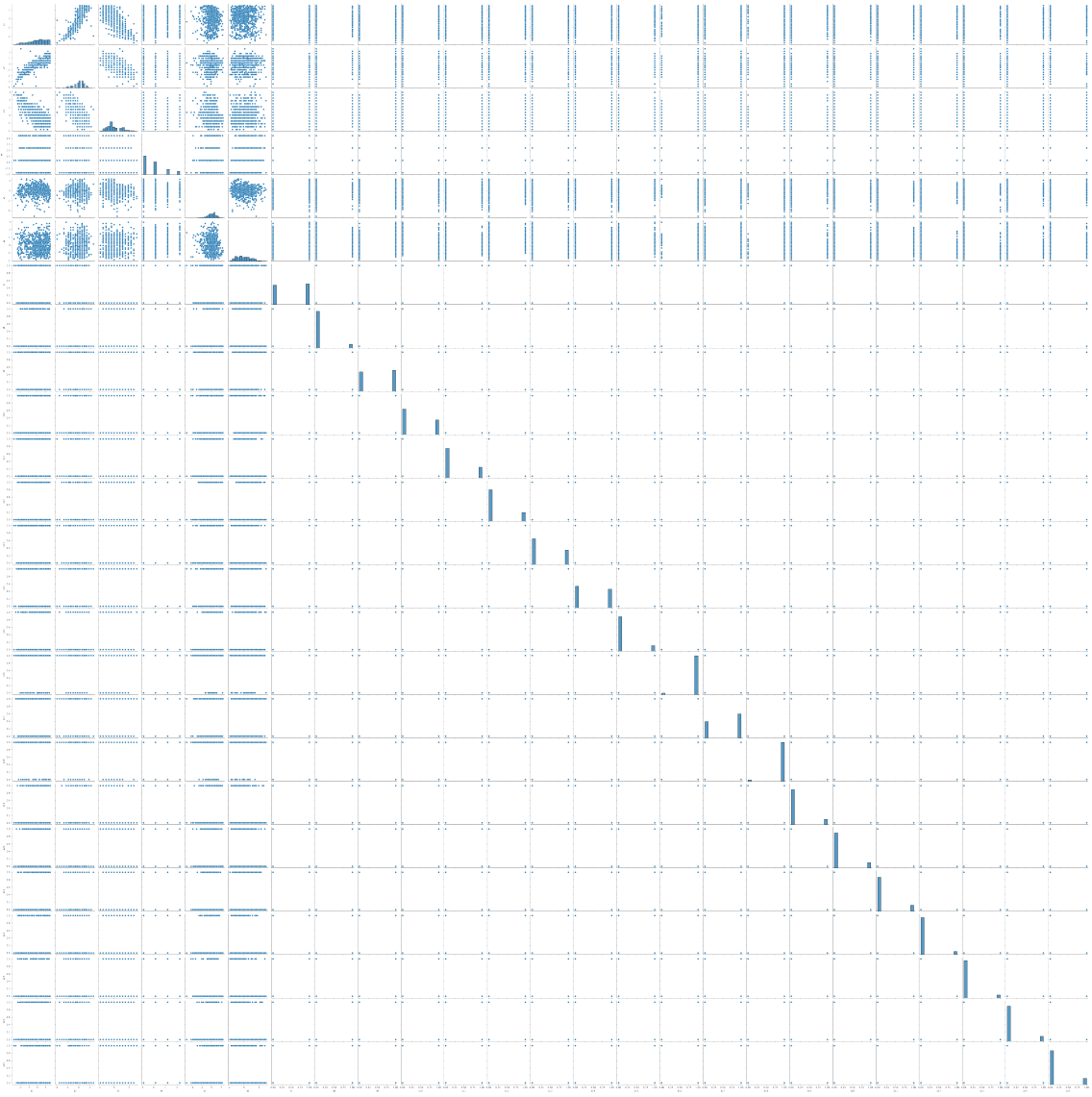

Figure 3: Scatterplot

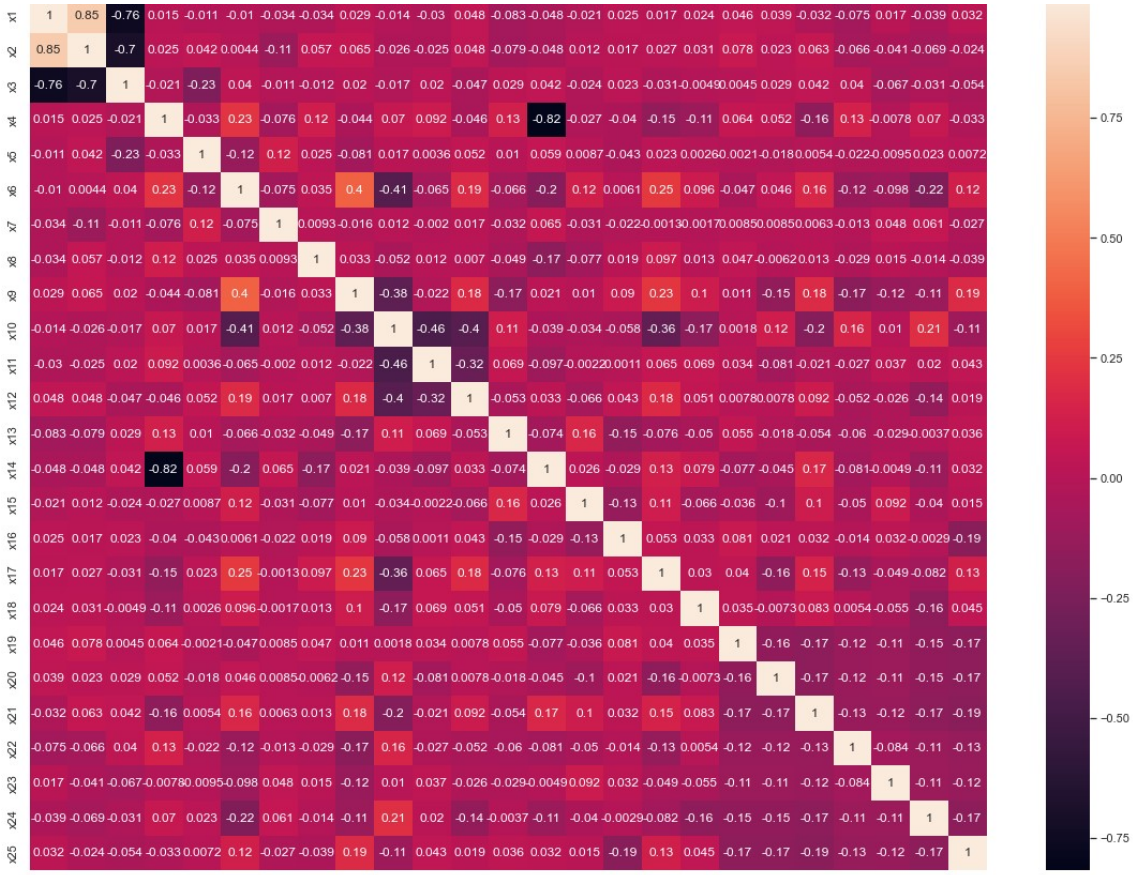


Figure 4: Heatmap

### 2.2 JOBS

The JOBS dataset is a combination of experimental and combinational data from [3] as a part of National Supported Work Program (NSWP) and Panel Study of Income Dynamics (PSID) [4] respectively. This dataset was provided by Dr Ana Matran-Fernandez as well. It is closer to real-life conditions as it provides with the employment status of an individual. The data consists a total of 3212 samples and 20 columns. There are 17 predictor variables (X) that captures people's basic characteristics about their circumstances. The treatment variable (t) represents if the individual received job training from the NSWP when t = 1 and vice-versa when t = 0. The response variable (y) records the employment status. The final column (e) gives information about whether a sample comes from experimental or observational data which is used for calculating specific performance metrics. The outcome variable is categorical; therefore, it is a classification problem where the label of the employment status needs to be predicted.

Plotting boxplot 5 reveals that all the features have almost similar scale.

The histogram 6 mainly brings forth that the data is highly imbalanced which is the case in most causal inference datasets.

In Figure 7 bout 85 percent of the samples are employed and 15 percent have an employment status of unemployed.

Many features are observed to have strong positive linear relationship amongst themselves in Figure 8.

The heatmap 9 points out the most positive correlation of 0.99 between x1 and x9 and, x9 and x10. On the other hand, x2 and x6 exhibit a negative correlation of -0.75.

## 3 Methodology

### 3.1 IHDP

In the IHDP dataset as the counterfactuals are provided, the causal question is given as:- How does family support (treatment) effect cognitive test score of infants (outcome)? In addition, as all the outcomes are simulated, there can be a further evaluation of CI estimators.

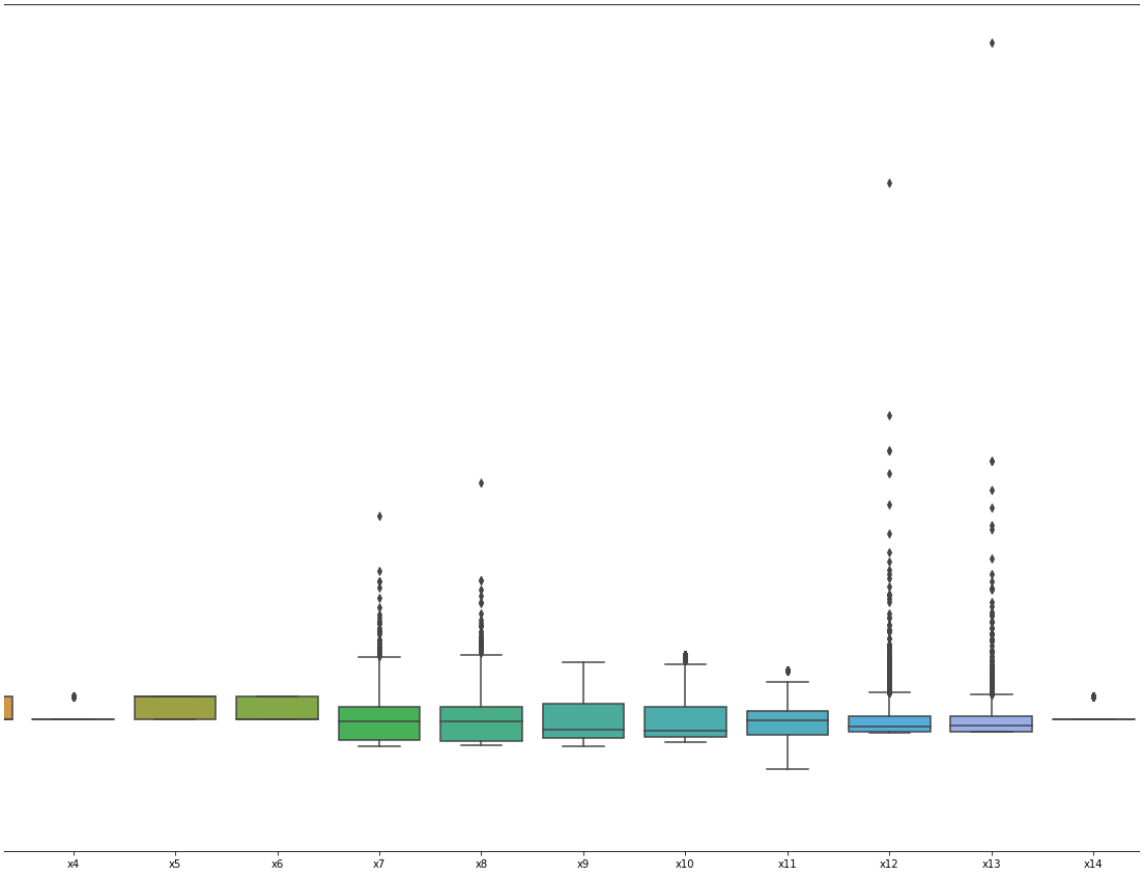


Figure 5: Boxplot

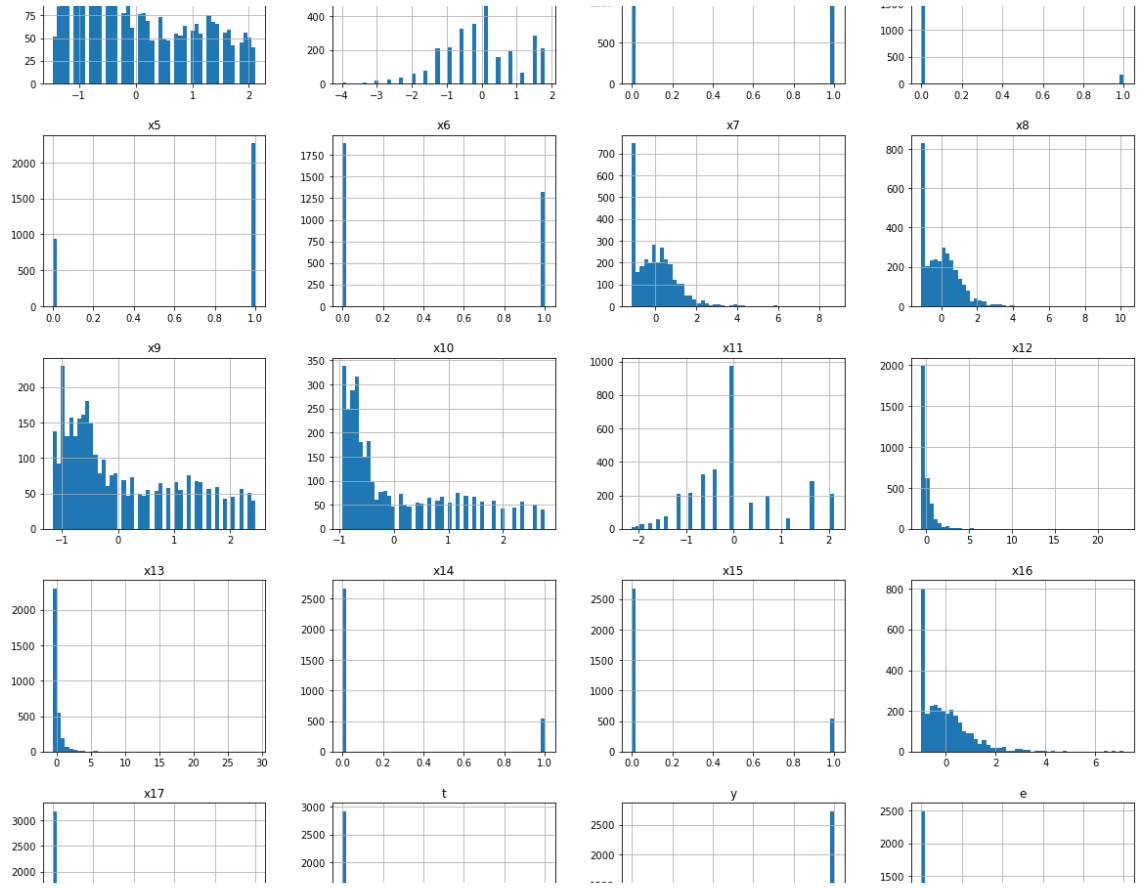


Figure 6: Histogram

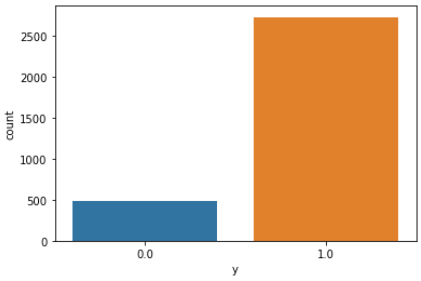


Figure 7: y

The data is split into two parts, namely, train and test data. It is also scaled during Data Preprocessing as there was variation in the feature scales. The train dataset comprises of 80 percent of the data and is used for training the model, whereas the test dataset is only 20 percent of the data on which the model built is tested. For estimating heterogeneous treatment effects using machine learning, the performance of S and T-learners are reviewed and then X-learner is approached which takes advantage of the unbalanced dataset. Moreover, a Random Forest regressor is preferably used as the data is not entirely linear, best to deal with noisy dataset and is powerful than decision trees. Hence, to solve the regression problem at hand, Random Forest Classifier is used by providing the X-learner with RF regressor as base learner. Then, to calculate the treatment effect the ITE i.e., the difference between the treated and controlled outcomes is calculated. In this case, for metric evaluation i.e., to check the accuracy of predicted effects the ATE test or PEHE test is performed on true ITEs and predicted ITEs to get the error measurements.

## 3.2 JOBS

As the JOBS dataset is mixed with Randomised Controlled Trials (RCTs), the causal question is given as:- What is the effect of job training received from NSWP (treatment) on individual employment status (outcome)? The data is randomly split into two parts, namely, train and test data. The train dataset comprises of 80 percent of the data and is used for training the model, whereas the test dataset is only 20 percent of the data on which the model built is tested. Due to heavy imbalance in the dataset, more sophisticated learner such as IPW is chosen to predict propensity scores for each individual. Decision Tree Classifiers is preferred over Logistic Regression Classifier to solve the binary problem of determining the employment status of each individual as LR constructs linear boundaries and it requires no multicollinearity. Hence, Decision Tree Classifier with Inverse Propensity Weighting is chosen to solve the classic binary classification problem. In the absence of counterfactuals, for performance metrics the error on ATT or Policy Risk is used for calculating the risk assessment.

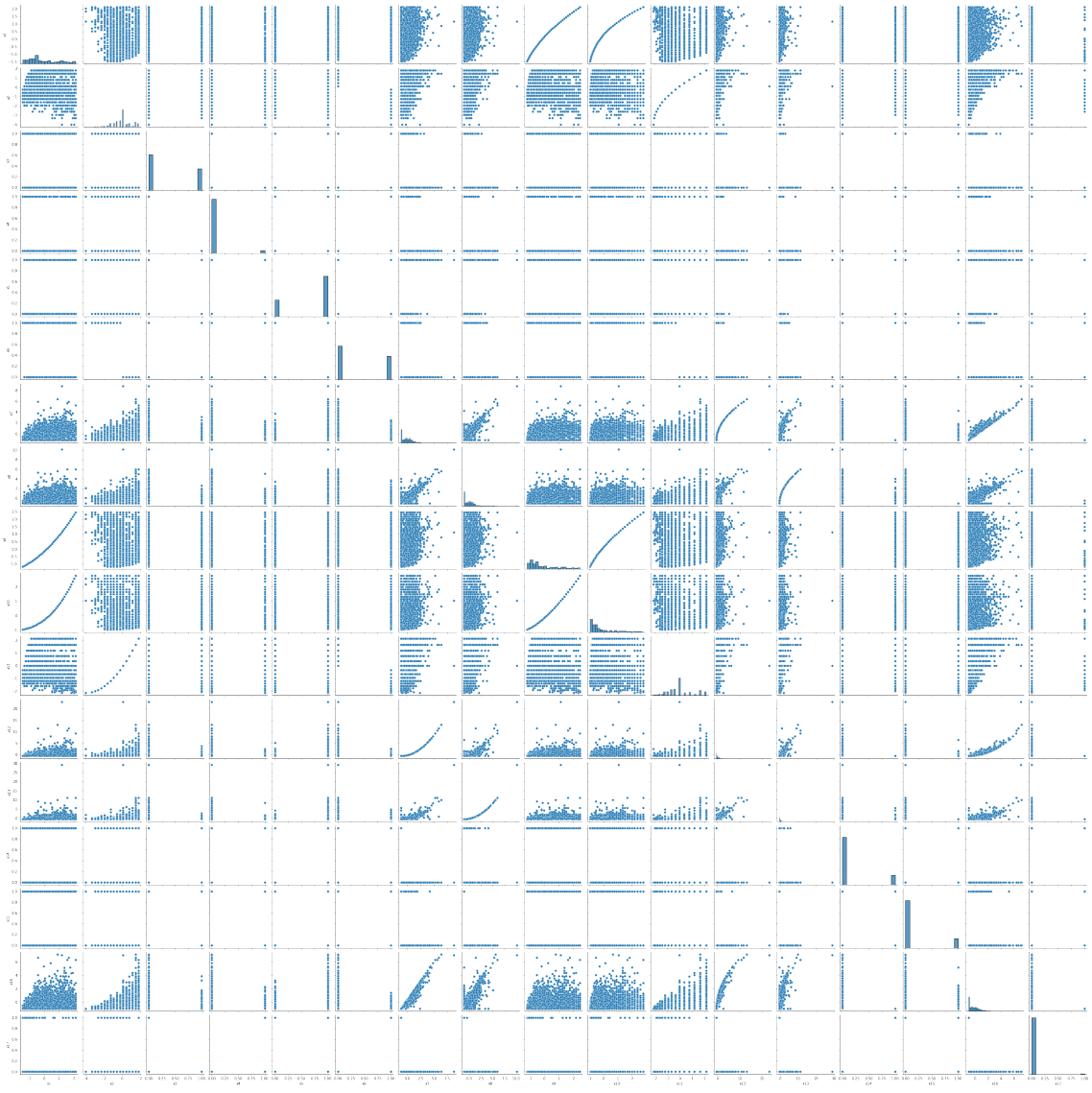

Figure 8: Scatterplot

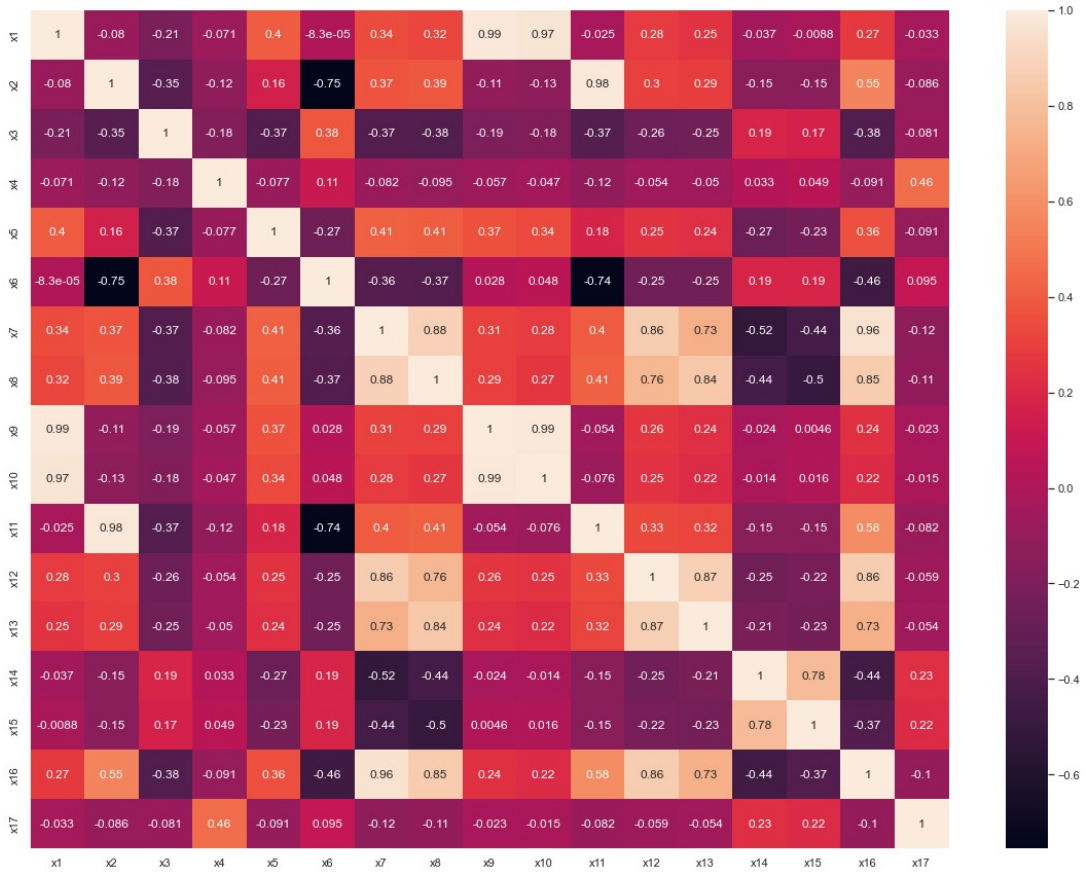


Figure 9: Heatmap

# 4 Conclusions

In this paper, we discuss about dealing with causality in datasets both in presence and absence of counterfactuals. Through data exploration and visualisation, we develop an appropriate methodology and decide on relevant modern machine learning algorithms and CI estimators to work with. I believe with proper training and testing we can make a good model for measuring the effect of treatments on the outcomes in the dataset.

# References

[1] J. L. Hill. Bayesian Nonparametric Modeling for Causal Inference). *Journal of Computational and Graphical Statistics*, 20:217–240, 2011.

[2] P. K. K. J. Brooks-Gunn, F. R. Liaw. Effects of early intervention on cognitive function of low-birth-weight preterm infants. *The Journal of Pediatrics*, 120:350–359, 1992.

[3] R. J. LaLonde. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76(4):604–620, 1986.

[4] S. W. R. H. Dehejia. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002.