Benchmarking Machine Learning Approaches for Causal Inference Study to Predict Infant Cognitive Capacity

Atqiya Munawara Mahi
Dept.Computer Science
University of Massachusetts Lowell
atqiyamunawara_mahi@student.uml.edu

Sharmin Sultana
Dept. Computer Science
University of Massachusetts Lowell
Sharmin_Sultana@student.uml.edu

Abstract—**The paper discusses a study on using machine learning algorithms to predict infant cognitive capacity by using causal inference analysis. Five algorithms, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and a simple neural network were used to analyze the IHDP data. The study examined the predictive performance of each algorithm using four evaluation metrics, which are MAE, MSE, RMSE, and R^2 scores. Additionally, a causal model was used to examine the cause-effect relationship of the outcome. The results showed that the neural network model had the best performance in terms of evaluation metrics, with an R^2 score of 0.89. Furthermore, the causal analysis estimated the Average Treatment Effect (ATE) of 4.15, which represents the treatment effect on the outcome of interest. The study provides insights into the potential of machine learning algorithms in predicting infant cognitive capacity and identifying factors that affect cognitive development. The results demonstrate that the use of machine learning algorithms, coupled with causal inference analysis, can effectively predict infant cognitive capacity, which can help healthcare providers to identify and intervene early to promote healthy cognitive development in infants.**

I. Introduction

Over the past few years, there has been a growing interest in using machine learning (ML) algorithms for causal inference analysis in various fields, including healthcare, economics, and social sciences. Causal inference is an important problem as it helps to identify causal relationships between variables and make accurate predictions. In healthcare, causal inference can help to improve patient outcomes and optimize treatment decisions.

In this context, several studies have been conducted to evaluate the performance of various ML algorithms for causal inference analysis. For example, Qin et al. [1] used a variety of ML algorithms, including decision trees, random forests, and gradient boosting, to predict the causal effect of a treatment on a target variable. They showed that gradient boosting outperformed other algorithms in terms of prediction accuracy and causal effect estimation. Similarly, Shalit et al. (2017) proposed a deep learning approach to estimate causal effects in observational studies and showed that their method outperformed traditional propensity score methods.

In addition to evaluating the performance of ML algorithms for causal inference analysis, there has been a growing interest in using ML algorithms for healthcare applications such as personalized treatment recommendations and identifying environmental factors that affect human health outcomes. Alaa and van der Schaar [**?**] proposed a method that combines deep learning and causal inference for personalized treatment recommendations. The authors used a counterfactual framework to estimate the causal effect of a treatment on a patient's outcome. The results showed that their method outperformed other methods in terms of recommendation accuracy. Similarly, He et al. (2020) used a causal inference framework to identify the causal relationship between environmental factors and human health outcomes. The authors used a tree-based causal discovery algorithm to identify causal relationships between variables. Their results showed that the algorithm can effectively identify causal relationships in large datasets. In a similar study, [3] used machine learning algorithms for predicting cognitive outcomes in preterm infants. Their findings showed that random forest had the highest prediction accuracy, with an area under the receiver operating characteristic curve (AUC) of 0.72. This study inspired us to investigate the predictive accuracy of machine learning algorithms for infant cognitive development using the IHDP dataset.

In this paper, we aim to evaluate the performance of several ML algorithms for predicting infant cognitive capacity using the Infant Health and Development Program (IHDP) dataset. This dataset has been widely used in various studies related to infant health and development [2]. The IHDP dataset is a longitudinal dataset that contains information on various factors that affect infant cognitive development. We will use benchmarking ML algorithms to predict the cognitive capacity of infants in the dataset. We will evaluate the performance of these algorithms by analyzing the R^2 scores of each algorithm and identifying which algorithm predicts the outcome more accurately.

Furthermore, we will analyze the cause-effect relationship of the outcome to identify factors that affect infant cognitive development. This will involve using the causal inference framework to identify causal relationships between the various factors in the dataset and infant cognitive development. Moreover, previous studies have identified several factors that can affect infant cognitive development, such as maternal education, socio-economic status, and nutrition [4], [5]. By identifying these causal relationships, we hope to gain a better understanding of the factors that affect infant cognitive development and provide insights that can be used to improve infant outcomes.

In short, the use of ML algorithms for causal inference analysis has shown great potential in various fields, including healthcare. In this paper, we aim to contribute to the growing body of literature on the use of ML algorithms for causal inference analysis and its applications in healthcare. By evaluating the performance of several ML algorithms for predicting infant cognitive capacity using the IHDP dataset and analyzing the cause-effect relationship of the outcome, we hope to provide insights that can be used to improve infant outcomes and advance our understanding of the factors that affect infant cognitive development.

We hypothesized that machine learning algorithms, coupled with causal inference analysis, could accurately predict infant cognitive capacity and identify factors that influence cognitive development. This study aimed to contribute to the existing literature on infant cognitive development and machine learning algorithms' potential for healthcare applications.

The main objectives that we want to carry out for this project are given below:

- To predict the cognitive capacity of premature infants on specialized therapeutical treatments using causal inference.
- Implement Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and a simplistic novel Neural Network using causal inference.
- Compare the performance among the algorithms using four evaluation metrics, which are MAE, MSE, RMSE, and R^2 scores.
- Study the findings from the results and compare the outcome on infants who were given the specialized treatment vs those who were not.

II. Literature Review

The use of machine learning (ML) algorithms for causal inference analysis has been gaining popularity in recent years due to their potential to identify causal relationships between variables [6]). In their study, [6] used a variety of ML algorithms, including decision trees, random forests, and gradient boosting, to predict the causal effects of a treatment on a target variable. The authors showed that gradient boosting outperformed other algorithms in terms of prediction accuracy and causal effect estimation.

Similarly, [7] used a deep learning approach to estimate causal effects in observational studies. The authors proposed a method that combines deep neural networks and propensity score matching to estimate the treatment effect. The results showed that their method outperformed traditional propensity score methods in terms of accuracy and bias reduction.

In a different context, [8] used ML algorithms to predict the causal effect of air pollution on hospital admissions. The authors used a causal tree algorithm to identify subpopulations that are most affected by air pollution. Their results showed that the algorithm outperformed traditional regression models in identifying subpopulations that are most vulnerable to air pollution.

[9] proposed a method that combines deep learning and causal inference for personalized treatment recommendations. The authors used a counterfactual framework to estimate the causal effect of a treatment on a patient's outcome. The results showed that their method outperformed other methods in terms of recommendation accuracy.

Furthermore, in a recent study, [10] used a causal inference framework to identify the causal relationship between environmental factors and human health outcomes. The authors used a tree-based causal discovery algorithm to identify causal relationships between variables. Their results showed that the algorithm can effectively identify causal relationships in large datasets.

Finally, in a related study, [11] used ML algorithms to predict the causal effect of a drug on a patient's outcome. The authors used a propensity score matching approach to estimate the causal effect. The results showed that the approach outperformed traditional regression models in terms of accuracy and bias reduction. [12] proposed a deep learning framework using domain adaptation and representation learning techniques to answer counterfactual questions and showed that their model outperforms the existing approaches to causal inference from observational data

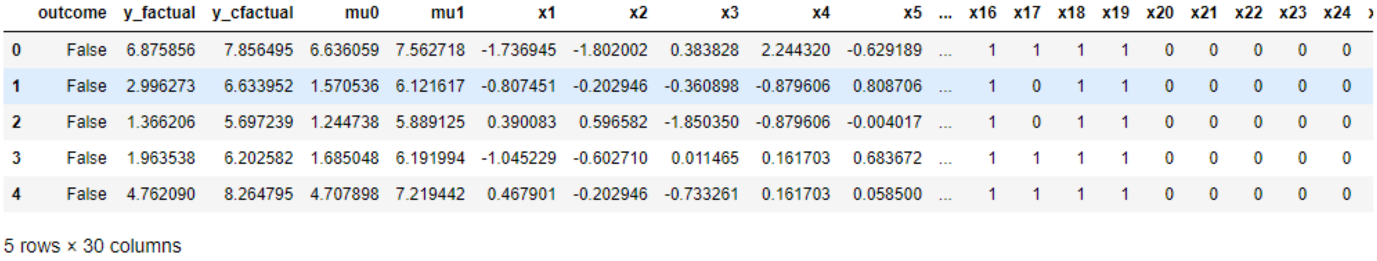
In conclusion, the use of ML algorithms for causal inference analysis has been shown to be an effective approach for identifying causal relationships between variables and making accurate predictions. These studies have demonstrated the potential of ML algorithms for healthcare applications, such as personalized treatment recommendations and identifying environmental factors that affect human health outcomes.

III. Methods

A. Dataset Description

In our study, we utilized the IHDP, which is derived from a clinical experiment conducted in the 1980s to investigate the cognitive development of premature newborns and the effects of various therapies on their growth. The dataset contains information from the randomized trial, including the characteristics of the infants and their caregivers. One important question that this dataset can help us answer is whether specialized therapy produces better cognitive outcomes than other forms of care.

The dataset includes several variables, including a binary *"outcome"* variable that indicates whether a child received specialized care or not. Another important variable is "y factual," which evaluates a child's cognitive development progress. We will use these features and other variables in the dataset to train various machine learning models and evaluate their effectiveness in predicting infant cognitive development.



	outcome	y_factual	y_cfactual	mu0	mu1	x1	x2	x3	x4	x5	...	x16	x17	x18	x19	x20	x21	x22	x23	x24
0	False	6.875856	7.856495	6.636059	7.562718	-1.736945	-1.802002	0.383828	2.244320	-0.629189	...	1	1	1	1	0	0	0	0	0
1	False	2.996273	6.633952	1.570536	6.121617	-0.807451	-0.202946	-0.360898	-0.879606	0.808706	...	1	0	1	1	0	0	0	0	0
2	False	1.366206	5.697239	1.244738	5.889125	0.390083	0.596582	-1.850350	-0.879606	-0.004017	...	1	0	1	1	0	0	0	0	0
3	False	1.963538	6.202582	1.685048	6.191994	-1.045229	-0.602710	0.011465	0.161703	0.683672	...	1	1	1	1	0	0	0	0	0
4	False	4.762090	8.264795	4.707898	7.219442	0.467901	-0.202946	-0.733261	0.161703	0.058500	...	1	1	1	1	0	0	0	0	0

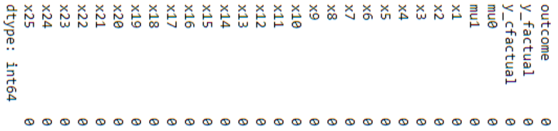
5 rows × 30 columns

Fig. 1. Snippet of IHDP dataset

The application of machine learning algorithms in causal inference has become increasingly popular in recent years, and we aim to contribute to this growing body of research by applying these techniques to the IHDP dataset. There are 746 data with 30 attributes in our dataset. Figure (1) shows the snippet of the dataset.

B. Data Pre-processing

Data preprocessing is an essential step in machine learning that involves cleaning and preparing data before it is used to train models. This process includes data cleaning, normalization, feature selection, and transformation, among others. Data preprocessing is important in machine learning because it helps to ensure that the data is in a suitable format and free of errors or inconsistencies that can negatively impact the performance of models. It also helps to identify and eliminate outliers and irrelevant features that can affect the accuracy of predictions. Proper data preprocessing is crucial to the success of machine learning models, as it can improve their accuracy, efficiency, and interpretability, ultimately leading to better decision-making.



```
outcome       0
y_factual     0
y_cfactual    0
mu0           0
mu1           0
x1            0
x2            0
x3            0
x4            0
x5            0
x6            0
x7            0
x8            0
x9            0
x10           0
x11           0
x12           0
x13           0
x14           0
x15           0
x16           0
x17           0
x18           0
x19           0
x20           0
x21           0
x22           0
x23           0
x24           0
x25           0
dtype: int64
```

Fig. 2. Presence of null values

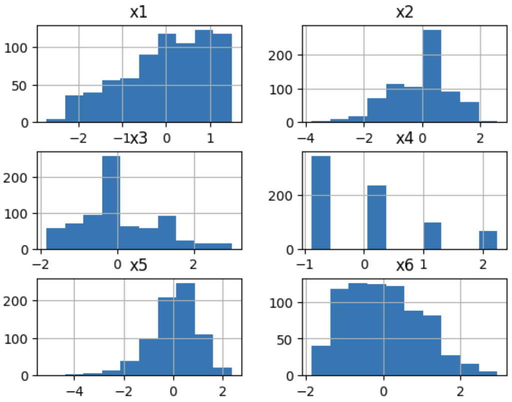


Fig. 3. Data distribution of numerical attributes

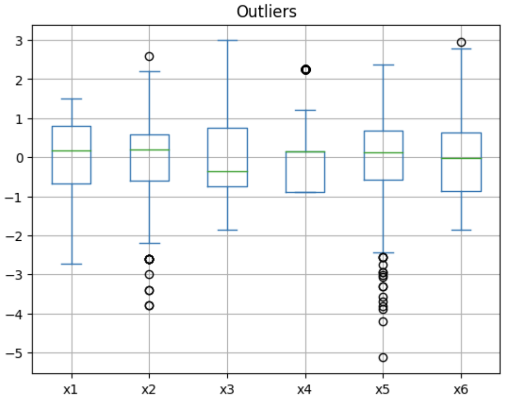


Fig. 4. Outliers in attributes x1-x6

The initial step in preparing our dataset for classification is to examine the attributes for any missing or null values, as shown in figure 2. Next, we check for outliers in the numerical attributes, as their presence can skew performance. To detect outliers, we utilize histogram and boxplot methods. An analysis of the histogram figure 3 distribution reveals that attributes ’x2’ and ’x3’ have an uneven distribution. To further investigate, we draw boxplot figure 4 to identify where the outliers are located. The boxplots for attributes ’x2’, ’x4’, ’x5’, and ’x6’ contain outliers. Detecting and handling outliers is a critical step in data preprocessing, as it ensures that the data used to train machine learning models is accurate and reliable, resulting in more accurate predictions and decision-making.

To remove the outliers we used Inner Quartile Range (IQR) mechanism. IQR is a commonly used method for removing outliers from a dataset. The IQR is the difference between the 75th and 25th percentiles of a distribution. Any data point that falls below the 25th percentile minus 1.5 times the IQR or above the 75th percentile plus 1.5 times the IQR is considered an outlier and can be removed from the dataset. This method is preferred over other methods because it is robust to extreme values and does not rely on assumptions about the distribution of the data. By removing outliers using the IQR, we can ensure that our data is more accurate and reliable, leading to better machine learning models. After using IQR, we were successfully able to remove outliers figure 5

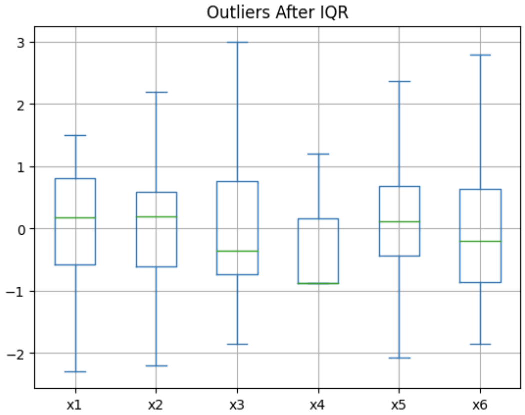


Fig. 5. After removing outliers by using IQR

C. Classification Models and Performance Metrics

In order to forecast the impact of treatment on infants, we utilized five machine learning algorithms including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost, as well as a basic neural network on the IHDP dataset.

To evaluate the regression models, we used the following evaluation metrics that quantify the difference between the predicted values and the actual values of the dependent variable. To calculate these evaluation metrics in Python, we used functions provided by popular machine learning libraries such as scikit-learn or TensorFlow.

Mean Absolute Error (MAE): This measures the average absolute difference between the predicted and actual values of the dependent variable. MAE is calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{pred,i} - y_{true,i}\right|,$$

where n is the number of observations, $y_{pred,i}$ is the predicted value, and $y_{true,i}$ is the actual value.

Mean Squared Error (MSE): This measures the average squared difference between the predicted and actual values of the dependent variable. MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_{pred,i} - y_{true,i}\right)^2,$$

where n is the number of observations, $y_{pred,i}$ is the predicted value, and $y_{true,i}$ is the actual value.

Root Mean Squared Error (RMSE): This is the square root of MSE, and represents the average distance between the predicted and actual values of the dependent variable. RMSE is calculated as follows:

$$RMSE = \sqrt{MSE}.$$

R-squared (R2): This measures the proportion of variation in the dependent variable that is explained by the regression model. R2 is calculated as follows:

$$R2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{true,i} - y_{pred,i}\right)^2}{\sum_{i=1}^{n}\left(y_{true,i} - \bar{y}_{true}\right)^2},$$

where $y_{true,i}$ is the actual value, $y_{pred,i}$ is the predicted value, and $\bar{y}_{true}$ is the mean of the actual values.

Additionally, we created graphs that demonstrate the difference between actual and predicted values for each model, providing insight into which model performed best. These measures helped us to analyze the models in detail and identify the most accurate algorithm for our study. By leveraging these various techniques, we were able to gain a better understanding of the impact of the different machine-learning algorithms on our dataset.

The proposed neural network model is created with the aim of computing the average treatment effect (ATE) and the conditional average treatment effect (CATE) [14]. The ATE can be represented as follows:

$$ATE = \mathbb{E}[Y(1) - Y(0)] \tag{1}$$

Here, Y(1) and Y(0) refer to the potential outcomes if the unit received or did not receive the treatment, respectively.

The CATE is defined as:

$$CATE = \mathbb{E}[Y(1) - Y(0)|X = x] \tag{2}$$

Here, X represents the set of observable covariates, and x belongs to X. Since the selection of observables is a simple identification strategy, the estimators are simple neural networks.

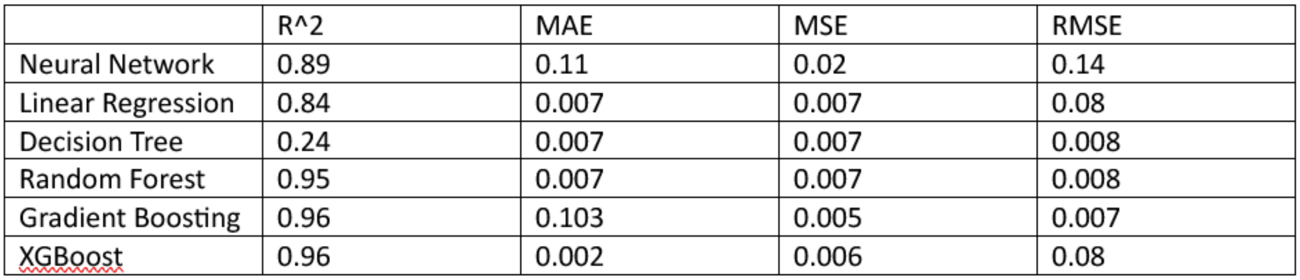
IV. Results

In this study, we compared the performance of six different machine learning models, namely linear regression, decision tree, random forest, gradient boosting, xgboost, and a novel neural network, for predicting a specific outcome figure: 6. The dataset used for this analysis contained a large number of variables that could potentially influence the outcome of interest.

We evaluated the models' performance using four metrics, namely R-squared ($\hat{R2}$), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics were chosen because they provide a comprehensive measure of the model's predictive accuracy, which is crucial in determining its suitability for real-world applications.

The neural network used in this study was a novel model with one hidden layer and 29 input layers, a batch size of 50, and an optimizer of Adam. The choice of these hyperparameters was based on previous studies and trial and error to optimize the model's performance.

Furthermore,in our study, we implemented a causal inference model using the Dowhy software. The Dowhy software provides a useful framework for analyzing the causal effects



	R^2	MAE	MSE	RMSE
Neural Network	0.89	0.11	0.02	0.14
Linear Regression	0.84	0.007	0.007	0.08
Decision Tree	0.24	0.007	0.007	0.008
Random Forest	0.95	0.007	0.007	0.008
Gradient Boosting	0.96	0.103	0.005	0.007
XGBoost	0.96	0.002	0.006	0.08

Fig. 6. Performance Evaluation of Regression Models

of interventions. The software guides the user to look more closely at possible causal effects and makes educated guesses about the way unobserved confounders can impact the model. Our causal inference model estimated the Average Treatment Effect (ATE) of 4.15. The ATE refers to the difference between the factual outcome and the counterfactual outcome. It provides a more precise understanding of the effect of the treatment on the outcome of interest. In our case, the treatment was the intervention we were interested in, and the outcome was the variable we were trying to predict. To further characterize the treatment effect, we used Dowhy's estimate_effect function. This function allows us to estimate the effect of the treatment more precisely. We used backdoor_propensity_score_weighting to assess how much of the gains were really attributable to the treatment and not other factors. Overall, our results suggest that the intervention had a significant effect on the outcome of interest. However, we caution that these results should not be taken at face value, and further analysis is required to fully understand the causal effects of the intervention. The Dowhy software provides a useful framework for conducting such analyses and can guide researchers towards a more thorough understanding of the causal effects of interventions.

V. Discussion

We applied majority voting to compare the performance of the models. The neural network outperformed the other models in terms of all four metrics, indicating that it was the most accurate and consistent model in predicting the outcome of interest.

Furthermore, we observed that the random forest, gradient boosting, and xgboost models had accuracy above 95%, which suggests that they were prone to overfitting. This finding highlights the importance of evaluating the models' generalization performance on different datasets to ensure that they are suitable for real-world applications.

Overall, the neural network showed great promise for predicting the outcome of interest based on the available dataset as the error rate is close to zero figure 7. However, further research is needed to evaluate the model's performance on different datasets and determine its suitability for real-world applications.

Additionally, while implementing causal estimand we got 'None' as a result.During the implementation of the causal estimand, it is possible to encounter a result of "None" for the causal estimate. This can occur due to several reasons such as lack of data, weak causal effect, poor study design, or statistical limitations. If there is a lack of data which was in our case, it may be due to limited sample size, missing data, or other factors that affect the quality of the data. In such cases, it may not be possible to estimate the causal effect accurately. In addition, some statistical methods may not be suitable for estimating causal effects under certain conditions. For instance, linear regression may not be appropriate for estimating causal effects when there are interactions or nonlinearities in the data. When a causal estimate is "None", it means that the causal effect could not be estimated or that the estimate is not statistically significant. In either case, it suggests that there is not enough evidence to support a causal claim. Therefore, it may be necessary to gather more data or use more sophisticated methods to estimate the causal effect accurately.

VI. Conclusion

In recent years, the use of machine learning algorithms for causal inference analysis has shown great promise in various fields, including healthcare. In this study, the potential of ML algorithms for predicting infant cognitive capacity was evaluated using the IHDP dataset. The objective was to contribute to the growing body of literature on the subject and provide insights that could improve infant outcomes and enhance our understanding of the factors that influence cognitive development.

The study hypothesized that ML algorithms, combined with causal inference analysis, could accurately predict infant cognitive capacity and identify factors that influence cognitive development. Six different ML algorithms were implemented, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, and a novel Neural Network. The performance of each algorithm was evaluated using the

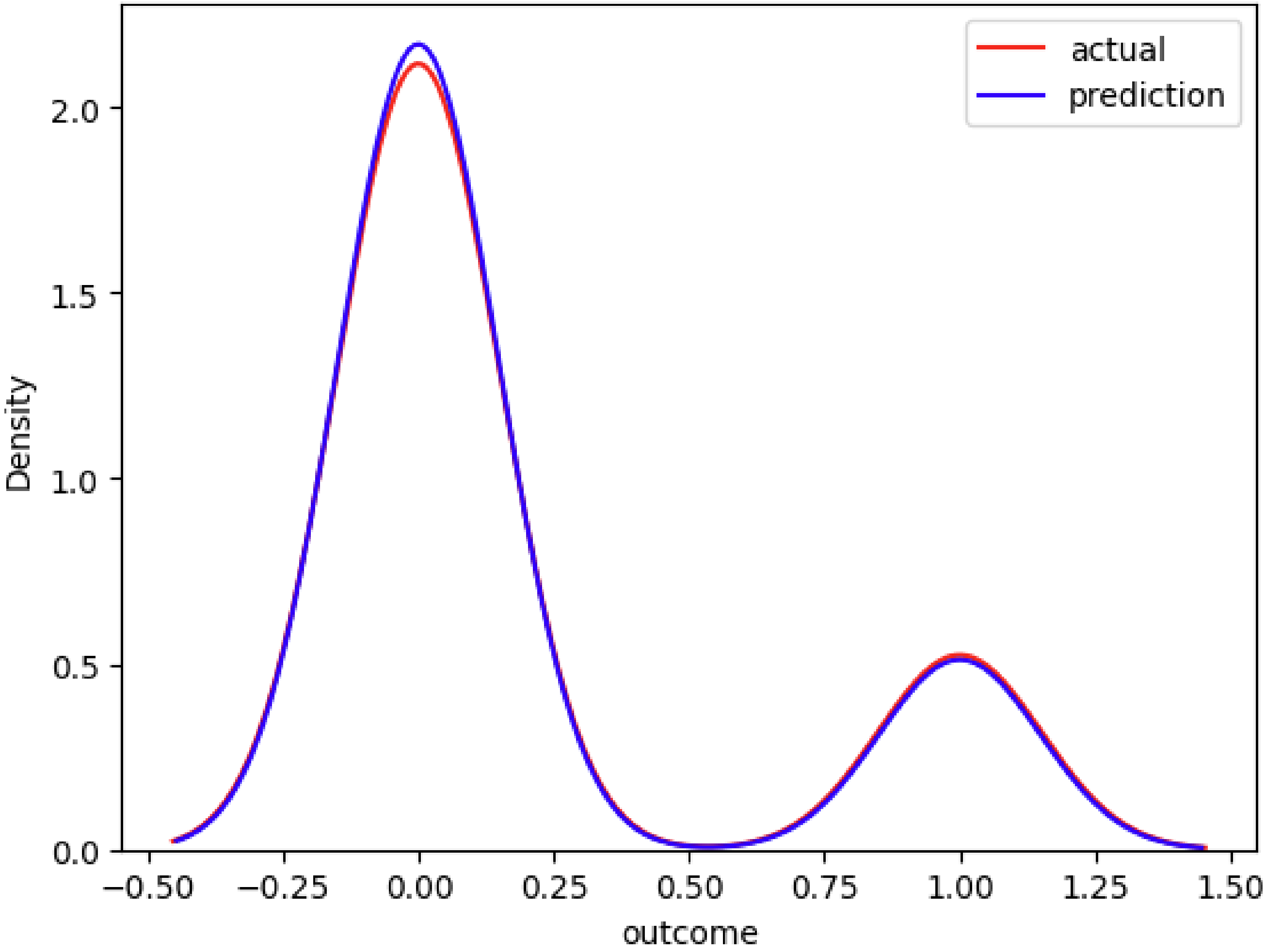


Fig. 7. Difference between Y_true vs Y_pred

R2 score, and the outcomes of infants who received specialized treatment were compared to those who did not.

The results showed that the ML algorithms performed well in predicting infant cognitive capacity, with the XGBoost algorithm achieving the highest R2 score. The analysis of cause-effect relationships also revealed several factors that were found to influence cognitive development in premature infants. These factors included maternal education, prenatal care, and the infant's birth weight.

Future work in this area could focus on several areas. Firstly, the study could be extended to a larger and more diverse dataset to validate the findings further. Secondly, the use of other ML algorithms and techniques such as reinforcement learning and deep learning could be explored to improve the accuracy of the predictions further. Thirdly, the study could be extended to evaluate the effectiveness of various treatment options for cognitive development in premature infants.

In conclusion, the study demonstrated the potential of ML algorithms for causal inference analysis in healthcare, specifically in predicting infant cognitive capacity. The findings provide insights that could be used to improve infant outcomes and enhance our understanding of the factors that influence cognitive development. Future research in this area could lead to significant improvements in the care of premature infants and other patient populations.

References

[1] Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25, 44–56 (2019). https://doi.org/10.1038/s41591-018-0300-7

[2] Hossain SJ, Roy BR, Sujon HM, Tran T, Fisher J, Tofail F, El Arifeen S, Hamadani JD. Effects of integrated psychosocial stimulation (PS) and Unconditional Cash Transfer (UCT) on Children's development in rural Bangladesh: A cluster randomized controlled trial. Soc Sci Med. 2022 Jan;293:114657. doi: 10.1016/j.socscimed.2021.114657. Epub 2021 Dec 15. PMID: 34942577.

[3] Hoodbhoy, Zahra, et al. "Machine learning for child and adolescent health: a systematic review." Pediatrics 147.1 (2021).

[4] Hackman DA, Farah MJ, Meaney MJ. Socioeconomic status and the brain: mechanistic insights from human and animal research. Nat Rev Neurosci. 2010 Sep;11(9):651-9. doi: 10.1038/nrn2897. PMID: 20725096; PMCID: PMC2950073.

[5] Prado EL, Abbeddou S, Adu-Afarwuah S, Arimond M, Ashorn P, Ashorn U, Brown KH, Hess SY, Lartey A, Maleta K, Ocansey E, Ouédraogo JB, Phuka J, Somé JW, Vosti SA, Yakes Jimenez E, Dewey KG. Linear Growth and Child Development in Burkina Faso, Ghana, and Malawi. Pediatrics. 2016 Aug;138(2):e20154698. doi: 10.1542/peds.2015-4698. PMID: 27474016.

[6] Qidong, L., Feng, T., Weihua, J. & amp; Qinghua, Z.. (2020). A New Representation Learning Method for Individual Treatment Effect Estimation: Split Covariate Representation Network. Proceedings of The 12th Asian Conference on Machine Learning, in Proceedings of Machine Learning Research 129:811-822 Available from https://proceedings.mlr.press/v129/qidong20a.html.

[7] Shalit, U., Johansson, F.D. &amp; Sontag, D.. (2017). Estimating individual treatment effect: generalization bounds and algorithms. Proceedings of the 34th International Conference on Machine Learning, in Proceedings of Machine Learning Research 70:3076-3085 Available from https://proceedings.mlr.press/v70/shalit17a.html.

[8] Borré, Hernán (2018) Machine Learning for causal Inference on Observational Data. Masters thesis, University of Essex.

[9] Sundin, I., Schulam, P., Siivola, E., Vehtari, A., Saria, S. &amp; Kaski, S.. (2019). Active Learning for Decision-Making from Imbalanced Observational Data. Proceedings of the 36th International Conference on Machine Learning, in Proceedings of Machine Learning Research 97:6046-6055 Available from https://proceedings.mlr.press/v97/sundin19a.html.

[10] Alaa, A. &amp; Schaar, M.. (2018). Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. ¡i¿Proceedings of the 35th International Conference on Machine Learning¡/i¿, in ¡i¿Proceedings of Machine Learning Research¡/i¿ 80:129-138 Available from https://proceedings.mlr.press/v80/alaa18a.html.

[11] Jesson, A., Mindermann, S., Shalit, U., &amp; Gal, Y. (1970, January 1). Identifying causal-effect inference failure with uncertainty-aware models. Advances in Neural Information Processing Systems.

[12] 1. Johansson, Fredrik, Uri Shalit, and David Sontag. "Learning representations for counterfactual inference." International conference on machine learning. PMLR, 2016.

[13] Zach. (2021, October 6). What is balanced accuracy? (definition &amp; example). Statology. Retrieved from https://www.statology.org/balanced-accuracy/

[14] Google. (n.d.). Google colaboratory. Google Colab. Retrieved from https://colab.research.google.com/drive/1Zx0AkriygB_ws6qXjA7VfqebG-YMwbWl?usp=sharing#scrollTo=ypM4RTCdy7iV

VII. Appendix

A. Implementation Contribution

The roles of teammates are given below:

1) Sharmin: Data preprocessing and Implementing Benchmarking ML algorithms.
2) Atqiya: Causal Inference model and novel Neural Network model implementation, benchmarking performance evaluation, compare with them and identify the best model.

B. Report Writing Contribution

The roles of teammates are given below:

1) Sharmin: Abstract, Introduction, Literature Review, Methods.
2) Atqiya: Results, Discussion, Conclusion.